Chapter 10. 다변량 자료의 탐색적 분석

김남형 응용통계학과

가천대학교

nhkim@gachon.ac.kr

- 다변량 자료의 탐색적 분석

실제 우리 주변에서 하나의 현상이나 자료를 가지고 설명하기 힘든 경우가 많음

다변량 자료 : 3가지 이상의 변수를 가지고 있는 자료로 측정대상으로부터 여러 개의 변수들을 측정하여 구하는 자료

다변량 자료 분석

1) 다변량 사이의 상관관계를 보려는 경우

2) 어느 한 변량에 대하여 영향을 미치는 다른 여러 변량들과의 회귀관계를 보려는 경우

- 산점도 행렬

여러 변량을 동시에 분석하는 것은 결코 쉬운 일은 아니다.

다변량 자료를 위한 그래픽 기법으로

체르노프 얼굴(Chernoff's faces), 앤드류스 곡선(Andrews curves), 별 그림(star plot)

이 있지만 그다지 효율적이지 못하다.

p ⩾3 변량의 자료에 일반적으로 적용 가능한 그래프는 산점도 행렬(scatter plot matrix)

산점도 행렬과 브러싱(brushing)에 대하여 설명해 보자

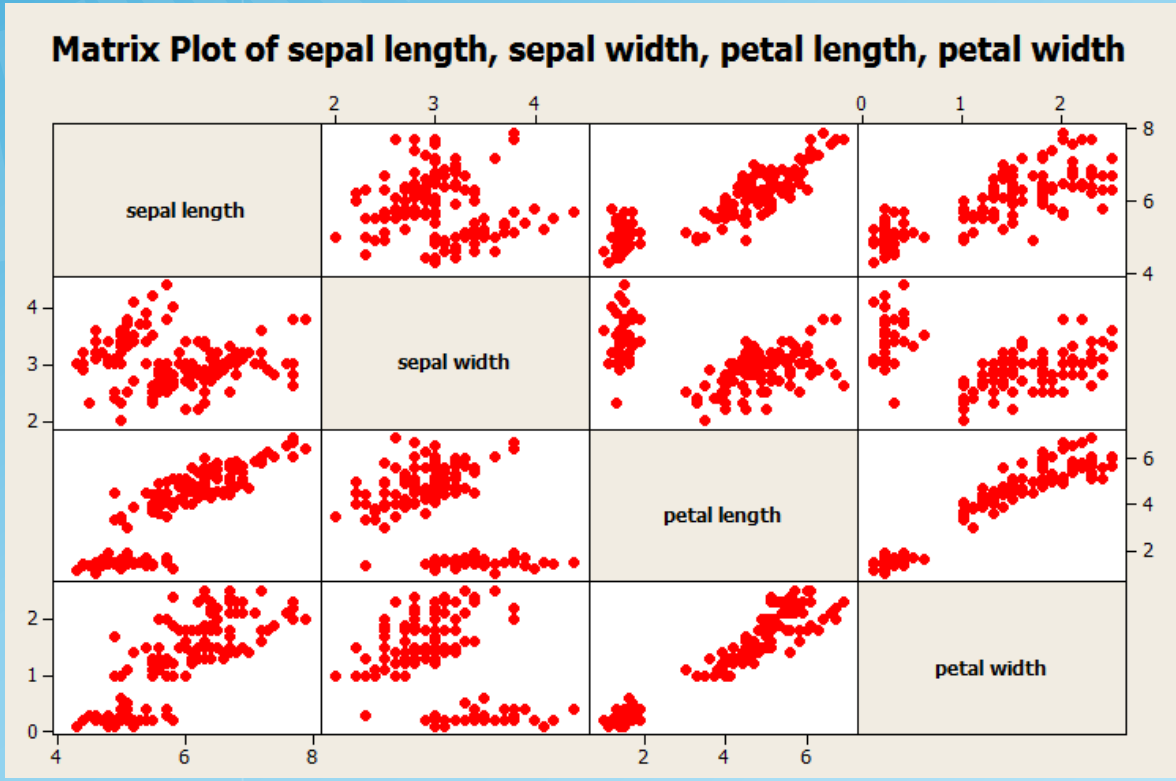
변수 $X_1, X_2, \cdots, X_p$ 에 대하여 모든 쌍의 산점도

$$\begin{matrix} (X_1, X_1)(X_1, X_2)\cdots(X_1, X_p) \\ (X_2, X_1)(X_2, X_2)\cdots(X_2, X_p) \\ \equiv \quad\quad \equiv \quad\quad \equiv \\ (X_p, X_1)(X_p, X_2)\cdots(X_p, X_p) \end{matrix}$$

행렬형태로 나타낸다

여기서 대각선 상에 놓이는 $(X_1, X_1)(X_2, X_2)\cdots(X_p, X_p)$ 의 그래프는 산점도적 의미가 없으므로 그리지 않을 수도 있다.

산점도 행렬의 (i, j) 번째 요소는 (X_i, X_j)의 산점도이고 (j,i) 번째 요소는 (X_j, X_i) 의 산점도 이므로 사실 중복된다 $(i \neq j)$. 그러므로 $p \times p$ 산점도 행렬에서 대각선 위의 산점도만 보는 것으로 충분하다.



