III. 다변량분석

1. *행렬의 이해*

1-1 행렬

■ 정의

✓ 형렬의 표현: matrix X of order nxp $\mathrm{X}_{n\times P}=\{x_{ij}\}$행렬: $\mathrm{X}_{n\times P}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$

- i 는 행을, j 는 열을 나타내며,
- x_{ij}를 행렬의 원소(element)라 한다.

✓ 벡터

- 열의 차수가 1 인 행렬을 열 벡터(column vector; $\underline{x}$)
- 행의 차수가 1 인 행렬을 행 벡터(row vector; $\underline{x}'$)

■ 특수한 행렬

✓ 정방 행렬: n=p 인 행렬

- 대각합(trace) tr(X) = $\sum_{i=1}^{n} x_{ii}=\sum_{i=1}^{n}\lambda_i, \lambda_i = eigenvalues(X)$
- 물리적 의미가 무엇인가?
  - 모든 inputs 의 분산 합
- tr(A+B) = tr(A) + tr(B)
- tr(AB) = tr(BA)

✓ 대각 행렬: $x_{ij}=0\ (i\neq j)$인 정방 행렬

✓ 삼각 행렬

- 상삼각 행렬: $x_{ij}=0\ (i<j)$인 정방 행렬
- 하삼각 행렬: $x_{ij}=0\ (i>j)$인 정방 행렬

✓ 항등 행렬: I_n

- 대각 행렬이면서, $x_{ii}=1\ (for\ all\ i)$인 행렬
- 대수에서의 1 과 같은 역할을 한다. (AI=IA=A)

✓ 영행렬: $x_{ij}=0\ (for\ all\ i,j)$

✓ 일행렬: $x_{ij}=1\ (for\ all\ i,j)$

1-2 행렬의 연산

■ 행렬과 벡터의 연산

✓ (3x3) 행렬 A 와 (3x1) 벡터 b 의 연산

$$\mathrm{Ab} = (A_1, A_2, A_3)\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = b_1 A_1 + b_2 A_2 + b_3 A_3$$

✓ (nxp)행렬 Z 와 (1xp) v 벡터의 연산

- 의미
  - 행렬 Z 는 basis-$\vec{\mathrm{x}} \in R^d$에서 측정된 관측치
    ① 각 input 에 대해 평균을 0 으로 이동한 관측치
  - v 는 좌표변환할 새로운 좌표의 basis unit vector
- 전체 변동을 최대한 설명하는 제 1 주성분 벡터를 구하자.

$$\max_v \sum_{i=1}^{n} \frac{(z_i{}'v)^2}{n} > 0$$

  - 전체 변동을 최대한 설명하는 것은 전체 관측치의 변동을 최소화하는 방향의 벡터를 구한 것
    ① 이는 각 관측치에 대한 사영의 값을 최대화하는 것이다.
  - $\sum_{i=1}^{n}(z_i{}'v)^2 = \sum_i (v'z_i)(z_i{}'v) = v'(\sum_i z_i z_i{}')v = v'Z'Zv = (Zv)'(Zv)$

$$\sum_{i=1}^{n}(z_i{}'v)^2 = \sum_{i=1}^{n}\left(\sum_{j=1}^{p} z_{ij}{}' v_j\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{p} v_j z_{ij} z_{ik} v_k$$

$$v'Z'Zv = v'\left[\sum_{i=1}^{n} z_{ji} z_{ik}\right]v = \sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{i=1}^{n} v_j z_{ij} z_{ik} v_k$$

  - Lagrange Multiplier 를 이용

$$\phi(v, \lambda) = \frac{v'Z'Zv}{n} - \lambda(v'v - 1)$$

$$\frac{\partial \phi}{\partial \mathrm{v}} = 0 \;\Rightarrow \frac{Z'Z}{n} v = \lambda v$$

  $v'Z'Zv$ is positive definite matrix → $\frac{Z'Z}{n} = VD_\lambda V'$ (고유값 분해)
- 따라서, v 는 첫번째 고유값의 고유벡터이다.
- 중요한 시사점

$$\boxed{\sum_{i=1}^{n}(z_i{}'v)^2 = \|Zv\|^2 = (Zv)'(Zv) = v'Z'Zv}$$

✓ 주성분점수와 특이값 분해

- 3rd 주성분 점수 벡터
  - n 개의 관측치의 new unit vector(v)에 대한 좌표값(정사영)이다.
  - n 개의 관측치의 3rd 주성분에 대한 사영값 벡터(s=3)

$$a_s = \begin{pmatrix} z_1{}'v_s \\ \vdots \\ z_i{}'v_s \\ \vdots \\ z_n{}'v_s \end{pmatrix} = \mathrm{Z}v_s$$

  - 특징
    ① $1_n{}'a_s(=0) = 1_n{}'\mathrm{Z}v_s = (\cdots, \sum_{i=1}^{n} z_{ji}, \cdots)v_s = (\cdots, 0, \cdots)v_s$

② $\frac{a_s{}'a_s}{n} = v_s{}'\frac{Z'Z}{n}v_s = v_s{}'\lambda_s v_s = \lambda_s$

③ $a_s{}'a_t = v_s{}'\frac{Z'Z}{n}v_t = v_s{}'\lambda_s v_t = 0, (\mathrm{s} \neq \mathrm{t})$

- 주성분 점수 행렬

$$\mathrm{A}_{(n\times p)} = [\overrightarrow{a_1} \quad \overrightarrow{a_2} \quad \overrightarrow{a_3} \quad \cdots] = \mathrm{Z}_{(n\times p)}\mathrm{V}_{(p\times p)}$$

$$\frac{A'A}{n} = D_\lambda$$

- 분산이 모두 1 이 되도록 척도화할 때

$$u_s = \frac{a_s}{\sqrt{\lambda_s}}$$

$$\mathrm{U} = [\overrightarrow{u_1} \quad \overrightarrow{u_2} \quad \overrightarrow{u_3} \quad \cdots] = \mathrm{A}D_{1/\sqrt{\lambda}}$$

$$\frac{U'U}{n} = I_p$$

- U 는 단위분산의 직교열을 갖는다.

$$\mathrm{U}_{(n\times p)} = \mathrm{A}D_{1/\sqrt{\lambda}} = ZVD_{1/\sqrt{\lambda}} \Longrightarrow Z_{(n\times p)} = UD_{\sqrt{\lambda}}V'$$

$$\frac{\mathrm{U}'\mathrm{U}}{\mathrm{n}} = \frac{\left(\mathrm{A}D_{1/\sqrt{\lambda}}\right)'\left(\mathrm{A}D_{1/\sqrt{\lambda}}\right)}{\mathrm{n}} = \mathrm{D}_{1/\sqrt{\lambda}}\mathrm{D}_\lambda\mathrm{D}_{1/\sqrt{\lambda}} = \mathrm{V}'\mathrm{V} = \mathrm{I}_\mathrm{p}$$

$\mathrm{U}: (\mathrm{n}\times\mathrm{p}), \quad \mathrm{V}: (\mathrm{p}\times\mathrm{p}), \quad \mathrm{D}: (\mathrm{p}\times\mathrm{p})$

- 행렬 Q(= $Z/\sqrt{n}$)의 특이값 분해
  - n>=p 인 경우

$$Z = UD_{\sqrt{\lambda}}V', Q = WD_{\sqrt{\lambda}}V'$$

① W(= $U/\sqrt{n}$)

② μ=$\sqrt{\lambda}$이며, 특이값(singular value)라고 부른다.

③ W 는 p 개의 (nx1) 정규직교 열로 구성되며,

④ V 는 p 개의 (px1) 정규직교 열로 구성된다.

■ 정칙 행렬(non-singular matrix)

✓ 정칙인 정방 행렬 A

- det(A)≠0
- 또는 rank(A)=dim(A)
- 또는 $\mathrm{A}\mathrm{A}^{-1} = \mathrm{A}^{-1}A = I$인 A^{-1}가 유일하게 존재

✓ 정칙 행렬 연산의 특징

- rank(AB)=rank(B)
- AB = AC 이면, B = C

■ 양정치 행렬과 반양치 행렬

✓ 양정치(positive definite) 행렬

- 정의
  - x≠0 인 벡터 x 에 대해,
  - x′Ax > 0 이면, 행렬 A 를 양정치 행렬이라 한다.

- 특징
  - A=T′T 를 만족하는 정칙행렬 T 가 존재한다.
  - 대각원소 $a_{ii} > 0$
- 양정치 대칭행렬 A 은 다음과 동치다.
  - x≠0 인 벡터 x 에 대해, x′Ax > 0
  - A 의 고유값은 모두 양수이다.
  - A 의 부분행렬 A_k에 대해, det(A_k)>0
- 양정치 행렬의 분해
  - 양정치 행렬 A, 정칙 행렬 Q, 대각 행렬 D 에 대해
  - A= $\mathrm{QD}Q' = \mathrm{Q}\sqrt{D}\sqrt{D}Q' = \left(\sqrt{D}Q\right)\left(\sqrt{D}Q\right)' = T'T$

✓ 반양치(positive semidefinite) 행렬
- 정의
  - x≠0 인 벡터 x 에 대해,
  - x′Ax ≥ 0 이면, 행렬 A 를 반양치 행렬이라 한다.
- 특징
  - 대각원소 $a_{ii} \geq 0$

■ 전치

✓ $\mathrm{X}_{n\times p}$에 대해 $\mathrm{X}'_{p\times n}$이 있어서
- $x_{ij} = x'_{ji}\ (for\ all\ i,j)$이면, 전치(transpose)라 한다.
- $(\mathrm{X}')' = \mathrm{X}$ 이다.

✓ 열 벡터의 전치는 행 벡터이다.

■ 대칭 행렬

✓ 개념: $X' = \mathrm{X}$인 행렬을 symmetric matrix 라 한다.

✓ 예
- $\mathrm{X}'X$, $X\mathrm{X}'$은 항상 양정치 대칭 행렬이다.
- 공분산 $\sum$는 양정치 대칭 행렬이다.
- 대칭행렬 $\sum$에 대해 $\mathrm{A}\sum\mathrm{A}'$는 항상 대칭 행렬이다.

✓ 특징
- 벡터 a 에 대해: $\mathrm{a}\sum\mathrm{a}' = \sum_i {a_i}^2 s_{ii} + \sum_{i\neq j} a_i a_j s_{ij}$

■ 합 연산

■ 곱 연산

✓ 곱의 성질
- 일반적으로 AB ≠ BA 이다.
- $(\mathrm{AB})' = \mathrm{B}'\mathrm{A}'$

- Tr(AB) = Tr(BA) = $\sum_{i=1}^{N}\sum_{j=1}^{D} a_{ij}b_{ji}$
  - If $B = A'$, Tr(AB) = $\sum_{i=1}^{N}\sum_{j=1}^{D} a_{ij}{}^2$

✓ 곱의 연산 법칙
- 결합법칙: (AB)C = A(BC) = ABC
- 배분법칙: A(B+C) = AB + AC

✓ 멱등 행렬
- 정의
  - $M^2 = M$인 행렬 → $M^k = M$
  - 이때 M'도 멱등행렬이다.
  - 이때, $MM' = I$인가? ← $M^2M'^2 = MMM'M' = M(MM')M' = MM'$
- 특징
  - 변환을 한번 더 하면, 원래로 돌아오는 변환
  - 대칭 변환, 180 도 회전
  - <참고> [실수 → 허수], 90 도 회전(K)은 KKKK=K, KK 는 멱등 행렬

✓ 항등 벡터
- $\underline{e_i}$: i 번째 원소만 1 이고 나머지는 모두 0 인 벡터
  - <참고> $\underline{e_4}' = (0\ 0\ 0\ 1\ 0\ \ldots 0)$이다.
- 항등행렬 $I_n = \sum_{i=1}^{n} \underline{e_i}\underline{e_i}'$

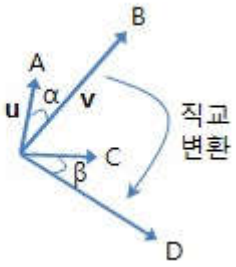
✓ 모든 원소가 1 인 벡터와 행렬: $\underline{1}_n$, $J_n = 1_n = \underline{1}_n\underline{1}_n'$

■ 직교 행렬

✓ $MM' = M'M = I$인 행렬
- 회전 변환을 나타내는 행렬은 직교행렬이다.
  - M'는 M 과 반대방향의 회전 변환을 나타낸다.
  - 돌리고(M), 다시 반대로 돌리면(M') 원래(I)로 돌아온다.
- 대칭 변환도 직교행렬이다.
  - 직교행렬간의 곱도 직교행렬이다.
- 회전 + 대칭 변환도 직교행렬이다.
  - (AT)′ (AT)= T′A′ AT= T′T = I

✓ 특징
- 직교행렬 A 를 곱해도 내적이 변하지 않음 → 회전하면 당연함
  - < Ax,Ay > = < x,y >
- 직교행렬 A 를 곱해도 벡터 길이는 변하지 않음
  - ‖Ax‖ = ‖x‖
- 직교행렬의 열벡터들이 R^n 정규직교 기저를 이룸

놈 보존 : $\|\mathbf{u}\|=\|T(\mathbf{u})\|$ $\|\mathbf{v}\|=\|T(\mathbf{v})\|$
거리 보존 : $d(A,B)=d(C,D)$
각도 보존 : $\alpha=\beta$



- 이와 같은 특징을 가지는 변환이 강체운동이다.

