제 1 장 서론

Introduction

목차

1.1 다변량 자료의 표현

1.1.1 자료행렬 (Data Matrix)

1.1.2 그래프적 표현

1.2 자료의 대수적 요약

1.2.1 표본평균 (Sample Mean)

1.2.2 표본분산과 표준편차 (Sample Variance, Standard Deviation)

1.2.3 표본공분산 (Sample Covariance)

1.2.4 표본상관계수 (Sample Correlation Coefficient)

1.2.5 중심화와 표준화

1.1 다변량 자료의 표현

1.1 자료행렬 (Data Matrix)

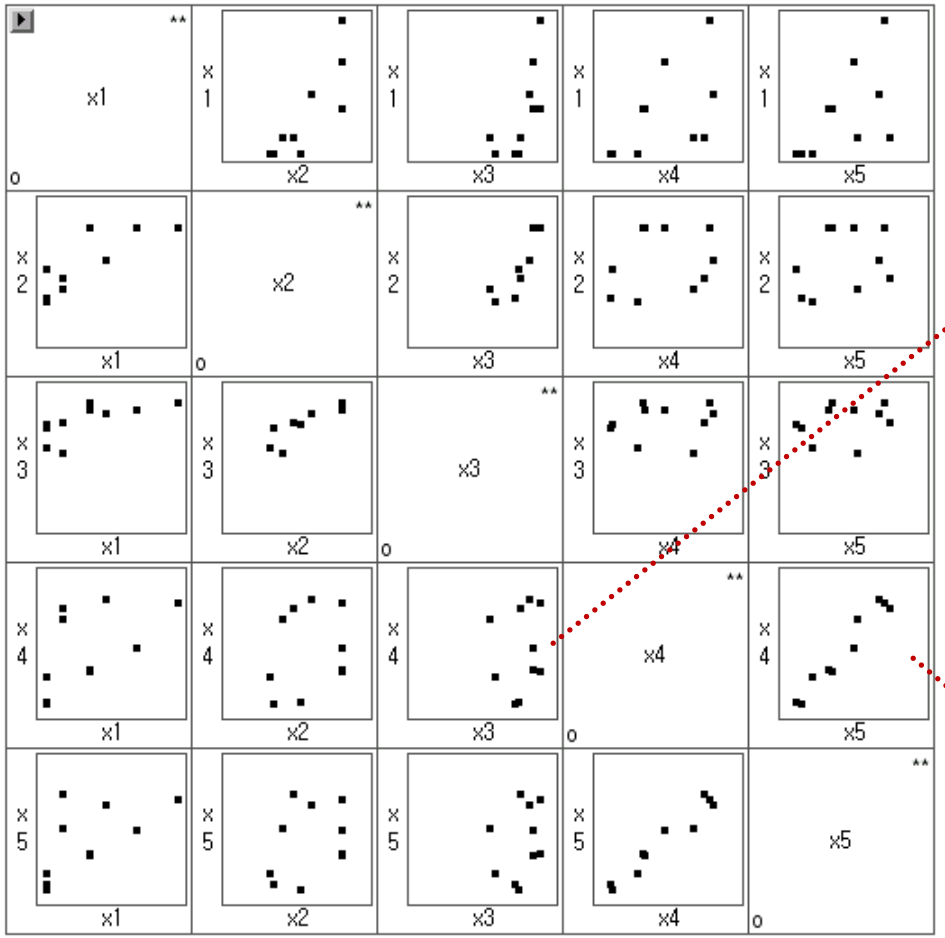
$$\underset{n\times p}{\mathbf{X}} = \begin{array}{c} \\ \text{관} \\ \text{찰} \\ \text{개} \\ \text{체} \end{array} \begin{array}{c} \\ (1) \\ (2) \\ \vdots \\ (i) \\ \vdots \\ (n) \end{array} \overset{\text{반응변수}}{\begin{array}{cccccc} (1) & (2) & \cdots & (j) & \cdots & (p) \end{array}} \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