- 상관계수 행렬

피어슨의 상관계수(K. Pearson) – 저항성이 없는 것이 결점

$$r_p = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{[\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2]^{1/2}}$$

스피어만의 상관계수(Spearman) – 비모수적 상관계수

$$r_s = \frac{\sum_i (u_i - \bar{u})(v_i - \bar{v})}{[\sum_i (u_i - \bar{u})^2 \sum_i (v_i - \bar{v})^2]^{1/2}}$$

u_i 는 $\{x_i; i = 1, \cdots, n\}$ 중에서
x_i 가 차지한 순위
v_i 는 $\{y_i; i = 1, \cdots, n\}$ 중에서
y_i 가 차지한 순위

예를 들어, 변량 1과 변량 2의 자료값이

{(−2,−5), (−3,−4), (0,1), (10,3), (5,5)}, $(\bar{x}, \bar{y})$ = (2,0) 인 경우

피어슨의 상관계수는

$$r_p = \frac{[(-4)(-5)+(-5)(-4)+(-2)(1)+(8)(3)+(3)(5)]}{[\{16+25+4+64+9\}\cdot\{25+16+1+9+25\}]^{1/2}} = 0.8131$$

원 자료를 순위값으로 대치하면

{(2,1), (1,2), (3,3), (5,4), (4,5)}, $(\bar{u}, \bar{v})$ = (3,3)

스피어만의 상관계수

$$r_s = \frac{[(-1)(-2)+(-2)(-1)+(0)(0)+(2)(1)+(1)(2)]}{[\{1+4+0+4+1\}\cdot\{4+1+0+1+4\}]^{1/2}} = 0.8$$

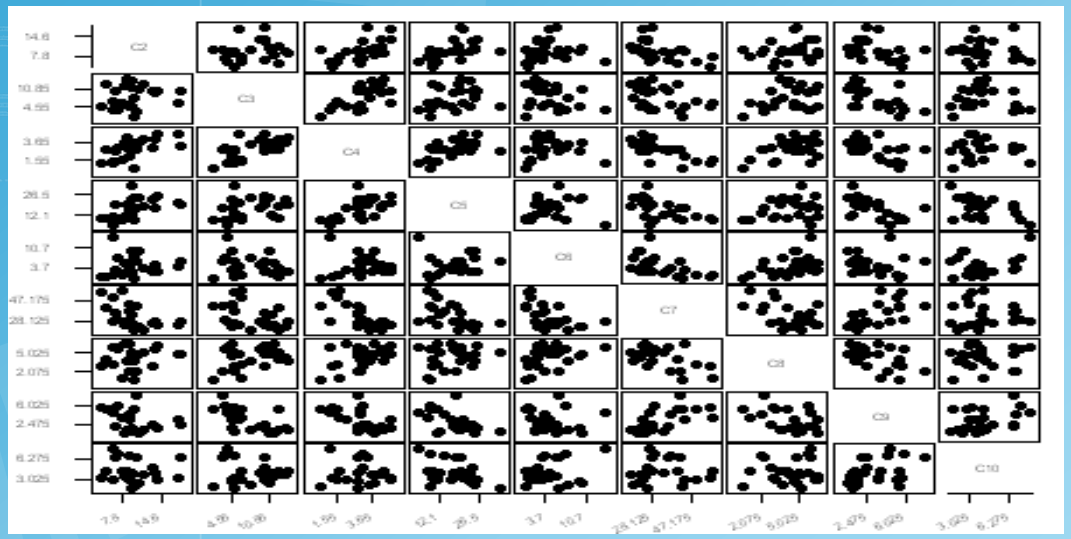
<도표 4> 유럽 25개국의 단백질 섭취 자료

피어슨 상관계수 Stat > Basic Statistics > Correlation...

스피어만 상관계수 먼저, Data > Rank...

모든 자료를 순위화 하여 순위화 한 자료를 이용하여 상관계수를 구한다.

단백질 섭취 자료의 산점도 행렬 Graph > Matrix Plot....



피어슨 상관계수 행렬

	C2	C3	C4	C5	C6	C7	C8	C9
C3	0.153							
C4	0.586	0.620						
C5	0.503	0.281	0.576					
C6	0.061	-0.234	0.066	0.138				
C7	-0.500	-0.414	-0.712	-0.593	-0.524			
C8	0.135	0.314	0.452	0.222	0.404	-0.533		
C9	-0.349	-0.635	-0.560	-0.621	-0.147	0.651	-0.474	
C10	-0.074	-0.061	-0.046	-0.408	0.266	0.047	0.084	0.375

스피어만의 상관계수

	C11	C12	C13	C14	C15	C16	C17	C18
C12	0.097							
C13	0.580	0.600						
C14	0.611	0.325	0.553					
C15	0.239	-0.249	0.212	0.305				
C16	-0.528	-0.411	-0.689	-0.691	-0.547			
C17	0.056	0.167	0.290	0.119	0.396	-0.366		
C18	-0.327	-0.605	-0.505	-0.638	-0.267	0.688	-0.413	
C19	-0.200	0.064	0.028	-0.392	0.081	0.247	0.028	0.372

예를 들어, 변량 1인 쇠고기(C2, C11)와 변량 4인 우유(C5, C14)사이의 피어슨 상관계수는 0.503이고, 스피어만의 상관계수는 0.611이다. 이는 우유가 소로부터 나온다(또는 소가 우유로부터 나온다)는 사실을 생각하면 자연스러운 결과이다. 그러나, 이런 식으로 모든 수치들을 해석하려 든다면 그것 또한 골치 아픈 일.