- det(직교행렬) = 1 or -1
- 직교행렬의 고유값은 실수 또는 공액복소수이고 절대값은 1

■ 역 행렬

✓ 나눗셈의 개념

✓ nxn 행렬 A 에 대한 행렬식 → 스칼라

$$\det(\mathrm{A}) = |\mathrm{A}| = \sum_{i=1}^{n} a_{ij}(-1)^{i+j}|M_{ij}| = \sum_{j=1}^{n} a_{ij}(-1)^{i+j}|M_{ij}|$$

- M_{ij}는 a_{ij}의 소행렬(minor), $(-1)^{i+j}|M_{ij}|$를 여인자(cofactor)라 한다.
  - M_{ij}는 행렬 A 에서 행 i 와 열 j 를 제거하고 얻는 (n-1)x(n-1)행렬
  - $|M_{ij}|$는 원소 a_{ij}의 소행렬식이라 한다.
  - 원소 a_{ij}의 여인자(여인소; cofactor) $A_{ij} = (-1)^{i+j}|M_{ij}|$이다.
- $\det(\mathrm{A}) = \prod_i \lambda_i$
- SAS 를 이용하여 간단하게 행렬식의 값을 얻는 방법이 있다.

✓ 행렬식의 성질

- $|\mathrm{A}'| = |\mathrm{A}|,\ |\mathrm{AB}| = |\mathrm{A}||\mathrm{B}|,\ |\mathrm{AB}| = |\mathrm{BA}|$
- $|\mathrm{A}^2|=|\mathrm{A}|^2$, 그러나 일반적으로 $|\mathrm{A}+\mathrm{B}| \neq |\mathrm{A}|+|\mathrm{B}|$
- 행렬 A 의 두 행 또는 두 열이 같으면 행렬식은 0 이다.
- 한 행(열)의 상수를 곱하여 다른 행에 더해도 행렬식의 값은 그대로
- 한 행(열)을 다른 행(열)의 선형결합으로 표현할 수 있으면 행렬식의 값은 0 이다.
  - 즉 각 행 또는 열이 독립이 아니면 행렬식은 0 이다.

✓ 역 행렬의 개념

- $|\mathrm{A}| \neq 0$인 정방행렬 A 에 대해 AB=BA=I 를 만족하는 행렬 B 를 행렬 A 의 역행렬이라 하며 A^{-1}로 나타낸다.
- $\mathrm{A}^{-1} = \frac{1}{|\mathrm{A}|}\mathrm{adj} = \frac{1}{|\mathrm{A}|}[A_{ij}]'$이다.

$$\mathrm{adj} = [A_{ij}]' = \begin{bmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{bmatrix}$$

  - adj는 A 의 수반 행렬(adjoin)로서, 여인자 행렬의 전치 행렬이다.
  - 따라서, 역행렬의 성분은 다음과 같이 표시할 수 있다.

$$A^{-1}{}_{(ij)} = \frac{(-1)^{i+j}|M_{ji}|}{\sum_i a_{(ij)}(-1)^{i+j}|M_{ij}|}$$

- $\det(A^{-1}) = \det(\mathrm{A})^{-1}$
- 역행렬이 존재하려면 정방행렬이며, 행렬식이 0 이 아니어야 한다.

✓ 연립방정식

- $A\underline{x} = \underline{b}$
- 해법
  - $\mathrm{A}^{-1}\mathrm{A}\underline{\mathrm{x}} = \underline{\mathrm{x}} = \mathrm{A}^{-1}\underline{\mathrm{b}}$
  - SAS/IML 을 이용하여 해법을 구할 수 있다.

✓ 역 행렬의 성질

- 역 행렬은 단 하나만 존재한다.
- $|\mathrm{A}^{-1}| = {}^{1}/_{|\mathrm{A}|}$
- $(\mathrm{A}^{-1})^{-1} = A$
- $(\mathrm{A}')^{-1} = (\mathrm{A}^{-1})'$
- $(\mathrm{AB})^{-1} = \mathrm{B}^{-1}\mathrm{A}^{-1}$

✓ 계수(Rank)

- 행렬에서 선형 독립인 행(열)의 수이다.
- 선형 독립인 벡터
  - $a_1\underline{x}_1 + a_2\underline{x}_2 + \cdots + a_p\underline{x}_p = 0$이 모든 $a_i = 0$일 때만 만족하는 경우
    ① 벡터 $\underline{x}_1, \underline{x}_2, \ldots . \underline{x}_p$는 선형 독립(linearly independent)이라 한다.
  - 그렇지 않은 경우, 상호 종속인 벡터는 다른 벡터의 선형 결합으로 표시할 수 있다는 것을 의미한다.
  - Row 가 column(inputs)보다 적은 경우, input 들이 실제 상호 독립일지라도, 데이터량이 적어 상호독립이라고 볼 수 없다.
- Full rank 의 의미
  - $\mathrm{X}_{n\times n}$정방행렬의 rank 가 n 인 경우 full rank 행렬이라 부른다.
  - 즉, rank($\mathrm{X}_{n\times n}$)=n 이면 full rank 이다.

✓ 정리

역 행렬이 존재한다.	역 행렬이 존재하지 않는다.
Full-rank 이다. Rank(A)=n	Full-rank 아니다. Rank(A)<n
A 는 non-singular 이다.	A 는 singular 이다.
|A|≠0	|A|=0
Ax=b 의 유일해가 존재한다.	Ax=b 의 유일해가 존재하지 않는다.

■ 행렬 미분

✓ 벡터 미분

- 상수 벡터 $\underline{a} = \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_p \end{bmatrix}$, 확률변수 벡터 $\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_p \end{bmatrix}$라 하면,
  - $\frac{\partial}{\partial \underline{x}}(\underline{a}'\underline{x}) = \underline{a}$
  - $\frac{\partial}{\partial \underline{x}}(\underline{x}'\underline{a}) = \underline{a}$
  - $\frac{\partial}{\partial \underline{x}}(\underline{x}'A\underline{x}) = A\underline{x} + A'\underline{x}$ ➔ (A 가 대칭행렬이면) $= 2A\underline{x}$

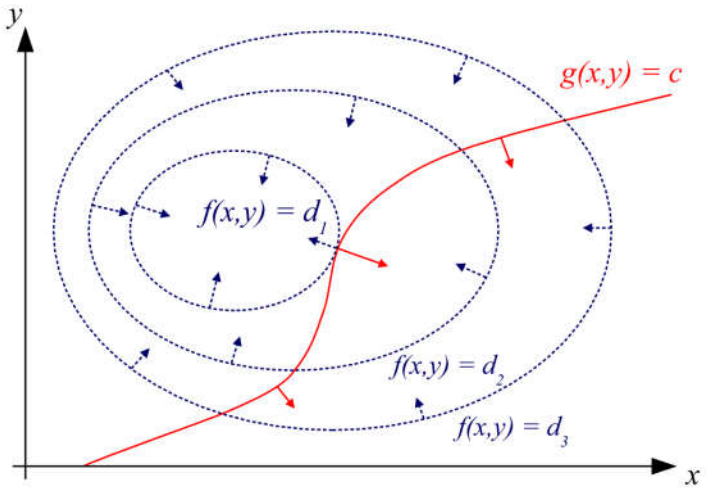
✓ 행렬 미분

-

■ Lagrange Multiplier(라그란지 승수)

✓ 개념

- 어떤 제약 조건(constraint)에서 목적함수의 최대/최소값을 구하는 방법
  - Constraint: $g(\vec{x}) = 0$
  - 목적함수: $f(\vec{x})$를 최대화하는 $\vec{x}$를 찾고, 그 최대값을 구하는 방법
- 해법: $\frac{\partial}{\partial \vec{x}}(f(\vec{x})) = \lambda \frac{\partial}{\partial \vec{x}}(g(\vec{x}))$, $g(\vec{x}) = 0$
  - $f(\vec{x})$가 최대(최소)가 되는 지점에서 $f(\vec{x})$와 $g(\vec{x})$의 기울기가 같다.
  - 즉, 그 위치에서 $\frac{dx_j}{dx_i}$의 값이 $f(\vec{x})$와 $g(\vec{x})$가 모두 같다.
  - 이때, 임의의 상수 $\lambda(\neq 0)$에 곱해져도 비율은 변하지 않는다.



✓ Lagrangian

- 제약함수가 여러 개 일 때,
  - $g_1(\vec{x}) = 0$, ..., $g_k(\vec{x}) = 0$
- 목적함수: $f(\vec{x})$
- $L(\vec{x}, \lambda_1, .., \lambda_k) = f(\vec{x}) - \lambda_1 g_1(\vec{x}) - \cdots - \lambda_k g_k(\vec{x})$

✓ 일반화 해법

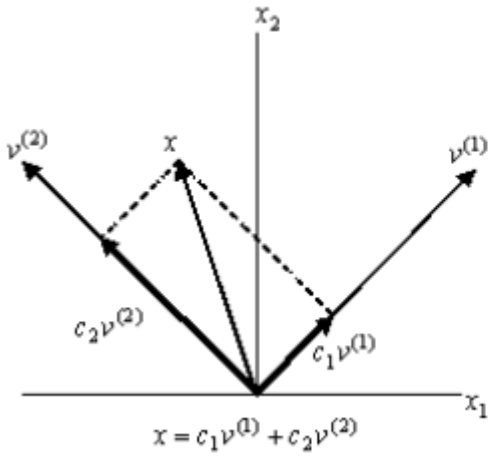
$$\vec{\nabla}L = \vec{\nabla}f(\vec{x}) - \lambda_1 \vec{\nabla}g_1(\vec{x}) - \dots - \lambda_k \vec{\nabla}g_k(\vec{x}) = 0$$

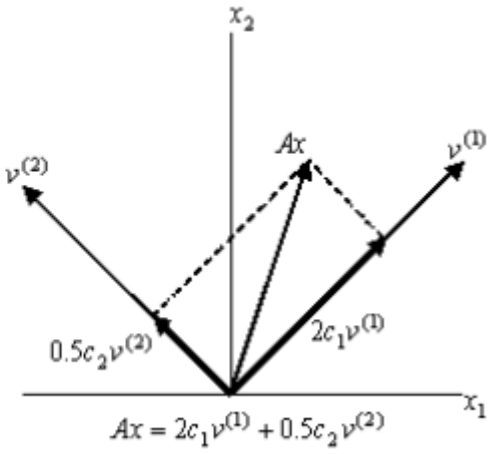
- $\vec{\nabla}f(\vec{x}) = \lambda_1 \vec{\nabla}g_1(\vec{x}) + \dots + \lambda_k \vec{\nabla}g_k(\vec{x})$
- 제약조건: $g_i(\vec{x}) = 0$, for all i

1-3 고유치와 고유벡터

■ 고유방정식(characteristic equation)

✓ 의미





- NxN 정방행렬은 보통 N 차원 좌표축의 변환에 사용된다.
- 어떤 선형 변환의 **고유벡터**는 변환 후에 크기만이 변하고 방향은 일정한 벡터를 가리킨다.
  - <참고> 삼차원 회전변환의 고유벡터는 그 회전축 상에 놓여 있다.
    ① 지구의 자전축은 자전변화에 대한 고유벡터이다.
  - 아래 팽창막의 경우에는 선형변환이지만 회전변환은 아니다.
  - 고유벡터의 **고유값**은 변환 전과 후의 고유벡터의 크기 비율이다.

✓ 고유벡터를 구하는 고유방정식은 $\boldsymbol{A\underline{x} = \lambda\underline{x}}$이다.

- 고유방정식의 해석
  - Ax(변환한 벡터가) 원래 벡터(x)보다 크기가 λ 만큼 크지만,
  - 방향은 동일(=)한 벡터(고유벡터)가 있다.
- 고유방정식으로부터 모든 고유벡터를 구한다.
- 특성방정식 $(A - \lambda\mathrm{I})\underline{x} = \underline{0}$으로 변형될 수 있고,
- $\det(A - \lambda\mathrm{I}) = 0$ 과 같이 바꿔쓸 수 있다.
  - 이 식으로부터 n 개의 고유값 λ를 구하고,
  - 이 λ를 $(A - \lambda\mathrm{I})\underline{x} = \underline{0}$에 대입하면, 고유벡터 $\underline{x}$를 구할 수 있다.
  - 이때, 하나의 고유값 λ에 대해 $\underline{w}$, $\underline{x}$가 고유벡터라면,
    ① $\underline{w} + \underline{x}$도 고유벡터이고
    ② $k\underline{x}$도 고유벡터이다.
    ③ 같은 고유값에 해당하는 벡터들은 0 벡터와 함께 하나의 벡터공간을 이루며, 이 공간을 고유값 λ에 대한 고유공간이라 함.
    ④ 이것을 이용하여 요인분석에서 요인회전이 가능하다.
- 또한 정방행렬 A의 전치 A'는 A와 같은 고유값을 가진다.
- 고유값이 가장 큰 고유벡터를 주 고유벡터라 한다.