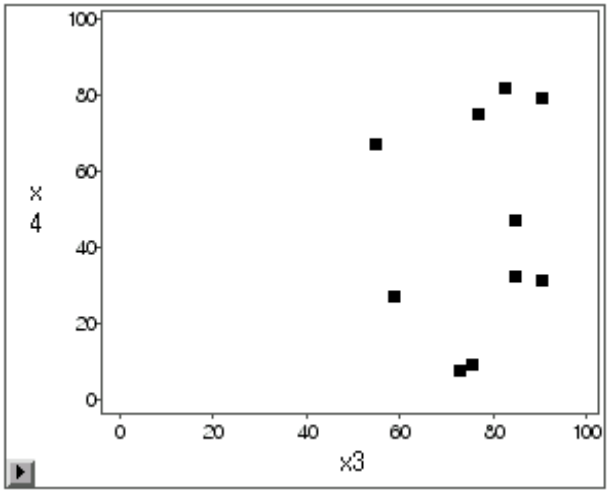
$\mathbf{x}_j$ (열벡터)　　$\mathbf{x}_i'$ (행벡터)

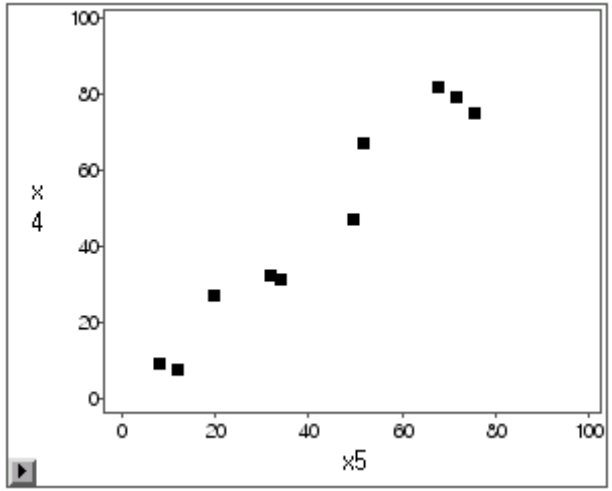
$$\underset{10\times 5}{\mathbf{X}} = \begin{array}{c} \text{국어} \\ \text{영어} \\ \text{제2외국어} \\ \text{수학} \\ \text{과학} \end{array} \begin{pmatrix} 3 & 33 & 73 & 8 & 12 \\ 3 & 30 & 59 & 28 & 20 \\ 35 & 83 & 91 & 32 & 34 \\ 35 & 83 & 85 & 33 & 32 \\ 15 & 40 & 55 & 68 & 52 \\ 3 & 53 & 76 & 10 & 8 \\ 68 & 83 & 85 & 48 & 50 \\ 15 & 47 & 77 & 76 & 76 \\ 46 & 60 & 83 & 83 & 68 \\ 98 & 83 & 91 & 80 & 72 \end{pmatrix}$$

1.1.2 그래프적 표현

- 산점도 (Scatter Plot)