쉽게 하는 방법이 없을 까?

이를 위하여 고유값-고유벡터 분해(2절)와 상관도(3절)를 학습할 필요가 있다.

- 고유값-고유벡터 분해(eigenvalue-eigenvector decomposition)

양정치 대칭행렬(positive definite symmetric matrix)의 고유값-고유벡터 분해

$p \times p$ 대칭행렬 A 가 양정치라고 하자

이것은, 길이 p 인 임의의 벡터 $u(\neq 0)$ 에 대하여 A 가 $u'Au > 0$ 인 조건 만족

양정치 대칭행렬의 A 의 고유값-고유벡터 분해를 설명해 보자

우선, $p=2$ 인 경우에 대하여 그리고 행렬 A 가 대각행렬인 경우에 대하여

$x'A^{-1}x=1$ 의 기하적 의미를 살펴보자. A 의 역행렬 A^{-1} 은

$$A^{-1} = \begin{pmatrix} a_{11}^{-1} & 0 \\ 0 & a_{22}^{-1} \end{pmatrix}$$

이므로 $x=(x_1, x_2)'$ 으로 표기하면 $x'A^{-1}x=1 \Leftrightarrow x_1^2/a_{11} + x_2^2/a_{22} = 1$ 이 된다.

$a_{11} \succ a_{22}$ 인 경우 이 타원의 장축의 길이가 $2\sqrt{a_{11}}$ 이고 단축의 길이가 $2\sqrt{a_{22}}$ 이며 장축과 단축은 직교한다. 즉, $x'A^{-1}x=1$ 은 a_{11} 과 a_{22} 가 클수록 크기가 큰 타원 방정식이 된다.

타원의 크기는 장축의 길이 l_1 과 단축의 길이 l_2 로 결정 되므로, 마찬가지로 그것들이 2 X2행렬의 크기를 특성화 한다.

만약, 양정치행렬 A 가 대각행렬이 아닌 경우에도 $x'A^{-1}x=1$ 의 자취에 대하여 똑 같이 말할 수 있는가? 답은 '그렇다'

타원의 장축 절반 길이와 단축 절반 길이의 제곱값이 행렬 A 의 고유값이다.

고유벡터란 장축과 단축의 방향을 나타낸다. 타원의 장축과 단축은 회전 후에도 직교하므로 고유벡터끼리의 내적은 0이 된다.

$p \geq 3$ 인 경우로 확장하면 다음과 같다.

◎ $p \times p$ 양정치 대칭행렬 A 의 크기는 p개의 고유값으로 나타내어지며 p 개의 고유값은 타원체 $x'A^{-1}x=1$ 의 p 개 축의 길이와 관련이 있다.

◎ p 개 축의 방향을 결정하는 것은 각기 해당하는 고유벡터이다.고유벡터는 서로 직교

[정의] $Ax=\lambda x(x \neq 0)$ 를 만족하는 실수 $\lambda > 0$ 를 $p \times p$ 양정치 대칭행렬 A의 고유값(eigenvalue)이라고 하고 단위벡터 x를 고유벡터(eigenvector)라고 한다. $(x'x=1)$

$p \times p$ 양정치 대칭행렬 A의 경우 이와 같은 고유값과 고유벡터는 모두 p 개가 있으며 이들을 각각 $\lambda_1, \lambda_2, \cdots, \lambda_p$ 와 $x_1, x_2, \cdots, x_p$ 라고 하면