✓ 특징

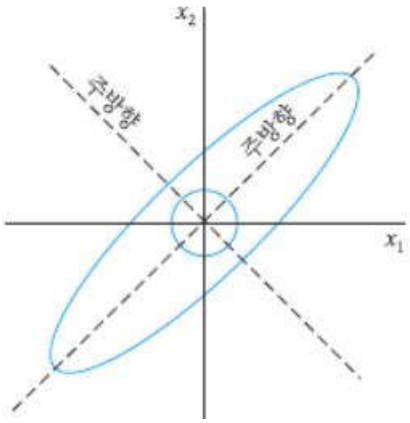
- $\mathrm{tr}(\mathrm{X}) = \sum_{i=1}^{n} x_{ii} = \sum_{i=1}^{n} \lambda_i$
- $\det(\mathrm{X}) = \prod_i \lambda_i$

■ 고유값 문제의 응용

✓ 예 1: 탄성막의 팽창 → 선형변환

- 경계로서 $x_1{}^2 + x_2{}^2 = 1$을 갖는 x_1x_2평면상의 탄성막
- 이 탄성막을 잡아당겨 점 $\underline{\mathrm{x}}(\mathrm{x}_1, \mathrm{x}_2)$가 $\underline{\mathrm{y}} = \begin{pmatrix} \mathrm{y}_1 \\ \mathrm{y}_2 \end{pmatrix} = \mathrm{A}\underline{\mathrm{x}} = \begin{pmatrix} 5 & 3 \\ 3 & 5 \end{pmatrix} \begin{pmatrix} \mathrm{x}_1 \\ \mathrm{x}_2 \end{pmatrix}$로 이동
  - 주방향 즉, 고유벡터($\underline{\mathrm{x}}$의 방향과 같은 $\underline{\mathrm{y}}$)을 구해보자
  - 이런 변형하에 경계원은 어떤 모양을 갖는지 알아보자
- 특성방정식

  $\begin{vmatrix} 5-\lambda & 3 \\ 3 & 5-\lambda \end{vmatrix} = (5-\lambda)^2 - 9 = 0$ ➔ 고유값 λ=8, λ=2
  - λ=8 일 때 고유벡터 $\underline{\mathrm{a}} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$
  - λ=2 일 때 고유벡터 $\underline{\mathrm{a}} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$
- 주방향은 양의 x_1축과 45^o와 135^o의 각을 이루는 방향(고유벡터의 방향)
  - 주방향으로 각각 8 과 2 만큼 팽창
  - 새로운 좌표계 $\mathrm{y}_1 = 8\cos\vartheta,\ \mathrm{y}_2 = 2\sin\vartheta$



$\frac{{y_1}^2}{8^2} + \frac{{y_2}^2}{2^2} = 1$ (타원)

✓ 예 2: Markov 과정에서 발생하는 고유값 문제

- 장기발전모델에 있어 변환후에도 자기자신이 되는 $A\underline{x} = \underline{x}$모델
- 이때, 고유값은 무조건 1 이 된다.

$A = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.1 & 0.9 & 0 \\ 0 & 0.2 & 0.8 \end{bmatrix}$ ➔ $A\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.1 & 0.9 & 0 \\ 0 & 0.2 & 0.8 \end{bmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$

  - A 의 고유값은 1
  - 이때 고유벡터는 $\begin{pmatrix} 2 \\ 6 \\ 1 \end{pmatrix}$
- 확률행렬 A 가 변하지 않는다는 가정하에 상업:공업:거주지역 비율은
  - 2:6:1 로 근사된다.

■ 대칭행렬의 활용

✓ 다변량에서 이용하게 될 공분산행렬이나 상관행렬은 모두 대칭행렬이다.

- <참고> 좌표계의 회전을 나타내는 행렬은 대칭행렬이 아니다.

✓ **대칭행렬의 성질**

- 고유치는 실수이다.
  - 대각 원소가 양이면, 고유치는 양수이다.
- 대칭행렬은 대각화(Diagnosable)가 가능하다.
  - $A = U^{-1}D_{\lambda}U,\ A^{-1} = U^{-1}D_{1/\lambda}U$
  - D 는 대각원소가 A 의 고유치인 대각행렬
  - U 는 직교행렬($U^T = U^{-1}$)
- 고유벡터는 서로직교(orthogonal)다. 즉, $\underline{u}_i{}'\underline{u}_j = 0, for\ \ i \neq j$
  - 서로 직교인 고유벡터로 basis 를 변경시 대각화가 가능하다.
- Rank (of the Matrix) = # of eigenvalue(nonzero)
  - 0 인 고유치가 존재하는 행렬은 full-rank 가 아니며, 역행렬이 없다.

✓ 공분산 행렬의 생성

- nxm(n>m)이고 Rank=m 인 행렬 X 에 대해
  - $X^TX = A$는 대각 원소가 모두 양인 대칭행렬이 된다.
- 공분산 행렬과 상관 행렬은 이와 같은 형태로 만들어진다.

■ 평균벡터

✓ 자료행렬

$$\mathrm{X}_{n\times P} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = [\underline{x}_1 \quad \underline{x}_2 \quad \cdots \quad \underline{x}_p],\ \underline{x}_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \cdots \\ x_{ni} \end{bmatrix}$$

- p 개의 변수벡터 만큼 존재하는 $\underline{x}_i$는 한 변수벡터의 모든 관측치이다.
- $\underline{x}_i$의 각 요소는 변수 i 의 n 개의 관측치를 나타낸다.

✓ 이에 대한 평균벡터는 다음과 같이 정의된다.

$$\overline{\underline{x}} = \left[\left(\frac{1}{n}\right)\underline{1}_n{}'\ \mathrm{X}_{n\times P}\right]' = \begin{bmatrix} \Sigma x_{i1}/n \\ \Sigma x_{i2}/n \\ \cdots \\ \Sigma x_{ip}/n \end{bmatrix}$$

- **각 변수들에 대한 n 개의 관측치에 대한 평균값을 나타낸다.**
  - 다음 가족수(X_1), 학벌(X_2), 일용돈(X_3) 변수의 평균 벡터를 구한다면,

$$\overline{\underline{x}}' = \left(\frac{1}{4}\right)[1\ \ 1\ \ 1\ \ 1]\begin{bmatrix} 2 & 3 & 2 \\ 4 & 2 & 3 \\ 2 & 1 & 1 \\ 5 & 2 & 2 \end{bmatrix} = \left(\frac{1}{4}\right)[13\ \ 8\ \ 8] = [3.25\ \ 2\ \ 2]$$

■ 공분산 행렬과 상관 행렬

✓ 공분산 행렬

- j 번째 변수와 k 번째 변수의 공분산은

$$\sigma_{jk} = cov(x_j, x_k) = \frac{1}{n-1}\sum_{1=1}^{n}(x_{ij} - \overline{x_j})(x_{ik} - \overline{x_k})$$

- j=k 이면, 공분산은 분산이 된다.

$$\sigma_{kk} = \mathrm{var}(x_k) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \overline{x_k})^2$$

- 그러므로 공분산 행렬 Σ는 다음과 같이 나타낼 수 있다.

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

- 표본 데이터는 σ대신 s 를 사용하고, ∑기호 대신 S 를 사용한다.

✓ 상관 행렬

- j 번째 변수와 k 번째 변수의 상관계수:

$$r_{jk} = \frac{cov(x_j, x_k)}{\sqrt{var(x_j)var(x_k)}}$$

- 상관계수 행렬: $R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \cdots & \cdots & 1 & \cdots \\ r_{n1} & r_{n2} & \cdots & 1 \end{bmatrix}$

✓ 공분산 행렬의 고유치

- 고유치는 $\det(\Sigma - \lambda I) = 0$의 해이다.
- 공분산 행렬은 대칭행렬이므로
  - 고유치의 값은 실수이고, 행렬 차수 만큼의 고유치가 존재한다.

✓ 공분산 행렬의 고유벡터

- 공분산 행렬 Σ의 고유치를 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$라 하면
  - $\Sigma\underline{u_i} = \lambda_i\underline{u_i}$를 만족하는 고유벡터 $\underline{u_i}$를 구할 수 있다.
  - 고유벡터는 무수히 존재할 수 있으나, 다변량분석에서 고유벡터는 다음식을 만족하는 것을 고유벡터로 사용한다.

    $\underline{u_i}'\underline{u_i} = 1,\ \underline{u_i}'\underline{u_j} = 0, for\ i \neq j$ ← **전체 분산이 변하지 않도록 함.**
- 고유벡터의 고유치는 원 변수의 변동량에 대한 설명력이다.

■ 고유치의 기하학적 해석

✓ 평균 $\underline{\mu} = \begin{bmatrix} 10 \\ 5 \end{bmatrix}$이고 공분산 행렬 $\Sigma = \begin{bmatrix} 9 & 2 \\ 2 & 4 \end{bmatrix}$인 2 변량 정규분포

- Σ의 고유치는 $\lambda_1 = 9.7,\ \lambda_2 = 3.3$
  - $\mathrm{Tr}(\sum) = \sum_i \lambda_i$
- 고유벡터는 $\underline{u_1} = \begin{bmatrix} 0.94 \\ 0.33 \end{bmatrix}$, $\underline{u_2} = \begin{bmatrix} -0.33 \\ 0.94 \end{bmatrix}$이다.

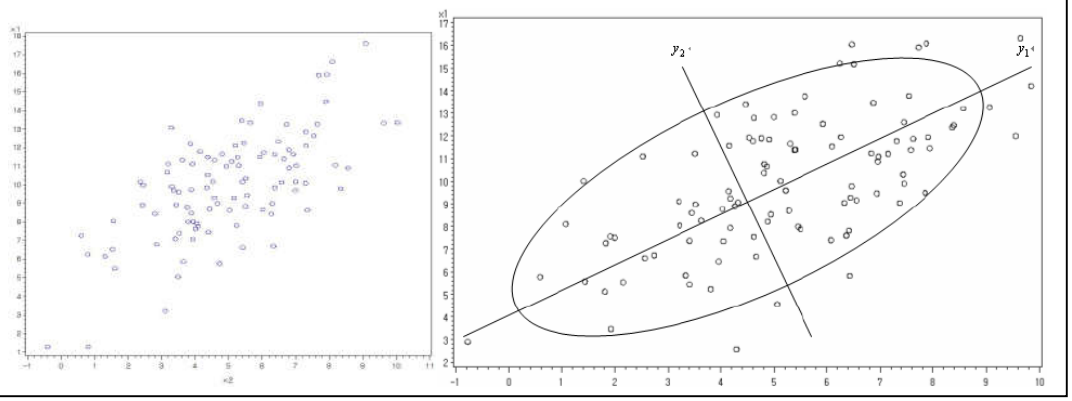
✓ 2 변량 자료에 대한 산점도를 그리면 아래와 같다. – 검토 필요

- 타원의 방정식은 $\frac{{y_1}^2}{\lambda_1}+\frac{{y_2}^2}{\lambda_2}=3(3=p+1)$이고
- 장축의 길이는 $2\sqrt{3\lambda_1}=10.8$, 단축의 길이는 $2\sqrt{3\lambda_2}=6.2$이다.
- y_1방향에서의 분산은 $\lambda_1=9$이고, y_1방향에서의 분산은 $\lambda_2=4$

✓ 고유방정식$[\lambda^2-(V_1+V_2)\lambda+(V_1V_2-COVxy^2)=0]$으로부터,

- $\lambda_1=\lambda_2$이면
  - $V_1=V_2$, $r^2=0$가 되는 완전제곱식이어야 한다.
  - 따라서 자료의 산점도는 원의 형태가 되고 상관계수는 0 이다.
- $\lambda_2=0$이라면,
  - $V_1V_2-COVxy^2=0$ → 상관계수 $r^2=\frac{COVxy^2}{V_1V_2}=1$이 된다.
  - 즉, λ_2 → 0 면 자료는 직선 상에 모이게 되고, r →1 이다.
  - 이와 같이 고유치의 값은 두 변수간의 상관관계를 나타내는 지표가 되며 다변량은 이를 일반화한 것이다.
- 고유벡터 y_1과 y_2는 주성분이며,
  - 타원의 길이는 주성분의 원변수변동에 대한 설명력이다.

```
data one;
   do i=1 to 100;
   z1=rannor(0);
   z2=rannor(0);
   x1=10+3*z1;
   x2=5+2/3*2*z1+2*sqrt(1-(2/3)**2)*z2;
   output;
   end;
run;
proc gplot data=one;
   symbol v=circle;
   plot x1*x2;
run;
```



1-4 스펙트럼 분해

■ 개념

✓ (pxp) 대칭행렬 A 는 다음과 같이 분해된다.

- A = PDP′ = $\sum_i \lambda_i \hat{e}_i \hat{e}_i'$ → P′AP = D
  - 여기서 λ_i는 A 의 고유값, $\hat{e}_i$는 A 의 단위 고유벡터

- P 는 각 열이 $\hat{e}_i$로 구성된 직교 행렬
- D 는 A 의 고유값으로 구성된 대각 행렬

▪ = $A^{-1} = PD^{-1}P' = \sum_i {}^{1}/_{\lambda_i}\, \hat{e}_i\hat{e}_i'$

✓ 대각화

▪ P'AP = D → "A 는 직교행렬 P 에 의해 대각화된다."라고 한다.

■ 활용

✓ 제곱근 행렬(square root matrix)

▪ 정의

- (pxp) 양 정치행렬에 대해
- $A^{1/2}A^{1/2} = \mathrm{A}$를 만족하는 $A^{1/2}$을 제곱근 행렬이라한다.

▪ 스펙트럼 분해를 이용하면, $A^{1/2} = PD^{1/2}P' = \sum_i \sqrt{\lambda_i}\hat{e}_i\hat{e}_i'$

- 여기서 λ_i는 A 의 고유값, $\hat{e}_i$는 A 의 단위 고유벡터
- P 는 각 열이 $\hat{e}_i$로 구성된 직교 행렬
- D 는 A 의 고유값으로 구성된 대각 행렬

✓ 멱등 행렬(idempotent matrix)

▪ $A^2 = \mathrm{A}$

- $\mathrm{A} = \mathrm{I} - \frac{1}{n}J$
- $A^2 = \left(\mathrm{I} - \frac{1}{n}J\right)\left(\mathrm{I} - \frac{1}{n}J\right) = I - \frac{2}{n}J + \frac{1}{n^2}JJ = \mathrm{I} - \frac{1}{n}J = A$

▪ $\mathrm{A} = \mathrm{X}(X'X)^{-1}X'$, 여기서 X 는 정방행렬이 아닐 수 있다.