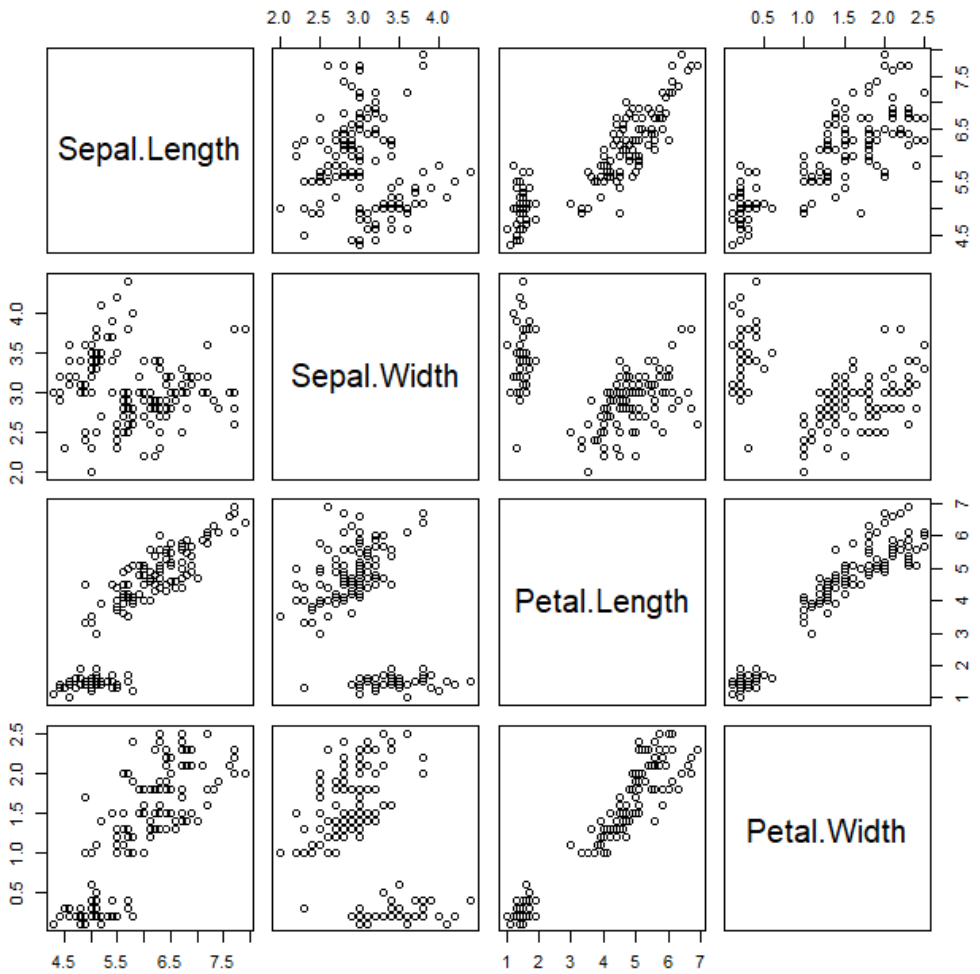




1.1.2 그래프적 표현

- R에 저장되어 있는 데이터를 불러와서 그래프를 그려보기.

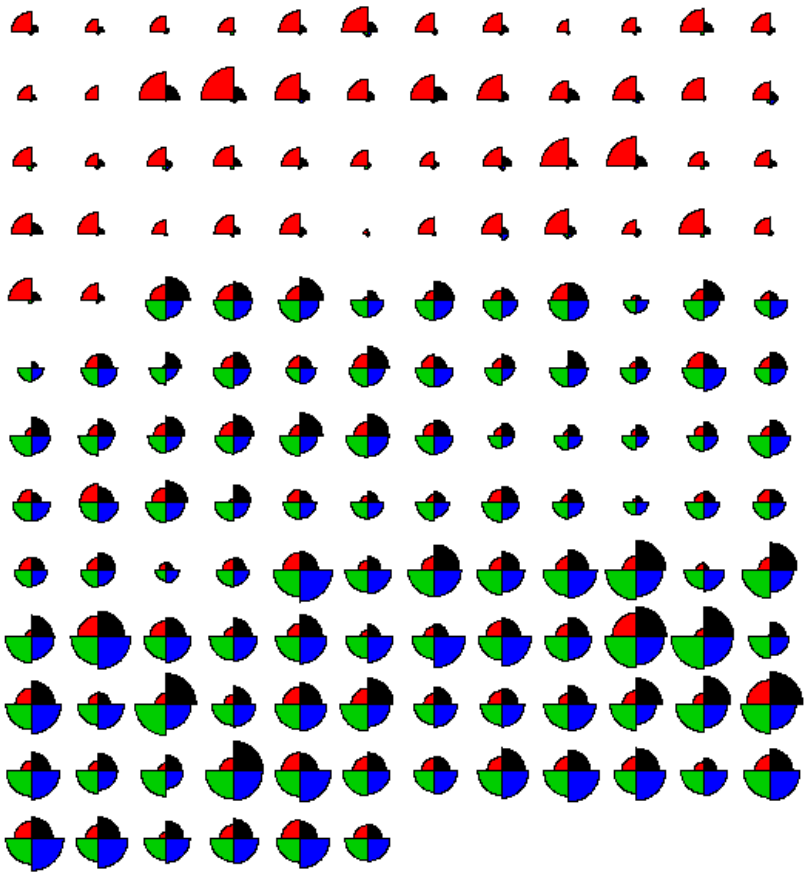
```
> data(iris)
> iris
> pairs(iris[,1:4])
```



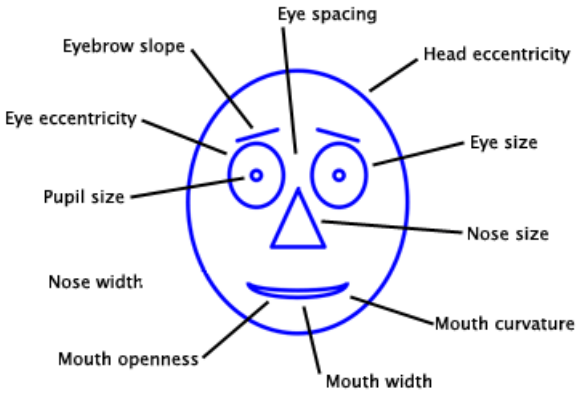

```
library(car)
scatterplotMatrix(iris[,1:4])
```