$$Ax_1 = \lambda_1 x_1, \quad (x_1' x_1 = 1)$$

$$Ax_2 = \lambda_2 x_2, \quad (x_2' x_2 = 1)$$

$$\bullet \quad \bullet \quad \bullet$$

$$Ax_p = \lambda_p x_p, \quad (x_p' x_p = 1)$$

의 관계에 있으며 고유벡터들은 서로 직교하게 된다. 즉 $x_i' x_j = 0 (i \neq j)$ 가 성립

대칭행렬 A 의 고유값-고유벡터 분해

$$AV = VD^2 \quad \text{또는} \quad A = VD^2 V'$$

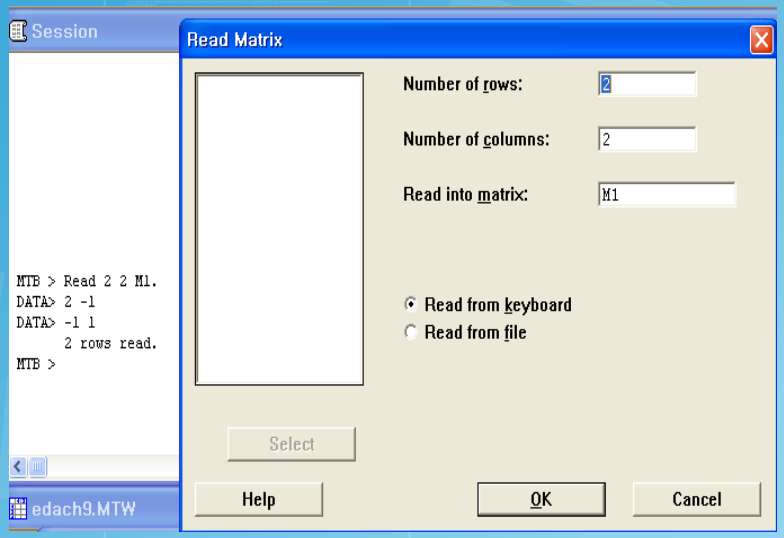
여기서, $D^2 = diag(\lambda_1, \lambda_2, \cdots, \lambda_p), V = (x_1, x_2, \cdots, x_p)$

고유벡터들을 열로하는 행렬 V는 $V'V = VV' = I$ 를 만족하는 직교행렬

미니탭을 이용하여 2X2행렬 $A = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$ 의 고유값-고유벡터 분해를 구하자.

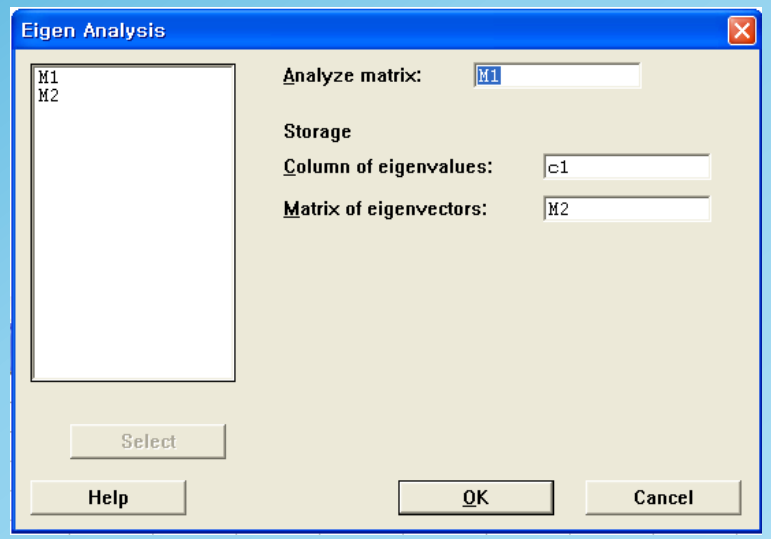
우선 행렬을 만들자

Calc > Matrices > Read...



고유값-고유벡터의 분해

Calc > Matrices > Eigen Analysis...



고유값과 고유벡터

<고유벡터>

```
Matrix M2

 0.850651   0.525731
-0.525731   0.850651
```

<고유값>

```
C1

2.61803    0.38197
```

- 상관도

$p \times p$ 상관계수 행렬을 R 이라고 할 때 이 행렬을 이용한 2차원(저차원) 평면상에 p 개 변량 간의 상관관계를 잘 나타내도록 할 수 없을 까?

Gabriel(1971)에 의하여 개발된 상관도(correlation plot ; 주성분 행렬도 : principle component biplot)방법을 설명

$p \times p$ 양정치 대칭행렬 R 을 고유값-고유벡터 분해식으로 표현

$$R = VD^2V'$$

로 표현이 된다. 여기서 D^2는 R 의 고유값 $d_1^2 \geq \cdots \geq d_p^2 (\geq 0)$ 을 대각원소로 하는 대각행렬이고 V 는 고유벡터를 각 열로 하는 직교행렬이다.

만약, 처음 두 고유값 d_1^2과 d_2^2 이 나머지 고유값들에 비하여 상당히 크다면

$$R = VD^2V' = d_1^2 v_1 v_1' + d_2^2 v_2 v_2' + \cdots + d_p^2 v_p v_p'$$
$$\cong d_1^2 v_1 v_1' + d_2^2 v_2 v_2'$$

로 근사된다.

$VD = H = (h_1, h_2, \cdots, h_p)$ 라고 놓으면 (즉 $h_j = d_j v_j$)

$R = VD^2V' = HH' \cong h_1h_1' + h_2h_2' = H_{(2)}H_{(2)}'$ 이다.

여기서 $p \times 2$행렬 $H_{(2)}$ 는

$$H_{(2)} \equiv (h_1, h_2) \equiv \begin{pmatrix} h_{1(2)} \\ h_{2(2)} \\ \vdots \\ h_{p-1(2)} \\ h_{p(2)} \end{pmatrix}$$

로 정의된다. 따라서, 변량 i 와 j 사이의 상관계수 r_{ij} 는

$r_{ij} \cong h_{i(2)}h_{j(2)}', \quad i, j = 1, \cdots, p$ ⇐ 내적으로 계산(근사)

내적 = 크기 * 크기 * cosine ⇐ 두 벡터 사이의 각이 작을수록 높은 상관관계

미니탭을 이용한 상관도 작성

$p \times 2$ 행렬 $H_{(2)}$ 를 구해보자.