- $A^2 = \mathrm{X}(X'X)^{-1}X'\mathrm{X}(X'X)^{-1}X' = \mathrm{X}(X'X)^{-1}X' = A$

1-5 SAS/IML 활용

■ 연립 방정식의 해를 구하기

✓ 연립방정식

$$\begin{Bmatrix} 2u + 2v - w = 5 \\ u + v - 2w = 1 \\ u - w = 4 \end{Bmatrix}$$

✓ Proc iml 사용하기

```
proc IML;
   a={2 2 -1, 1 1 -2, 1 0 -1};
   b={5, 1, 4}
   x=inv(a)*b;
   aprime=a`;
   print x, a;
run;
```

X	A		
5	2	2	-1
-1	1	1	-2
1	1	0	-1

▪ Print 옵션은 행렬의 output 을 윈도우에 출력한다.

- 만약 print x a;를 사용하면 a 가 x 옆에 출력된다.

▪ Proc iml; 문장 다음으로 reset print;를 사용하면 모든 결과가 출력된다.

■ 공분산 행렬과 고유치, 고유벡터를 구하기

```
data one;
    input x1 x2;
    cards;
    1 3
    3 7
    5 9
    7 13
;
run;
proc corr data=one cov outp=out1;
run;
proc print data=out1;
run;
```

데이터에는 변수가 2개 관측치가 4개이다. COV 옵션에 의해 공분산을 구한다. 그 결과를 OUT1에 저장한다.

TYPE	_NAME_	x1	x2
COV	x1	6.6667	10.6667
COV	x2	10.6667	17.3333
MEAN		4.0000	8.0000
STD		2.5820	4.1633
N		4.0000	4.0000
CORR	x1	1.0000	0.9923
CORR	x2	0.9923	1.0000

```
data out2;
    set out1;
    if _type_="COV";
    keep x1 x2;
run;

proc print data=out2;
run;
```

자동 생성 변수 _TYPE_에서 "COV"인 관측치만 선택하고 변수는 X1, X2만 남긴다.

x1	x2
6.6667	10.6667
10.6667	17.3333

```
proc iml;
reset print;
use out2;
read all into x;
call eigen(m,e,x);
run;
```

USE 옵션은 SAS data를 사용하여 행렬을 만들 때 사용한다. READ는 SAS data 데이터를 into 행렬 X로 만들라는 의미이다. M은 고유치, E는 고유벡터를 출력한다.

```
X    2 rows    2 cols    (numeric)
  6.6666667 10.666667
  10.666667 17.333333

M    2 rows    1 col     (numeric)
  23.925696
  0.0743041

E    2 rows    2 cols    (numeric)
  0.5257311 0.8506508
  0.8506508 -0.525731
```

■ SAS - exercise

- ✓ 특수 행렬의 성질
  - **A 가 직교 행렬이면 |A|=±1**
  - **A 가 멱등 행렬이면 |A|=0 또는 1**
  - $|\boldsymbol{A}|^{\boldsymbol{k}} = |\boldsymbol{A}^{\boldsymbol{k}}|$
  - $\mathbf{tr}(\mathbf{AB}) = \mathbf{tr}(\mathbf{BA})$ **(단, A B 가 정방행렬)**
  - **XX'은 대칭행렬**
- ✓ SAS/IML 을 이용한 계산
- ✓ 상관행렬과 고유치, 고유벡터 구하기

2. *주성분분석*

2-1 개념

■ 의의

✓ 기성복 바지를 살 때 우리 몸의 치수를 모두 알아야 할까?

- 허리둘레와 기장만 알만 충분하다.
- Principal component analysis(주성분분석)이 숨어있다.
- 변수 정보를 축약한 변수를 주성분 변수라 한다.

✓ 주성분분석의 의의

- 다양한 변수를 축약하여 1~2 개의 지표 변수를 만들 때 사용한다.
- **p≥3 인 변수를 1~2 개의 주성분 변수로 줄이고**, 이를 정의하는데 목적
  - **전체의 80%를 설명할 수 있으면 충분하다**.

■ 맛보기

✓ 수식

구분	Original Input	Transform matrix	주성분
설명	ist 관측치[D, 1]	각 열이 ∑의 고유벡터[D, L]	ist 관측치[L, 1]
주성분 변환식 '='가 성립 하려면 D=L	$\vec{x_i}\left[=\begin{pmatrix} x_{i1}\hat{e}_1 \\ x_{i2}\hat{e}_2 \\ x_{ij}\hat{e}_j \\ \vdots \\ x_{iD}\hat{e}_D \end{pmatrix}\right] =$	$W\left[=\begin{pmatrix} ev(\Sigma)_1 & ev(\Sigma)_2 & ev(\Sigma)_j & ev(\Sigma)_L \\ \downarrow & \downarrow & \downarrow & \downarrow \\ \downarrow & \downarrow & \downarrow & \downarrow \\ \vdots & \vdots & \vdots & \vdots \\ \downarrow & \downarrow & \downarrow & \downarrow \end{pmatrix}\right]$	$\vec{z_i}\left[=\begin{pmatrix} z_{i1}\hat{e}'_1 \\ z_{i2}\hat{e}'_2 \\ z_{ij}\hat{e}'_j \\ \vdots \\ z_{iL}\hat{e}'_L \end{pmatrix}\right]$
When $\vec{z_i} = \hat{e}'_1$: 1st 주성분은 1st 고유벡터	$\begin{pmatrix} ev(\Sigma)_1 \\ \downarrow \\ \downarrow \\ \vdots \\ \downarrow \end{pmatrix}$	$\begin{pmatrix} ev(\Sigma)_1 & ev(\Sigma)_2 & ev(\Sigma)_j & ev(\Sigma)_L \\ \downarrow & \downarrow & \downarrow & \downarrow \\ \downarrow & \downarrow & \downarrow & \downarrow \\ \vdots & \vdots & \vdots & \vdots \\ \downarrow & \downarrow & \downarrow & \downarrow \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$
일반화	$\vec{x_i} =$	$z_{i1}\overrightarrow{W_1} + z_{i2}\overrightarrow{W_2} + \cdots + z_{i2}\overrightarrow{W_2}$	$\overrightarrow{W_1} = ev(\Sigma)_1 = \hat{e}'_1$

■ 주성분분석의 활용

✓ 데이터 스크린

- p≥3 인 변수를 가진 데이터를 저차원(2 차원) 그래프로 나타내어 개체들의 특성이나 이상치의 존재 여부를 알아보는 것이다.

✓ 군집

- p≥3 인 다변량 데이터는 산점도 만으로는 해석에 어려움이 있음
- 주성분에 의해 개체를 분류하거나, **군집분석 결과에 대한 해석**으로 주성분분석을 사용한다.

✓ 판별분석

- 측정변수가 너무 많으면 계산이 오래걸려 주성분 분석 방법으로 변수의 수를 줄여 판별분석을 실시한다.
- 이 경우 컴퓨터의 발달로 인해 주성분 분석을 거의 사용하지 않는다.

✓ 회귀분석

- 다중 회귀분석에서 설명변수간의 상관관계가 높으면 다중공선성에 의한 회귀계수의 분산이 커져 최소자승 추정치를 믿을 수 없게 된다.
  - 이 경우 **문제가 되는 설명변수를 제외**한다.
  - Ridge(능형) Regression 으로 추정치의 불편성을 희생하고 최소분산을 갖는 추정치를 구한다.
  - **주성분 변수를 설명변수로 이용하여 회귀분석을 실시**한다.
    ① 주성분 변수들은 서로 독립이라는 성질을 이용하는 방법이다.
    ② 주성분이 명확히 해석되지 않으면 회귀분석의 해석이 어렵다.

2-2 주성분 구하기

■ 기본 원칙

✓ **주성분변수간에는 서로 상관관계가 전혀 존재하지 않는다.(독립)**

✓ 첫 주성분은 데이터의 변동(분산, 정보)을 가장 많이 설명한다.

- 차례로 다음 주성분은 나머지 정보들을 설명하고 점점 설명하는 크기는 줄어든다.
- 원 변수 벡터의 선형 결합 형태로 주성분을 표현할 수 있다.
- 주성분벡터 표현형식: $\underline{y} = L\underline{x}$

■ 주성분 정의

✓ 원변수의 선형결합인 주성분변수로 원변수의 공분산 구조를 설명하는 방법

- **원 변수의 변동 합과 주성분 변수의 변동 합은 동일**하다.

✓ 이때 선형 계수(L)를 구하는 방법은?

- 변수 벡터 $\underline{x}$가 공분산 행렬 $\sum$를 갖는다고 하자.
- 공분산 행렬의 고유치를 $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p$, 고유벡터를 $\underline{e}_1$ $\underline{e}_2$ $\cdots \underline{e}_p$라 하자.
  - $|\Sigma - \lambda I| = O$ ← 변수 x_i들의 변동합은 고유치의 합과 동일하다.
  - $\Sigma \underline{e}_i = \lambda_i \underline{e}_i$ $(\underline{e}_i' \underline{e}_i = 1, \underline{e}_i' \underline{e}_j = 0)$
- 이때, $\boldsymbol{y_i = \underline{e}_i' \underline{x}}$라 하면,
  - 각 관측치의 주성분 $\underline{e}_i$에 대한 사영인 y_i의 분산은 $\underline{e}_i$의 고유치이다.
  - $Var(y_i) = \underline{e}_i' \Sigma \underline{e}_i = \lambda_i$ ← 주성분 사영 벡터 y_i의 변동합은 x_i의 변동합과 같다.
  - $\mathrm{Cov}(y_i, y_j) = \underline{e}_i' \Sigma \underline{e}_j = 0, \text{for } i \neq j$
  - 따라서, k 번째 주성분의 원변수의 변동 설명 비율은 $\frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}$이다.

■ **주성분 계산**

✓ 첫번째 주성분(first principle component)

- $\underline{e_1}'\underline{e_1} = 1$를 만족하면서 $\underline{e_1}'\left(\underline{x} - \underline{\mu}\right)$**의 분산을 최대화하는** $\underline{e_1}$을 찾는다.
  - $\underline{y_1} = \underline{e_1}'\left(\underline{x} - \underline{\mu}\right)$**이 첫 번째 주성분이다.**

✓ 두번째 주성분

- $\underline{e_2}'\underline{e_2} = 1, \underline{e_1}'\underline{e_2} = 0$ ← 첫번째 주성분과 독립이어서 설명변동이 다르다.
- $\underline{e_2}'\left(\underline{x} - \underline{\mu}\right)$의 분산을 최대화하는 $\underline{e_2}$을 찾는다.
  - $\underline{y_2} = \underline{e_2}'\left(\underline{x} - \underline{\mu}\right)$을 두 번째 주성분이라 한다.

✓ 세번째 주성분

- $\underline{e_3}'\underline{e_3} = 1, \underline{e_1}'\underline{e_3} = 0, \underline{e_2}'\underline{e_3} = 0$이고
- $\underline{e_3}'\left(\underline{x} - \underline{\mu}\right)$의 분산을 최대화하는 $\underline{e_3}$을 찾는다.
  - $\underline{y_3} = \underline{e_3}'\left(\underline{x} - \underline{\mu}\right)$을 세 번째 주성분이라 한다.

✓ 이를 변수의 수 만큼(p 번) 반복하여 주성분들을 구한다.

- 주성분은 변수의 수 만큼 존재하며, 각 주성분은 서로 독립이다.
- $\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_p \end{pmatrix} = \begin{pmatrix} \underline{e_1}' \\ \underline{e_2}' \\ \cdots \\ \underline{e_p}' \end{pmatrix} \left(\underline{x} - \underline{\mu}\right)_{p\times 1}$

✓ 주성분 계수벡터 $\underline{e_i}$는 고유벡터이다.

- **주성분** y_i**의 분산은 고유치** λ_i**와 같다.**
- tr(∑)은 원 변수 x_i의 변동(분산)의 합이다. ➔ 따라서 **tr(∑)=** $\sum_{i=1}^{p}\lambda_i$
- 주성분 y_k의 원 변수들의 전체 변동에 대한 설명력은 $\frac{\lambda_k}{\sum_{i=1}^{p}\lambda_i}$이다.

✓ 주성분의 추정

- 모집단의 분산, 공분산 행렬 ∑는 알 수 없으며, 이에 대한 추정치로 표본에 대한 분산, 공분산 행렬 S 를 사용한다.
- 즉, S 의 고유치와 고유벡터를 구해 이를 주성분의 추정치로 사용한다.

2-3 주성분 점수와 이름

■ **주성분 점수**

✓ 각 측정치(개체)에 대한 주성분 점수를 계산하고, 개체를 좌표화할 수 있다.

- 이를 통해 이상치를 발견할 수 있다.