상관계수 행렬이 M1에 입력되어 있다면

① Calc > Matrices > Eigen Analysis.... 명령어에 의하여 M1(=R)의 고유값이 열 C10에, 고유벡터가 행렬 M2(=V)에 만들어져 들어간다.

즉, M1=M2 X diag(C10) X M2' 이 된다.

② 다음으로 미니탭 명령어 Calc > Calculator...

에 의하여 열 C11에 열 C10의 제곱근 값을 저장 ($C11 = \sqrt{C10}$)

③ Calc > Matrices > Diagonal...

에 의하여 열 C11의 각 값을 대각요소로 하는 대각행렬 M3를 만들고

④ Calc > Matrices > Copy...

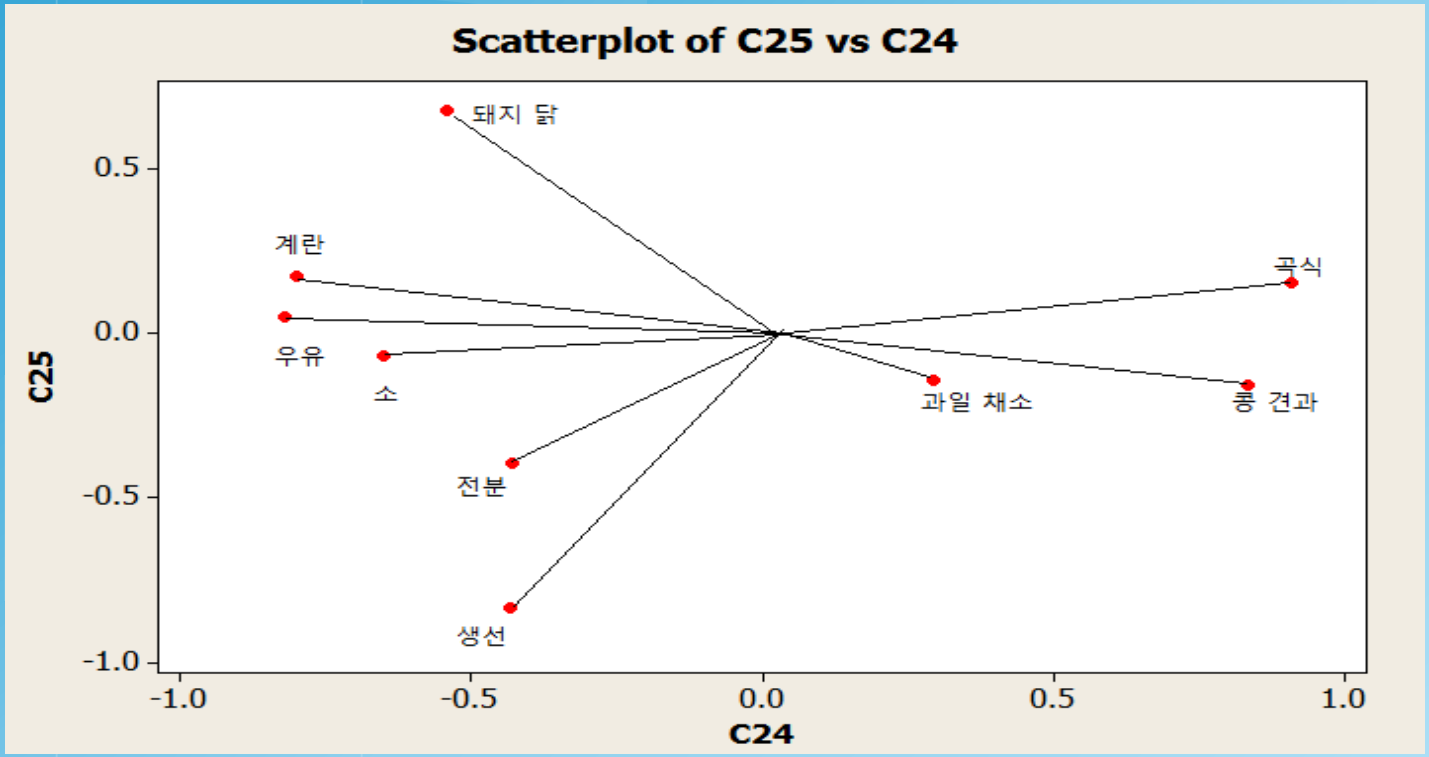
에 의하여 두 행렬 M2, M3를 곱하여 행렬 M4를 만들면 이것이 바로 행렬 H(9X9)

⑤ Calc > Matrices > Copy... 명령어를 사용하여 행렬 M4의 9개 열의 원소를 각각 9개 열 C21, C22, ··· , C28, C29로 복사해서 저장

이에 따라 처음 두 열 C21과 C22에 h_1 과 h_2 를 만들어 넣을 수 있다.