✓ **r 번째 측정치에 대한 j 번째 주성분의 점수**(주성분의 계수)

- $y_{rj} = \underline{e_j}'\left(\underline{x_r} - \underline{\mu}\right)$**,** $\underline{x_r}$**는 r 번째 측정치(조사행렬의 r 번째 행)**

✓ 보통은 주요한 주성분(고유치가 1 이상인)에 대한 점수만을 사용하여,

- 이를 서열화 하여, 만족도가 가장 높은 측정치를 선택한다.
- 여기서 측정치는 결국 특정 응답자 또는 지원자가 된다.

■ **주성분 부하 벡터**

✓ **성분 부하 벡터**의 정의

- $\underline{c}_j = \sqrt{\lambda_j}\underline{e}_j$
- 주성분의 이름을 부여하는데 사용할 수 있다.
- 성분 부하 값이 크다는 것은 원 변수에 대한 영향력이 크다는 의미이다.
- 성분 부하 값이 큰 변수를 고려하여 주성분의 이름을 부여하면 된다.
  - 사실 $\sqrt{\lambda_j}$과 무관하게 $\underline{e}_j$만으로도 어떤 원 변수에 대한 변동을 잘 반영하는지 추론할 수 있다.
  - 단지 $\sqrt{\lambda_j}$에 의해 주성분 변수들 간에 어떤 주성분 변수가 전체 변동에 더 큰 영향을 미치는 지 파악이 가능하다.
- 성분 부하 값으로 주성분 이름을 부여할 때 주의할 점
  - 이때 원 변수의 측정단위는 동일(유사)해야 한다.
  - **원 변수의 측정단위가 다른 경우는 공분산 행렬을 이용하지 않고, 상관계수 행렬을 이용하여 고유치, 고유벡터를 구해야 한다.**
  - 부하의 크기 비교는 하나의 주성분 내에서만 가능하며 주성분간 성분 부하 값을 비교하여 이름을 짓는 것은 의미없다.

2-4 주성분 개수

■ **주성분 개수 줄이기**

✓ 주성분 분석의 목적 중 하나는 변수의 차수를 줄이는데 목적이 있다.

- 주성분의 개수는 원 변수의 수 만큼 존재하지만,
- 단지 서로 독립인 변수들로 변형된 것 뿐이다.
- 원 변수의 변동을 어느 정도 설명하는(보통 80%) 주성분만을 선택하여 2 차 분석을 실시할 수 있다.

✓ 총 변동 설명비율

- 이때 고유치의 크기 순으로하여 고유치의 합의 80%가 될 때까지 선택하는 방법을 사용할 수 있다.
- 특히, **상관계수 행렬을 사용하는 경우** 보통, **고유치의 값이 1 이상**인 주성분만 사용하면 총 변동의 80% 정도를 설명한다.
  - 상관계수 행렬을 사용하면 측정 단위 조정으로 인해 변동에 대한 정보가 축소되는 경향이 있다.
  - 측정단위 차가 크지 않으면 그대로 사용하거나, 단위를 조정하여 공분산 행렬을 사용하는 것이 좋은 방법이다.
- 설명력

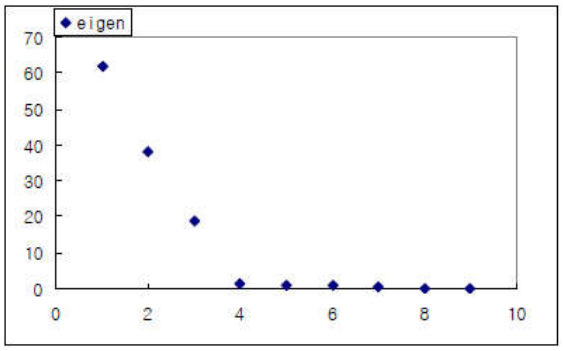
공분산 행렬 ∑	상관 계수 행렬 R

$\sqrt{\mathrm{tr}(\Sigma)} = \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$	λ_i / p (p 는 변수의 개수)

■ SCREE plot 사용

✓ 고유치의 값(y 축)과 고유치의 순서(x 축)으로 그린 SCREE plot

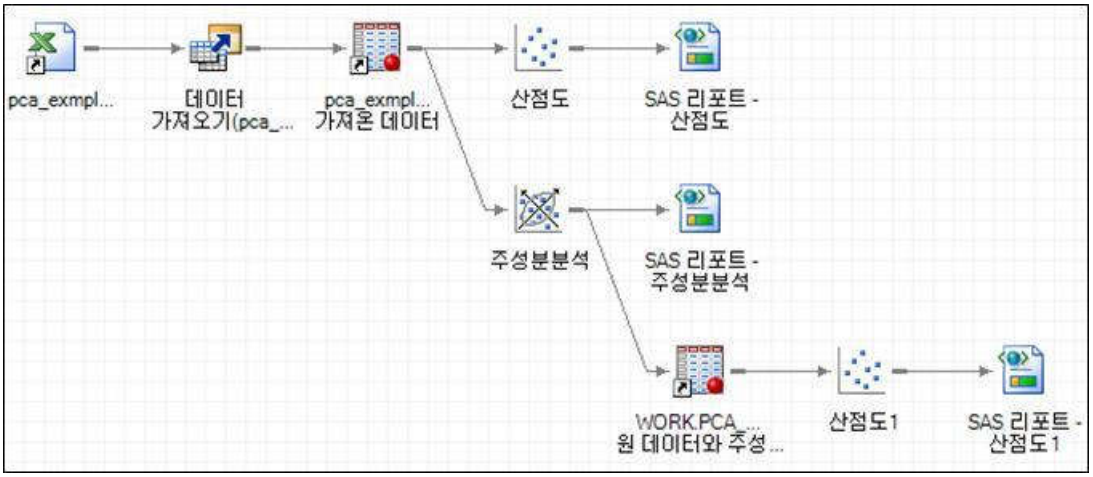
- 갑자기 떨어지면서 0 에 가까워지는 값 이전의 x 값을 개수로 선택한다.
- 그림에서는 주성분의 개수를 3 으로 보면 되겠다.



■ SAS 활용 예제

✓ 변수가 2 개인 경우

- 학생 data 를 통해 다음을 분석해 보자
  - IQ(y 축)과 몸무게(x 축)의 산점도를 그려보자.
  - IQ 와 몸무게의 공분산 행렬을 구하자.
  - 공분산 행렬로부터 고유치를 구하자.
  - 주성분 계산을 위한 고유치, 고유벡터로부터
    ① 주성분 점수(변수) y1 을 y 축, y2 를 x 축으로 한 산점도
    ② 원변수의 산점도와 비교해 보기
- SAS data set 의 생성 및 주성분 분석과 산점도 비교



  - 보통 주성분 분석시 SAS/EG 는 default 로 상관계수행렬을 사용한다.
    ① 따라서, 분석 탭에서 이를 공분산행렬(corr)로 변경시켜줘야 한다.
  - 분석결과를 data set 으로 저장하면, 주성분점수가 기록된다.
    ① 이 주성분점수에 대한 산점도를 그리면 된다.

② 이때, 주성분을 원변수와 비교하기 위해 $y_{rj} = \underline{e_j}' \left(\underline{x_r} - \underline{\mu} \right)$ 대신에 $y_{rj} = \underline{e_j}' \underline{x_r}$을 사용하기도 한다.

✓ 공분산 행렬의 사용

- 회사 지원자 48 명의 능력을 측정한 자료로 주성분 분석을 시행
  - 15 개 항목 중 6 개가 자기공선성이 강하다면, 이를 줄이는 것이 좋다.
    ① 분석의 주목적은 측정 변수를 2~3 개의 주성분 변수로 줄여서
    ② 이 주성분 변수를 통해 우수지원자를 선택함
- 주성분 분석 프로그램

```
DATA WORK.APPLICANT;
     INFILE 'C:₩APPLICANT.TXT'   ;
     INPUT id l ap aa li sc lc ho sm ex dr am gc po kj su;
RUN;
proc princomp data=applicant out=score cov;
   var l--su;
run;
```

  - Cov 는 covariance 의 alias 로 공분산으로 주성분분석을 하게 한다.

The PRINCOMP Procedure

Observations 48
Variables 15

Simple Statistics

	l	ap	aa	li	sc	lc	ho	sm
Mean	6.000000000	7.083333333	7.083333333	6.145833333	6.937500000	6.312500000	8.041666667	4.854166667
StD	2.673749459	1.966023455	1.987549901	2.805690140	2.418072470	3.170047654	2.534513536	3.439381336

	ex	dr	am	gc	po	kj	su
Mean	4.229166667	5.312500000	5.979166667	6.250000000	5.687500000	5.562500000	5.958333333
StD	3.308529167	2.947456531	2.935400827	3.035253854	3.183442871	2.657036023	3.300279378

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	66.5364216	48.3558875	0.5430	0.5430
2	18.1805340	7.5895485	0.1484	0.6914
3	10.5909855	3.8230376	0.0864	0.7778
4	6.7679478	2.7823032	0.0552	0.8330
5	3.9856446	0.3579742	0.0325	0.8656
6	3.6276704	0.7119224	0.0296	0.8952
7	2.9157480	0.0802174	0.0238	0.9190
8	2.8355306	0.8804081	0.0231	0.9421
9	1.9551225	0.3416313	0.0160	0.9581
10	1.6134912	0.4770192	0.0132	0.9712
11	1.1364720	0.2636018	0.0093	0.9805
12	0.8728702	0.1661951	0.0071	0.9876
13	0.7066751	0.1981398	0.0058	0.9934
14	0.5085353	0.2071663	0.0042	0.9975
15	0.3013690		0.0025	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
l	0.149129	0.371461	0.200481	-.277311	0.636939	0.143537	0.054334	0.188839
ap	0.132250	-.029296	0.041918	0.134231	0.042210	0.757039	0.121848	-.269341
aa	0.029611	0.101846	-.131030	0.603168	0.167474	-.002238	0.117237	0.365044
li	0.203126	-.093042	0.619733	0.126399	0.053473	-.018911	-.134527	0.046300
sc	0.231436	-.235740	-.189273	-.072088	-.025117	-.012126	0.247911	-.103262
lc	0.336870	-.195978	-.124714	0.052788	0.231817	-.367360	-.350780	-.318030
ho	0.120238	-.300549	0.447178	0.255587	-.334369	0.058714	0.177661	-.094947
sm	0.379017	-.090010	-.281581	-.172303	-.177778	0.093920	-.064038	-.086513
ex	0.164016	0.636212	0.025043	0.166245	-.191487	-.297426	0.480589	-.388399
dr	0.316050	0.012486	-.113315	-.134844	-.338054	-.118876	0.078867	0.577877
am	0.312106	-.122150	-.244517	-.147307	0.105416	0.226453	0.342601	0.037926
gc	0.338764	-.074347	-.050497	0.206271	0.258316	-.129935	-.121565	-.236845
po	0.357165	-.024920	0.041308	0.317232	0.108875	-.028784	0.023782	0.291193
kj	0.226076	-.044837	0.385206	-.459715	-.026846	-.148771	0.185820	0.030878
su	0.274483	0.470867	0.016815	-.015962	-.349972	0.254634	-.570749	0.003511

	Prin9	Prin10	Prin11	Prin12	Prin13	Prin14	Prin15
l	0.378998	0.224362	0.092656	-.190687	0.001965	-.135682	-.020959
ap	-.033120	-.028399	-.092894	0.086091	0.488649	0.206481	0.015430
aa	-.060954	0.450152	-.199223	0.387155	-.124848	0.099539	0.106257
li	-.517090	0.135003	0.321951	0.069673	0.044180	-.182957	-.311511
sc	0.264633	0.150084	0.477289	0.176181	-.220675	0.431442	-.436694
lc	0.073486	0.160766	0.213320	0.123700	0.357672	0.037881	0.442122
ho	0.539060	0.137682	-.038002	-.139078	-.155159	-.256097	0.232601
sm	-.241327	0.543114	-.297876	-.458581	-.115634	-.094137	-.113068
ex	-.087106	0.006795	0.072428	-.064734	0.122574	-.031972	-.001808
dr	0.162946	-.080129	0.015671	0.113254	0.544965	-.180792	-.156981
am	-.242374	-.237856	0.204549	0.237351	-.309290	-.448256	0.338946
gc	0.168305	-.356813	-.470429	0.160937	-.086183	-.226942	-.467283
po	-.101284	-.414023	0.055047	-.524681	-.106388	0.411102	0.176385
kj	-.103965	0.010538	-.436794	0.318139	-.095263	0.416567	0.208082
su	0.132791	-.061706	0.116204	0.223043	-.306782	0.069812	0.075094

- 결과의 해석과 지원자의 선발
  - **처음 4 개의 주성분만으로도 전체 변동의 83%를 설명함**
  - 처음 4 개의 주성분에 의한 값의 합으로 지원자 5 명을 선정

```
data prinscore;
  set score;
  tot4=sum(prin1--prin4);
```

id	tot4
40	13.6249
39	13.2953

```
   keep id tot4;
run;
proc sort data=prinscore;
   by descending tot4 ;
run;
proc print data=prinscore(obs=5) noobs;
run;
```

8	11.1717
10	10.3382
7	9.4973

✓ 상관행렬의 사용

- 측정 변수들의 측정 단위가 차이가 나는 경우, 이로 인한 분산의 크기의 차가 발생할 때 상관계수 행렬로 주성분 분석을 실시한다.
  - 지원자 예제의 경우 모두 리커트 척도를 사용하므로 불필요하다.
  - 가급적 측정단위를 바꾸어서라도 공분산 행렬을 사용하는 것이 좋다.
- 지원자 예제로 상관계수를 이용한 주성분분석을 해보자.

```
proc princomp data=applicant out=score;
   var I--su;
run;
data prinscore;
   set score;
   tot4=sum(prin1--prin4);
   keep id   tot4;
run;
proc sort data=prinscore;
   by descending tot4 ;
run;
proc print data=prinscore(obs=5) noobs;
run;
```

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	7.51379418	5.45749301	0.5009	0.5009
2	2.05630117	0.60048169	0.1371	0.6380
3	1.45581948	0.25792178	0.0971	0.7351
4	1.19789771	0.45874509	0.0799	0.8149

id	tot4
40	13.6249
39	13.2953
8	11.1717
10	10.3382
7	9.4973

- 상관계수 행렬을 사용할 때, proc princomp 에서 별도의 옵션 불필요
- 결과적으로 공분산분석의 결과와 차이가 없다.

2-5 주성분 해석하기

■ 주성분 이름 붙이기

✓ 주성분 변수 벡터 표현

- $\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_p \end{pmatrix} = \begin{pmatrix} \underline{e_1}' \\ \underline{e_2}' \\ \cdots \\ \underline{e_p}' \end{pmatrix} (\underline{x} - \underline{\mu})_{p\times 1}$ ➔ $\begin{pmatrix} \underline{e_1}' \\ \underline{e_2}' \\ \cdots \\ \underline{e_p}' \end{pmatrix} \underline{x} = \mathbf{L}\underline{x}$

- 주성분 변수는 원변수의 선형결합으로 표시되므로 주성분 변수의 이름은 주성분 계수(고유벡터)를 이용해 부여하는 것이 타당하다.
- 지원자 예제에서 공분산 행렬로부터 구한 고유벡터

Eigenvectors

	Prin1	Prin2	Prin3	Prin4
l	0.149129	**0.371461**	0.200481	-.277311
ap	0.132250	-.029296	0.041918	0.134231
aa	0.029611	0.101846	-.131030	**0.603168**
li	0.203126	-.093042	**0.619733**	0.126399
sc	0.231436	-.235740	-.189273	-.072088
lc	**0.336870**	-.195978	-.124714	0.052788
ho	0.120238	-.300549	**0.447178**	0.255587
sm	**0.379017**	-.090010	-.281581	-.172303
ex	0.164016	**0.636212**	0.025043	0.166245
dr	**0.316050**	0.012486	-.113315	-.134844
am	**0.312106**	-.122150	-.244517	-.147307
gc	**0.338764**	-.074347	-.050497	0.206271
po	**0.357165**	-.024920	0.041308	0.317232
kj	0.226076	-.044837	**0.385206**	**-.459715**
su	0.274483	**0.470867**	0.016815	-.015962

✓ 이름 붙이기

- 각 주성분 내에서 계수가 큰 변수들을 묶은 후,
  - 이 변수들이 함께 나타내는 지표를 이용하여 이름을 부여한다.
  - 전체 변동의 80%를 설명하는 4 개의 주성분의 이름을 정하자.
- 1 번째 주성분의 계수 크기에 의해
  - LC(명석), SM(판매능력), DR(돌파력), AM(야망), GC(개념파악), PO(잠재력) 변수의 계수 크기가 크므로
  - 제 1 주성분은 "정신적.지적 능력"으로 명명한다.
- 2 번째 주성분에 의해 EX(경험), SU(적합)로 "경험" 주성분이라 하자.
- 3 번째 주성분에 의해 LI(호감), HO(진실), KJ(사교)로 심성 변수라 하자.
- 4 번째 주성분에 의해 AA(성적), KJ(사교)로ㅇㄴ부터 성적 주성분이라 하자.

■ 주성분 계수 나타내기

✓ 주성분 계수의 크기로 원변수 분류를 위한 **주성분 계수의 산점도 그리기**

- Outstat= 옵션으로 주성분 분석에 의한 통계량 고유벡터(계수)를 저장

```
proc princomp data=applicant out=score outstat=out1 cov;
   var l--su;
run;
proc print data=out1;
run;
```

OBS	_TYPE_	_NAME_	l	ap	aa	li	...
19	SCORE	Prin1	0.1491	0.1323	0.0296	0.2031	...

  - _TYPE_변수의 값이 SCORE 인 관측만이 주성분 계수이다.
- 주성분 계수만 뽑아서 data 를 전치한다.

```
data out11;
   set out1;
   if (_type_="SCORE");
run;
proc transpose data=out11 out=out2;
run;
proc print data=out2;
run;
```

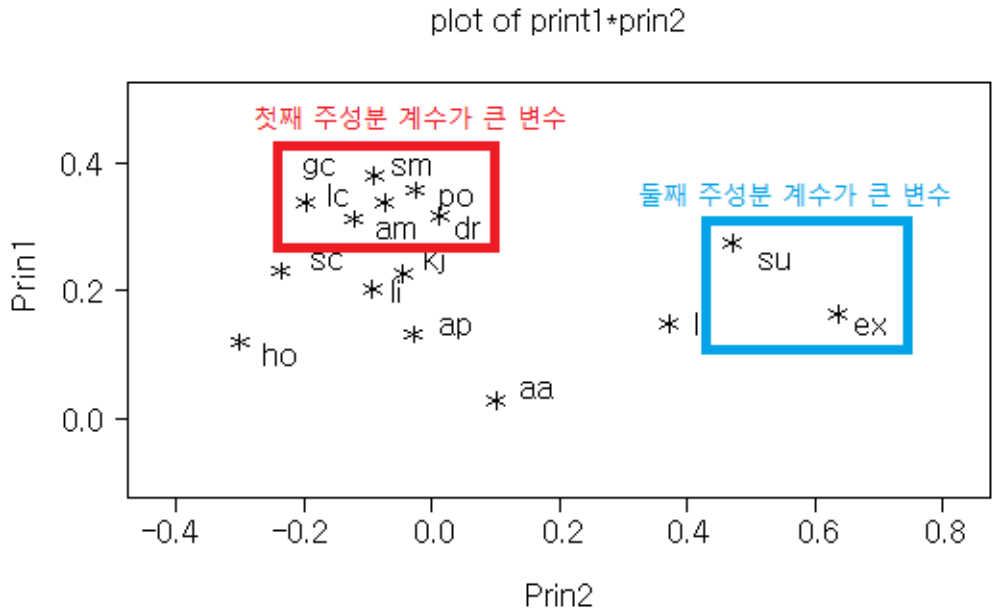
OBS	_NAME_	Prin1	Prin2	Prin3	Prin4	Prin5
1	l	0.14913	0.37146	0.20048	-0.27731	0.63694
2	ap	0.13225	-0.02930	0.04192	0.13423	0.04221

- 주성분 계수는 부하값과 동일하다.

✓ 이제 각 주성분 변수에 대해 산점도를 그리면 된다.

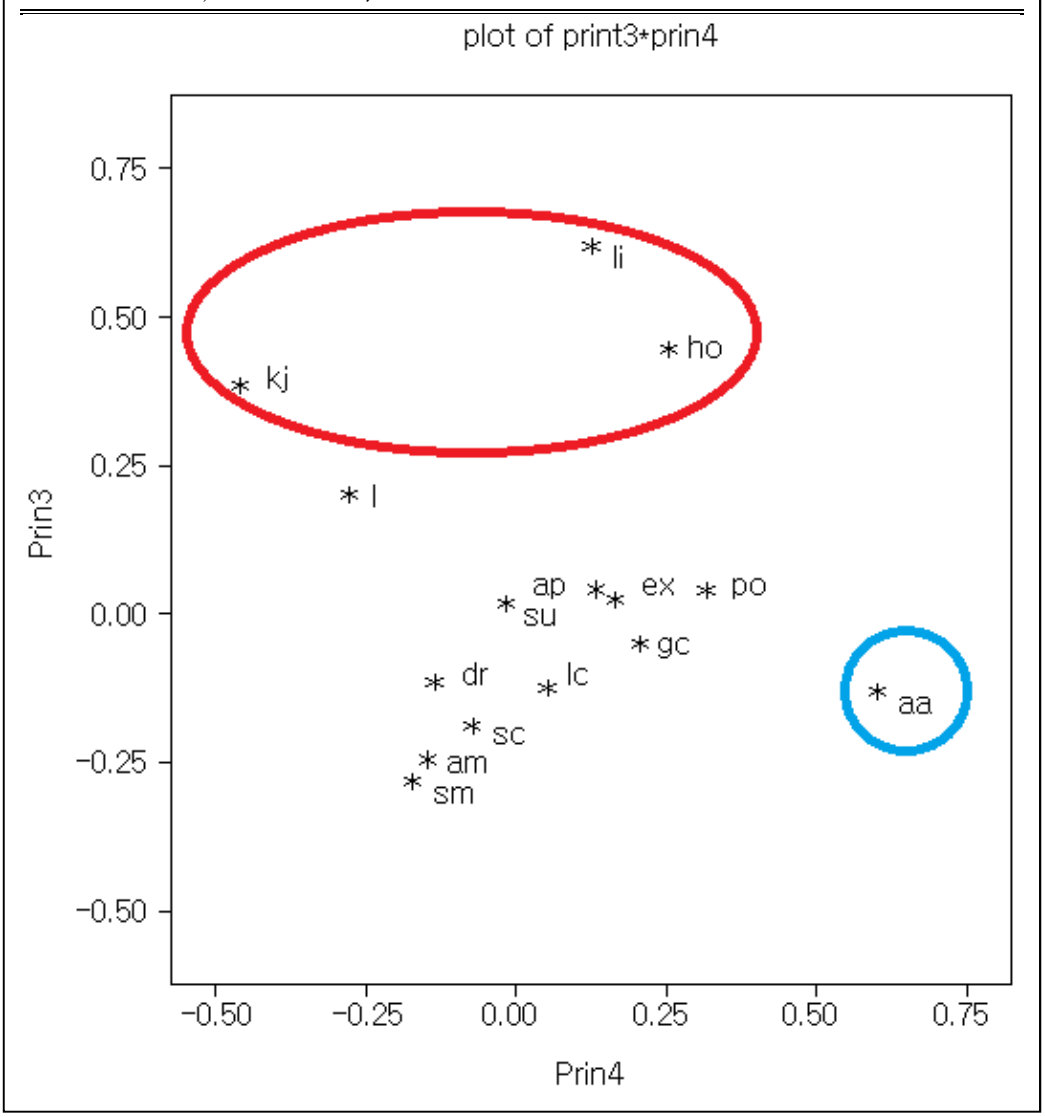
▪ 주성분 1 과 주성분 2 를 비교한 산점도를 그린다.

```
title "plot of print1*prin2";
%PLOTIT(data=out2, labelvar=_NAME_, plotvars=prin1 prin2,
   color=black, colors=blue);
```



▪ 주성분 3 과 주성분 4 를 비교한 산점도를 그린다.

```
title "plot of print3*prin4";
%PLOTIT(data=out2, labelvar=_NAME_, plotvars=prin3 prin4,
   color=black, colors=blue);
```



- 각 주성분에 대한 부하값이 큰 변수들로 주성분을 명명한다.