이것을 플롯한 것이 상관도인 <도표 7>이다.

<도표 7> 유럽 25개국 단백질 섭취원(변량)들의 상관도



① 변량 1(쇠고기), 변량 4(우유), 변량 3(계란), 변량 2(돼지, 닭고기) ← 축산업

② 변량 6(곡식), 변량 8(콩, 견과), 변량 9(과일, 채소) ← 밭농사

③ 변량 5(생선), 변량 7(전분) ← 농,어업 혼합형

상관행렬 M1의 고유값중 2개는 (4.01+1.41)/9=60.2%이다.

❑ 저항성 다중 선형회귀

다중회귀(multiple regression)이란

어느 한 변량의 예측 또는 설명을 위하여 이 변량에 영향을 줄 것으로 생각되는 다른 여러 변량들을 이용하는 통계적 분석기법

⇒ 저항성 직선 회귀를 2개 이상의 설명변량(=독립변량)으로 반응변량(=종속변량)을 설명하는 것 (sweeping 방법을 이용한 저항성 다중 선형회귀 적합)

<자료를 이용한 예 : stack loss dataset>

3개의 설명변량 AIR, TEMP, ACID와 1개의 종속변량 LOSS로 구성

이들 변수들은 암모니아(NH_3)를 산화하여 질산(HNO_3)을 만드는 화학 공정에 관한 것

X_1 = AIR = 공기흐름(속도)

X_2 = TEMP = 냉각수 온도

X_3 = ACID = 질산의 농축도

Y = LOSS = 암모니아 비수거분(손실분, %)

자료분석의 목적 및 적합 방법(sweeping)

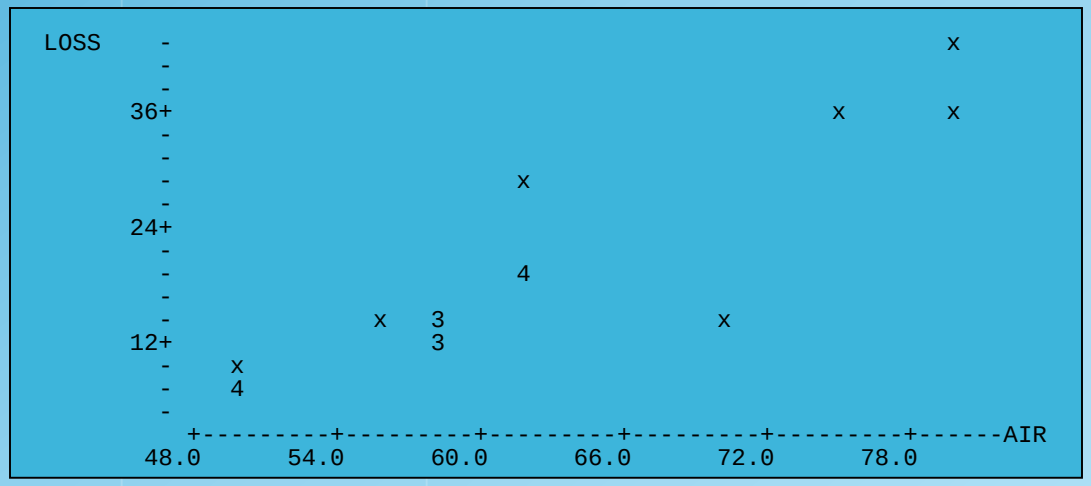
: 변량 LOSS(=Y)가 변량 AIR(=X_1), TEMP(=X_2), ACID(=X_3)에 의하여 어떤 영향을 받는가

=> 모든 설명변수를 동시에 적합시키지 않고 선형관계가 가장 높은 변수 하나씩 적합

=> 각 설명변량과 종속변량 사이의 산점도

edach10-2.MTW ***

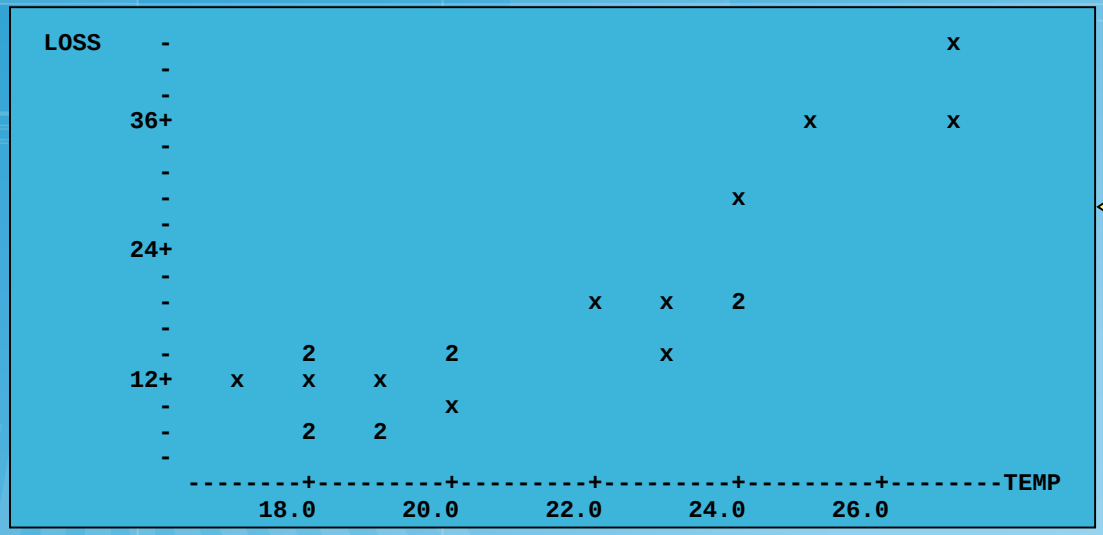
	C1 AIR	C2 TEMP	C3 ACID	C4 LOSS
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15



Y와 X_1과의 저항성 직선식

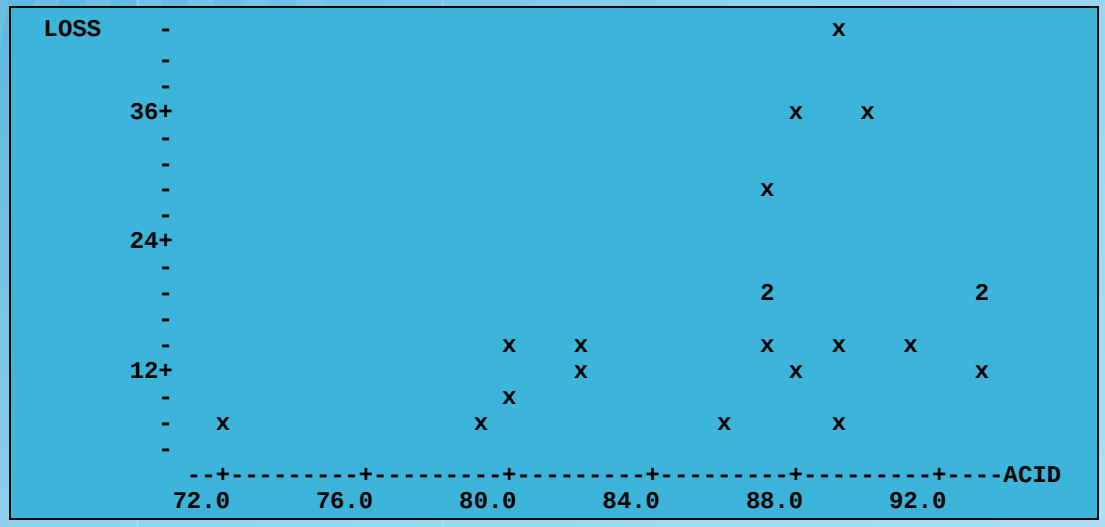
$Y=-40.93+0.9667X_1+Y_{.1}$

$Y_{.1}$: X_1 변수만 적합시키고 남은 잔차



Y와 X_2과의 저항성 직선식

$Y=4.78+0.2778X_2+Y._2$



Y와 X_3과의 저항성 직선식

$Y=48.25+0.625X_3+Y._3$

1. Y와 X_1과의 저항성 직선식 : $Y = -40.93 + 0.9667X_1 + Y_{.1}$

2. $Y_{.1}$ 과 X_2 또는 X_3 와 관계

   - X_2 또는 X_3 에는 X_1 과 중복되는 부분 남아있다
   - 중복되는 부분을 제외한 즉, 순수한 $X_{2.1}$ 또는 $X_{3.1}$ 을 찾는다
   - 저항성 직선을 계산 : $X_2 = 4.78 + 0.2778X_1 + X_{2.1}$

     $X_3 = 48.25 + 0.6250X_1 + X_{3.1}$

3. $Y_{.1}$ 과 $X_{2.1}$, $Y_{.1}$ 과 $X_{3.1}$ 사이의 산점도 확인, 저항성 직선 계산

   : $Y_{.1} = 0.40 + 0.60X_{2.1} + Y_{.12}$, $Y_{.12}$ 는 적합후 남은 잔차

4. $Y_{.12}$ 와 X_3 와의 관계는 단계 2 처럼 순수한 $X_{3.12}$ 를 사용

   : $X_{3.1} = 0.62 - 0.7418X_{2.1} + X_{3.12}$

5. $Y_{.12}$ 와 $X_{3.12}$ 사이의 산점도 확인, 저항성 직선 : $Y_{.12} = a + bX_{3.12} + Y_{.123}$

모든 과정을 정리하면

Y = −40.93 + 0.9667X1 + Y.1 에서

Y.1 = 0.40 + 0.60X2.1 + Y.12 , X2 = 4.78 + 0.2778X1 + X2.1 각각 대입하면

$\underline{Y = -43.53 + 0.80X_1 + 0.60X_2 + Y.12 : \text{최종적인 저항적인 회귀식}}$

, $Y.12$ 는 잔차 **r** , $\hat{Y} = -43.53 + 0.80X_1 + 0.60X_2$

- 최소제곱법에 의한 스위핑(sweeping)

  최소제곱법에 의한 다중선형회귀에서

  A) 단순 선형회귀의 알고리즘을 사용하는 방법(sweeping)