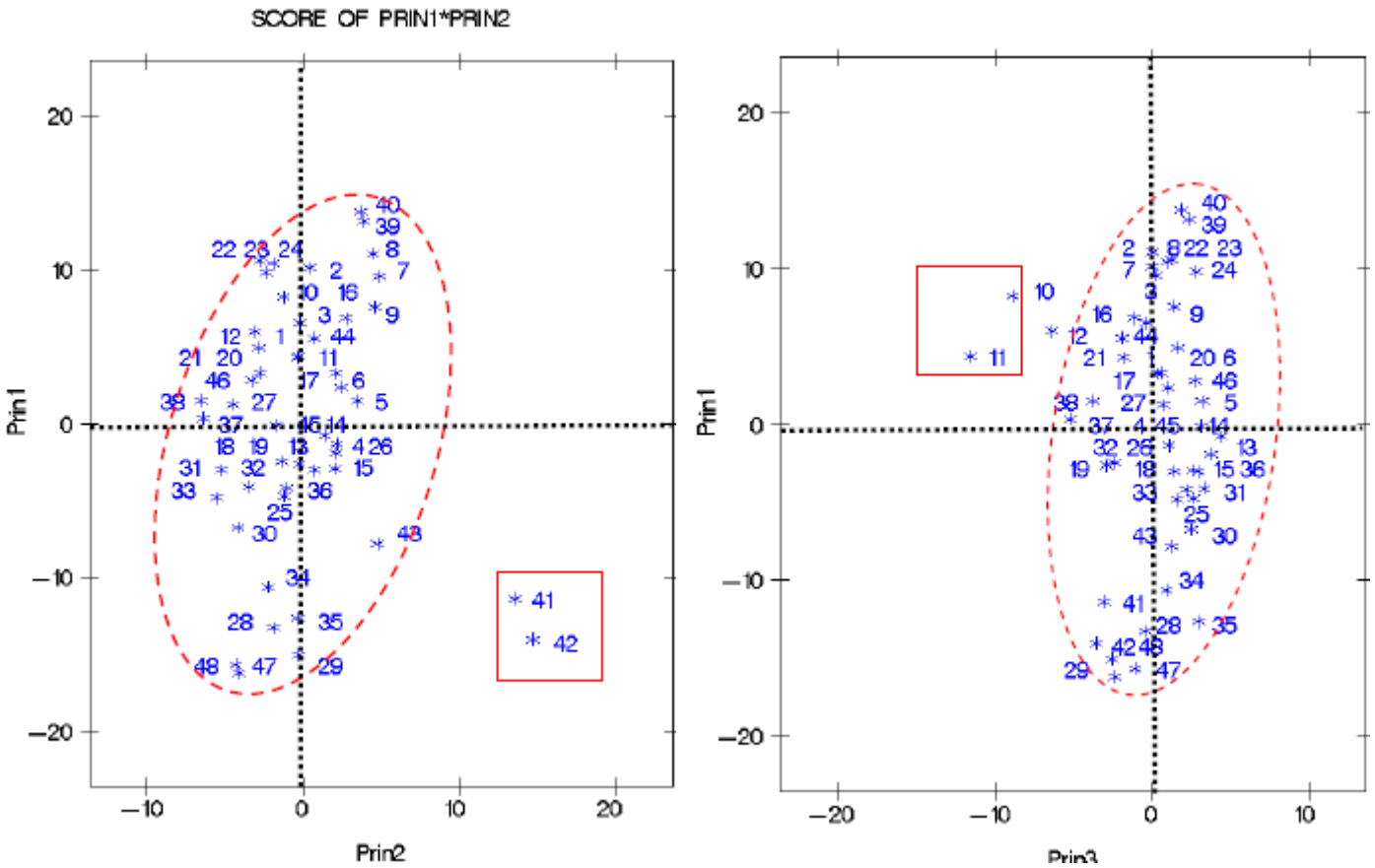
■ 주성분 점수 이용하기

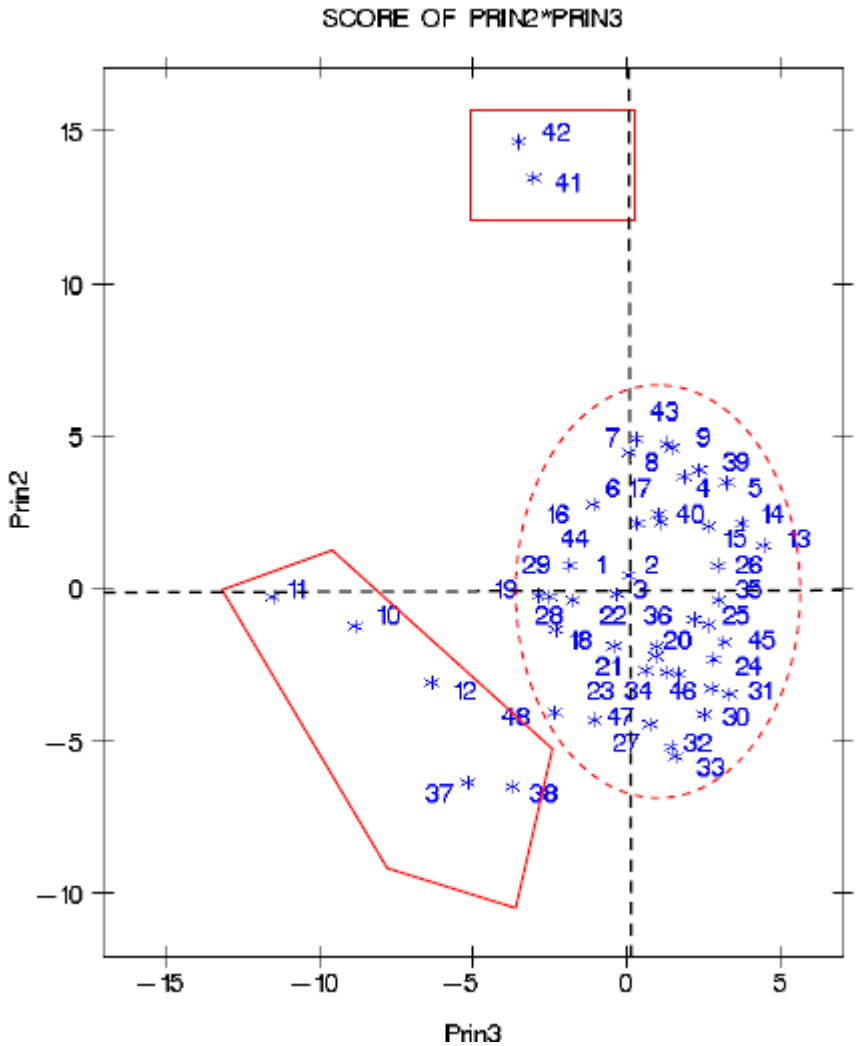
✓ 이상치 발견하기

- 주성분 점수를 이용하여 산점도를 그리면
  - 점의 분포는 타원의 형태를 가지며 타원의 길이는 고유치에 비례한다.
  - 이때 타원으로부터 떨어진 개체가 이상치가 된다.
- PLOTIT 매크로로 산점도를 그려보자.

```
title "SCORE of print1*prin2";
%PLOTIT(data=SCORE, labelvar=ID, plotvars=prin1 prin2,
   color=black, colors=blue);
title "SCORE of print1*prin3";
%PLOTIT(data=SCORE, labelvar=ID, plotvars=prin1 prin3,
   color=black, colors=blue);
title "SCORE of print2*prin3";
%PLOTIT(data=SCORE, labelvar=ID, plotvars=prin2 prin3,
   color=black, colors=blue);
```

- 타원에서 벗어난 점(관측치의 번호)이 이상치이다.





- (10, 11, 12, 37, 38) 지원자는 순한(mild) 이상치이고
  ① 다른 지원자에 비해 심성이 떨어진다.

- (41, 42) 지원자는 극심한 이상치이다.
  ① 이 두 지원자는 경험치가 매우 높다.

✓ 개체 분류
- 어떤 주성분을 위주로 하여 뽑느냐에 따라 합격자가 다르다.
- 그러나 1~4 개의 주성분의 합계로 뽑는 것이 가정 적당하다.

2-6 주성분 분석의 활용

■ 주성분 분석 요약

✓ 주성분을 이용해 변수의 개수를 줄이는 분석방법
- 주성분변수는 원변수의 선형결합으로 이뤄진다. $\underline{y} = L\underline{x}$
- 원 변수가 가지는 정보는 어떻게 표현하는가?
  - 정보는 변동으로 표현되므로 공분산 행렬을 이용한다.
- 계수 행렬(L)은 어떻게 구하는가?
  - 공분산 행렬로 고유치를 구하고
  - 크기순으로 정렬하고(고유치의 크기가 주성분의 설명력)
  - 각 고유치에 대해 $\underline{e}_i'\underline{e}_i = 1, \underline{e}_i'\underline{e}_j = 0$을 만족하는 고유벡터를 구하여 이를 계수로 사용한다.
- 주성분 변수의 개수를 결정
  - 총 변동의 80% 정도를 설명하는 고유치까지 선택한다.
  - 설명력

공분산 행렬 ∑	상관 계수 행렬 R
$\sqrt{\mathrm{tr}(\sum)} = \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$	λ_i/p (p 는 변수의 개수)

■ 일변량 분석

✓ 원변수들이 다변량 정규분포이면, 주성분 변수는 일변량 정규분포를 따른다.
- 또한, 주성분 변수가 정규 분포이면 원 변수는 다변량 정규분포이다.
- 주성분 변수의 정규성으로 원변수의 정규성을 확인할 수 있다.
  - 정규분포를 따르는지 확인(예: Shapiro-wilk W statistics)
- 이상치 존재 여부: 상자-수염 그림(Box-whisker plot)

✓ 4 개의 주성분 변수에 대한 일변량 분석과 결과 해석

```
proc univariate data=score normal plot;
   var prin1-prin4;
run;
```

- 먼저 W-통계량 결과를 살펴보자

정규성 검정

Shapiro-Wilk	----통계량----		-------p-값-------	
PRIN1	W	0.962353	Pr < W	0.1257 ← normal distribution
PRIN2	W	0.897215	Pr < W	0.0005
PRIN3	W	0.87729	Pr < W	0.0001
PRIN4	W	0.972061	Pr < W	0.3041 ← normal distribution

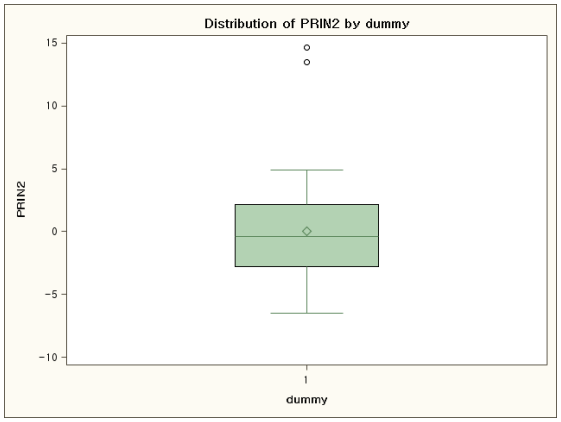
- Shaprio-Wilk 검정: **p>0.05 보다 크면 귀무가설(정규분포)를 채택**
  - ① 임의표본이 정규분포로 나왔는지 검증하는 W 통계치를 추정
  - ② 다른 적합도 검정과 비교시 잘 활용되는 방법
  - ③ sample 수가 5 개 이상, 2000 개 이하일 때 사용
  - ④ 가장 정확도가 높다가 알려져 있음
- kolmogorov-smirnov 검정
  - ① 표본의 누적확률분포와 가설로 설정된 모분포의 누적 확률분포를 상호비교하여 적합도를 검정
  - ② 자료 분포에 대한 어떠한 가정도 가지 않음(비 모수 검정)
  - ③ sample 수가 2000 개 이상일 때
  - ④ 보통 다른 검정법에 의해 정확도가 떨어져 잘 사용하지 않음
- 기타: Q-Q plot
  - ① 정규분포의 분위수와 이에 대응하는 자료 분포의 분위수를 좌표평면에서 각각 수평축과 수직축의 좌표로 표시한 것

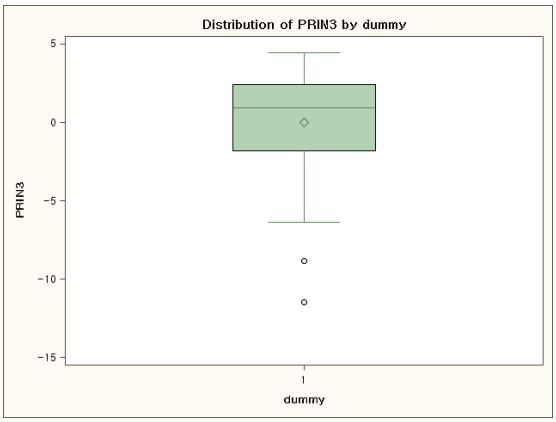
· 주성분 1 과 4 는 정규분포를 따르지만, 2 와 3 은 정규분포가 아니다.

· 따라서 원변수들은 다변량 정규분포를 따르지 않을 것이다.

✓ 박스-수염 그림을 통해 이상치를 확인해보자.

· SAS/EG>작업>기술>요약통계량 or 분포분석 또는 proc univariate





· 이상치는 41, 42 관측치의 경험치에 의한 것임을 확인할 수 있다.

· 이 이상치에 의해 PRIN2, PRIN3 는 왜도가 발생하게 된다.

■ 회귀분석에 이용

✓ 다중회귀모형에서 설명변수간 다중공선성이 있는 경우

· 즉, 설명변수간의 상관관계가 매우 유의하면 $|X'X| \approx 0$이 된다.

- 이때, $(X'X)^{-1} = \frac{1}{|X'X|} adj(X'X)$이므로, 이 값이 매우 커지게 된다.

- 회귀 계수의 추정치 $\hat{\beta} = (X'X)^{-1}X'Y$, 추정치의 분산 $s_{\hat{\beta}}^2 = \text{MSE}(X'X)^{-1}$이 불안해지고 t-검정이나 잔차분석에 의해서도 발견되지 않는다.

✓ 다중 공선성 발견방법

- 산점도 행렬이나 상관계수
  - 두 변수간 상관 관계만 존재할 때는 편리하지만
  - 두 변수와 다른 변수간 상관관계가 존재하는 것은 진단할 수 없다.
- VIF(Variance Inflation Index)와 Condition index 를 이용
  - 대략 10 이상이면 다중 공선성이 있는 것으로 판단

✓ 다중 공선성의 해결방법

- 상관 관계가 높은 변수 제외
- **주성분 분석 이용**
  - 주성분 변수를 설명변수로 사용하는 방법으로
  - 주성분 점수가 새로운 설명변수의 측정치가 된다.
  - 이때의 회귀모형: $Y_j = \beta_0 + \beta_1 Prim1_{1j} + \cdots + \beta_k Prim\text{k}_{kj} + \varepsilon_j$
  - 주성분 분석 활용 조건
    ① 다중 공선성 문제를 발생하는 설명변수가 꼭 필요하거나,
    ② 이 설명변수를 제거하면 변수가 너무 적은 경우
    ③ 주성분 이름을 부여하기 용이하고 제어가 편리한 경우
    ④ 회귀분석에 사용할 때는 모든 변수를 설명변수로 사용하되 변수 선택방법(stepwise, backward, forward)에 의해 선택하면 된다.
  - 회귀분석의 주목적이 예측치를 구하는데 있지 않다면 다중 공선성 문제해결로 주성분을 이용하는 것은 적절하지 않다.
- 능형 회귀
  - 다중 공선성이 회귀계수의 분산을 증가시키므로 불편성을 포기하고
  - MSE(Mean Square of Error)를 최소화하는 편기 추정량으로 계수 추정

2-7 Exercise

■ 1990 년 미 해군 학사장교 관사의 소요인력 추정을 위해 조사: navy.txt

- 25 개 지역(관측치)에 대해 7 개 분야(설명변수)에 대해 조사한 자료
- SITE 는 관측치 번호, MMH 는 종속변수이다.