  B) 다중 선형회귀의 표준적인 방법

  첫째 방법 A

  앞의 자료를 이용한 다중 회귀분석을 시도, 저항적인 방법의 결과 동일한 모형을 사용하여

  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ 을 적합하여 보자

  ① i) Y를 X1에 최소제곱 회귀시켜보자. 그 결과 최소제곱 회귀식은

  $Y = -44.132 + 1.0203X_1 + Y_{.1}$

  ii) X2를 X1에 최소제곱 회귀시켜 보자. 그 결과 최소제곱 회귀식은

  $X_2 = 4.807 + 0.26954X_1 + X_{2.1}$

  ② Y.1을 X2.1에 최소제곱 회귀시켜 보자. 그 결과 최소제곱 회귀식은

  $Y_{.1} = 0.000 + 1.2954X_{2.1} + Y_{.12}$

위의 단계별 결과 ① 과 ②를 종합하여 보면

$Y = -44.132 + 1.0203X_1 + Y_{.1}$

$= -44.132 + 1.0203X_1 + (1.2954X_{2.1} + Y_{.12})$

$= -44.132 + 1.0203X_1 + 1.2954(X_2 - 4.807 - 0.26954X_1) + Y_{.12}$

$= -50.36 + 0.6711X_1 + 1.2954X_2 + Y_{.12}$

두번째 방법 B

$$A = \pi r^2$$

이 방법에 따른 $\beta_0, \beta_1, \beta_2$ 의 해는 잘 알려진대로

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \quad \Longrightarrow \quad b = (X'X)^{-1}(X'y)$$

를 최소화 시켜서 얻게 된다.

$\hat{Y} = -50.4 + 0.671X_1 + 1.30X_2$

다중 저항성 회귀에 의한 잔차와 최소제곱 회귀에 의한 잔차 비교

	제곱잔차		절대값 잔차
	중위수	평균	합
다중 저항성 잔차 **r** :	1.96	11.06	44.1
최소제곱 잔차 **e** :	4.48	8.99	49.8

요약

⦿ 스피어만 상관계수 r_s 는 순위(rank)를 이용하여 구하는 데 특이값에 대하여 저항적
반면 피어슨의 상관계수 r_p 는 그렇지 않다.

⦿ 상관계수 행렬의 고유값-고유벡터 분해를 통하여 상관도를 그릴 수 있다.
변량들 사이의 상관계수는 상관도에서 변량점 벡터간의 내적으로 근사
두 변량점이 원점을 중심으로 이루는 사이각이 작은 경우 두 변수량간의 상관이 크다

⦿ 다중 선형회귀에서 저항성이 있는 회귀 적합(sweeping)

① i) 종속변량 Y와 설명변량 X_1, … , X_p의 각각의 산점도를 그린다.
그 중 Y와 가장 강한 선형관계를 갖는 변량 하나 선택
그 변량을 표기의 편의상 X_1이라고 하자

ii) Y와 X_1과의 회귀식을 구하고 잔차 $Y_{.1}$을 구한다
마찬가지로 X_j 를 X_1으로 회귀시키고 잔차 $X_{j.1}$을 구한다. (j=2, ⋯;p)

② i) 종속변량 $Y_{.1}$과 설명변량 $X_{2.1}, \cdots, X_{p.1}$의 각각의 산점도를 그려보고 그 중 $Y_{.1}$과 가장 강한 선형관계를 갖는 변량 하나를 선택한다. 그 변량을 표기의 편의상 $X_{2.1}$이라고 하자.

ii) $Y_{.1}$과 $X_{2.1}$과의 회귀식을 구하고 잔차 $Y_{.12}$를 구한다

마찬가지로 $X_{j.1}$을 $X_{2.1}$으로 회귀하고 잔차 $X_{j.12}$를 구한다.$(j=3,\cdots,p)$

③ i) 종속변량 $Y_{.12}$와 설명변량 $X_{3.12}, \cdots, X_{p.12}$의 각각의 산점도를 그려보고 그 중 $Y_{.12}$와 가장 강한 선형관계를 갖는 변량 하나를 선택한다

그 변량을 표기의 편의상 $X_{3.12}$라고 하자

ii) $Y_{.12}$와 $X_{3.12}$와의 회귀식을 구하고 잔차 $Y_{.123}$를 구한다.

마찬가지로 $X_{j.12}$를 $X_{3.12}$로 회귀하고 잔차 $X_{j.123}$를 구한다$(j=4,\cdots,p)$

④ 이상의 방법을 거듭하여

Y와 X_1과의 회귀식과 잔차 $Y_{.1}$, $Y_{.1}$과 $X_{2.1}$과의 회귀식과 잔차 $Y_{.12}$,

$Y_{.12}$와 $X_{3.12}$와의 회귀식과 잔차 $Y_{.123}$, $\cdots$

을 구할 수 있게 되는데 이것을 종합하여 Y를 $X_1, \cdots, X_p$에 회귀시킨 회귀 적합식과 잔차 $Y_{.12\ \ldots p}$를 계산할 수 있게 된다.