✓ 다중 공선성 확인방법

- 산점도와 상관계수로 1 차 진단한다.
  - 산점도: 설명변수간의 선형관계가 존재하는가?
  - 상관계수: 0.8 이상이면 두 변수간 공선성을 의심할 만하다.
    ① 유의한 상관관계가 있는 두 변수의 유무에 따른 회귀식에서 회귀계수의 변화가 심하게 나타나면 다중 공선성이 있다.

② 하나의 설명변수가 2 개 이상의 설명변수의 선형 결합으로 표현되어 발생하는 다중공선성 문제는 발견할 수 없다.

③ 두 변수간 상관관계가 유의하면, 종속변수와의 상관관계가 비교적 낮은 변수를 제거거나 해석이 용이한 변수를 남긴다.

- 산점도와 상관계수로는 Xi = a Xk + b 인 두 변수간의 상관관계만 찾을 수 있다.

· 분산팽창지수(VIF)나 상태지수(Condition index) 통계량 이용하기

- 하나의 변수가 여러 변수에 의해 설명되어지는 관계를 찾을 수 있다.
- VIF 의 의미
  ① $VIF_i = \frac{1}{(1-R_i^2)}$ → 그냥 R_i^2을 사용하지 않고 이렇게 하는 의미는?
  ② R_i^2: i 번째 독립변수를 종속변수로 하고 나머지 독립변수로 회귀모형을 구성할 때의 R_i^2값이다.
- 분산팽창지수(Variation Inflation Factor)의 활용
  ① 분산팽창지수가 크다는 것은 하나의 의심되는 독립변수가 다른 설명변수들의 선형 함수로 표현될 수 있다는 것이다.
  ② 이로 인해 다중공선성 문제가 발생한다.
  ③ 일반적으로 분산 팽창지수가 10 이상인 설명변수가 다중공선성 문제를 발생시킨다고 판단한다
- 상태지수의 의미
  ① $(X`X)^{-1}$ 의 대각을 1 로 정규화(상관계수 likely) 한 행렬의 고유치
  ② 어떤 주성분변수의 CI 는 $CI_i = \sqrt{\frac{\lambda_{MAX}}{\lambda_i}}$
  i. 여기서 고유치는 $(X`X)^{-1}$에 대한 고유치이다.
  ③ CI_{MAX}는 X 행렬의 condition number 이라 한다.
  ④ 공선성 문제는 높은 CI 를 가지는 주성분이 2 개 이상의 원변수의 변동에 강하게(proportion of variation 이 0.5 이상) 기여할 때 발생한다.
- 상태지수(condition index)를 활용하기
  ① 고유치는 원변수의 선형결합에 의해 만들어진 주성분 변수의 원변수 변동에 대한 설명력이다.
  ② 고유치가 크다는 것은 주성분의 원 변수 변동에 대한 설명력이 크다는 것을 의미하므로 주성분에 의해 원변수의 차수를 줄일 수 있다는 것을 의미하며,
  ③ 원변수 간에 상관관계(숨어있는 것도 포함)가 높음을 의미한다.
  ④ 매우 작은 고유치가 있다는 것은 설명력이 매우 작은 독립변수가 있을 수도 있지만, 이 독립변수의 대부분을 다른 변수에서 충분히

설명하고 있어, 매우 작은부분만이 독립적으로 기여가 된다는 의미가 될 수 있다.

⑤ Belsley, Kuh 와 Welsch(1980)은 이 값이 10 이 되면, 약한 상관관계가 회귀계수에 영향을 주기 시작하고, 이 값이 100 이 되면, 회귀계수는 평가할 수 없을 정도가 된다.

- 적용 예

```
proc reg data=data_9;
   model ansim deung chae mok apdari woodun seoldo satae yangzi galbi etal total
      = wh24 hh24 bl24 cd24 cw24 rw24 rl24 pw24 hw24 cg24 /vif collin;
run;
```

① VIF 값은 8 이하로서 공선성을 발견하기 어렵다.

② CI 값을 보면, 45 ~ 413 까지 매우 큰 값을 가져 공선성을 의심할 수 있다.

i. Proportion of variation(고유치가 설명한 설명변수의 변동에서 원변수가 기여하는 변동부분)에서 11 번째 주성분에 의해 wh24 와 hh24 의 변동의 94%를 설명하므로 이 변수들은 공선성을 가진다.

ii. 7 번째 주성분에 의해 bl24 에 의한 변동의 84%를 설명하지만, 이 변수에 대해서만 크게 설명하므로 문제가 되지 않는다.

③ 해결방법

i. 이 경우, wh24 와 hh24 둘 중 하나의 변수를 제거해야 하는데, 종속변수에 대한 설명력이 높은 변수를 남기거나

ii. 해석이 용이한 변수를 남긴다.

④ 옵션설명

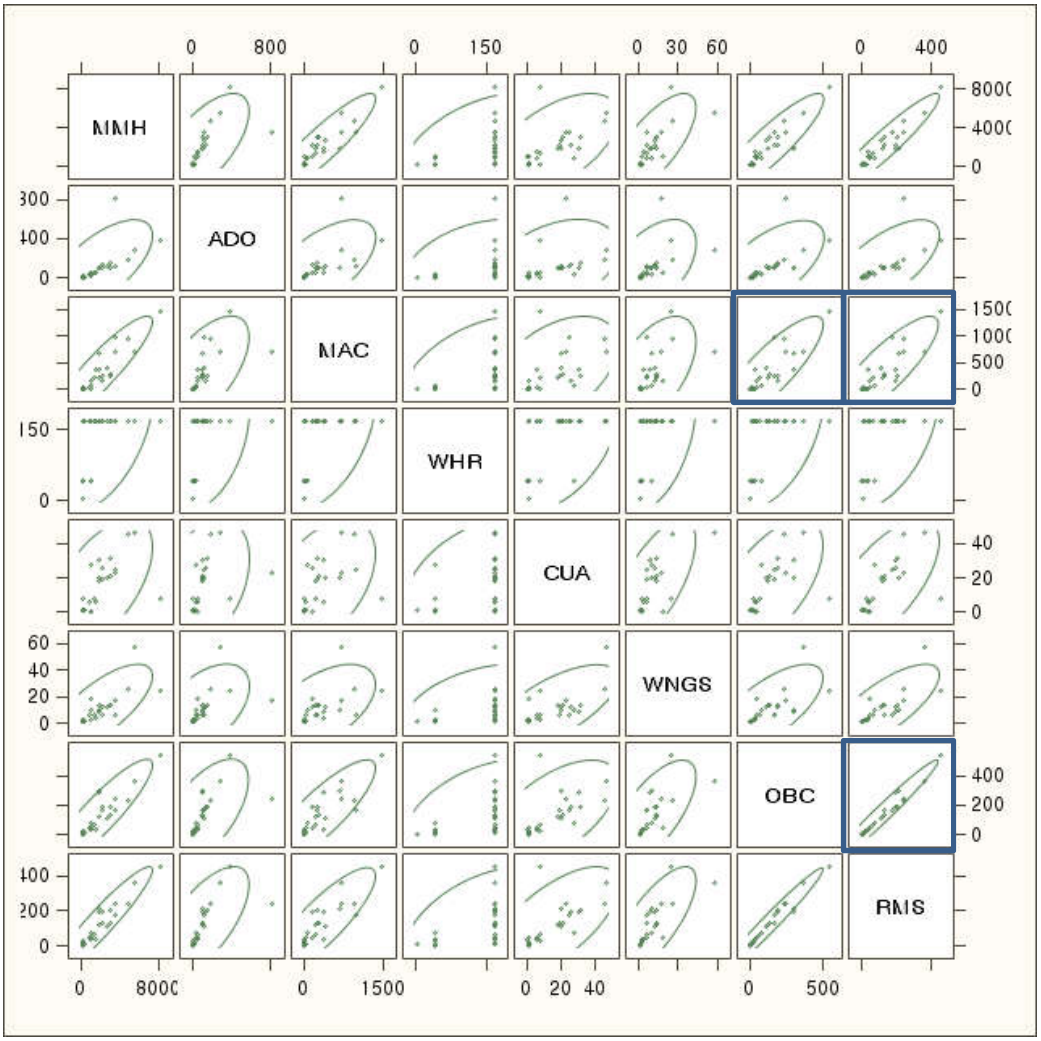
i. Collin 옵션은 collinearity 분석을 실시하도록 한다. 이때 $(X`X)^{-1}$ 을 1 로 대각화한다.

ii. VIF 옵션은 VIF 을 제공하도록 한다.

iii. TOL 옵션은 tolerance 를 제공하며 이 값은 1/VIF 이다.

✓ 7 개 분야를 설명변수로 하고 MMH 를 종속변수로 회귀분석시 다중 공선성이 있음을 보여라.

- 산점도 행렬 ← SAS/EG 로 작성



- 상관계수 행렬

```
data navy;
   infile 'c:₩navy.txt';
   input SITE ADO MAC WHR CUA WNGS OBC RMS MMH;
run;
proc corr data=navy nosimple;
   var ADO MAC WHR CUA WNGS OBC RMS;
run;
```

	ADO	MAC	WHR	CUA	WNGS	OBC	RMS
ADO	1.00000	0.61918 0.0010	0.34730 0.0889	0.38744 0.0557	0.48838 0.0132	0.62004 0.0009	0.67632 0.0002
MAC	0.61918 0.0010	1.00000	0.47139 0.0174	0.47319 0.0169	0.55245 0.0042	0.84953 <.0001	0.86076 <.0001
WHR	0.34730 0.0889	0.47139 0.0174	1.00000	0.38829 0.0551	0.38079 0.0604	0.47278 0.0170	0.49006 0.0129
CUA	0.38744 0.0557	0.47319 0.0169	0.38829 0.0551	1.00000	0.68614 0.0002	0.59383 0.0018	0.66189 0.0003
WNGS	0.48838 0.0132	0.55245 0.0042	0.38079 0.0604	0.68614 0.0002	1.00000	0.67632 0.0002	0.75894 <.0001
OBC	0.62004 0.0009	0.84953 <.0001	0.47278 0.0170	0.59383 0.0018	0.67632 0.0002	1.00000	0.97819 <.0001
RMS	0.67632 0.0002	0.86076 <.0001	0.49006 0.0129	0.66189 0.0003	0.75894 <.0001	0.97819 <.0001	1.00000

- 상관계수로부터 유의수준 0.05 하에

① MAC, OBC, RMS 변수는 모든 변수에 대해 상관관계가 있음

② WNGS 는 WHR 만 빼고 모든 변수에 대해 상관관계가 있음

③ CUA 는 WHR 과 ADO 만 빼고 모든 변수에 대해 상관관계가 있음

④ WHR 은 MAC, OBC, RMS 변수와 상관관계가 있음

⑤ ADO 는 WNGS, MAC, OBC, RMS 변수와 상관관계가 있음

- 설명변수 중 독립된 변수그룹만을 취한다.

① ADO, WHR, CUA 만을 설명 변수로 취한다.

· VIF 와 CI 값의 확인

```
proc reg data=navy;
   model mmh=ADO MAC WHR CUA WNGS OBC RMS /vif collin;
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	134.96790	237.81430	0.57	0.5778	0
ADO	1	-1.28377	0.80469	-1.60	0.1291	2.16276
MAC	1	1.80351	0.51624	3.49	0.0028	4.52397
WHR	1	0.66915	1.84640	0.36	0.7215	1.35735
CUA	1	-21.42263	10.17160	-2.11	0.0504	2.33264
WNGS	1	5.61923	14.74609	0.38	0.7079	3.65318
OBC	1	-14.48025	4.22018	-3.43	0.0032	37.12912
RMS	1	29.32475	6.36590	4.61	0.0003	63.70809

Collinearity Diagnostics

Number	Eigenvalue	Condition Index
1	6.46876	1.00000
2	0.60654	3.26572
3	0.36009	4.23846
4	0.26829	4.91034
5	0.14267	6.73346
6	0.08104	8.93414
7	0.06799	9.75445
8	0.00462	37.41602

Collinearity Diagnostics

	Proportion of Variation							
Number	Intercept	ADO	MAC	WHR	CUA	WNGS	OBC	RMS
1	0.00248	0.00489	0.00244	0.00214	0.00354	0.00278	0.00028046	0.00016059
2	0.09052	0.10388	0.01940	0.03919	0.00908	0.00075907	0.00090277	0.00043043
3	0.03785	0.23481	0.00762	0.02152	0.11895	0.11983	0.00009296	0.00015412
4	0.00037995	0.46835	0.13132	0.00121	0.04674	0.04596	0.00554	0.00093471
5	0.01653	0.00952	0.00634	0.00415	0.68837	0.44134	3.242197E-8	0.00002855
6	0.12221	0.00001004	0.52582	0.15120	0.01847	0.07182	0.05057	0.00988
7	0.72932	0.00005271	0.22011	0.78057	0.00895	0.01257	0.00844	0.00126
8	0.00070543	0.17849	0.08696	0.00002103	0.10590	0.30494	0.93418	0.98715

· 결론적으로 OBC, RMS 변수가 공선성이 있는 것으로 판단된다.

✓ 다중 공선성 문제를 일으키는 설명변수를 제외하고 회귀분석을 실시

· ADO, WHR, CUA 만을 설명변수로 하고 회귀분석을 실시

· 어떤 설명변수가 가장 영향을 많이 미치는지 알아보자(표준화 회귀계수)

✓