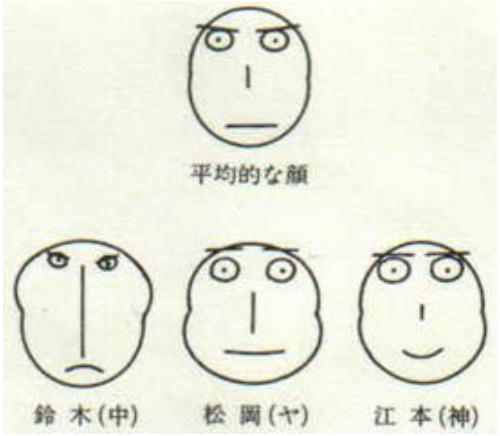
별도표 (Star Chart)

```
> stars(iris[,1:4], draw.segments = T)
```



체르노프의 얼굴 (Chernoff's face)





2005 National League

	PCT	H	HR	BB	SB
ARI	.475	1419	191	606	67
ATL	.556	1453	184	534	92
CHI	.488	1506	194	419	65
CIN	.451	1453	222	611	72
COL	.414	1477	150	509	65
FLO	.512	1499	128	512	96
HOU	.549	1400	161	481	115
LAD	.438	1374	149	541	58
MIL	.500	1413	175	531	79
NYM	.512	1421	175	486	153
PHI	.543	1494	167	639	116
PIT	.414	1445	139	471	73
SDP	.506	1416	130	600	99
SFG	.463	1427	128	431	71
STL	.617	1494	170	534	83
WAS	.500	1367	117	491	45

win pct
face height, smile curve, hair styling

hits
face width, eye height, nose height

home runs
face shape, eye width, nose width

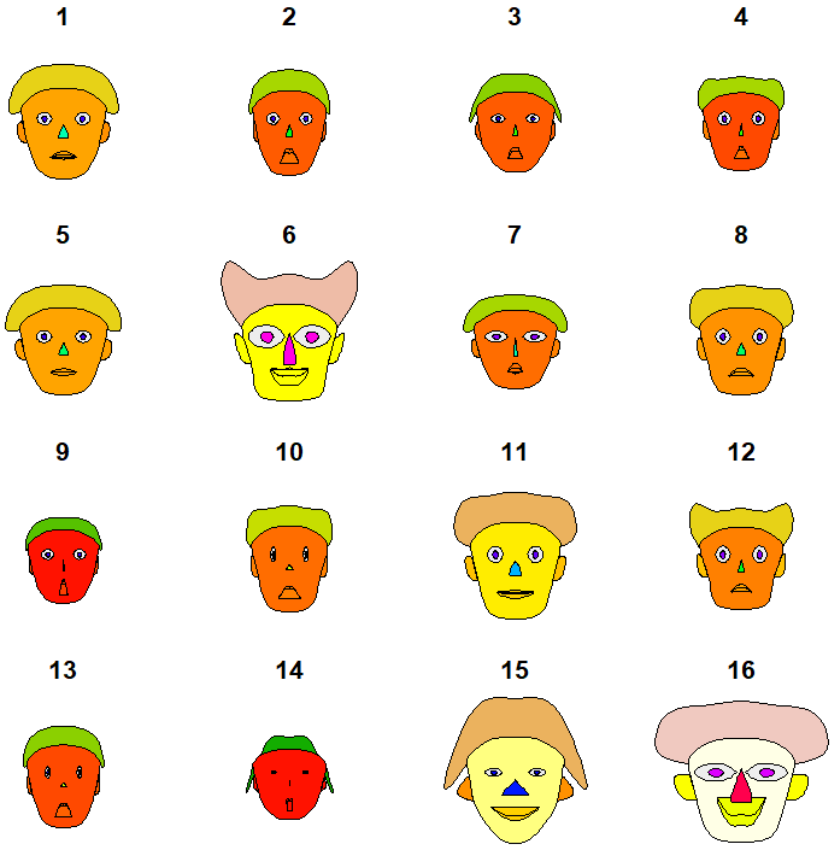
walks
mouth height, hair height, ear width

stolen bases
mouth width, hair width, ear height

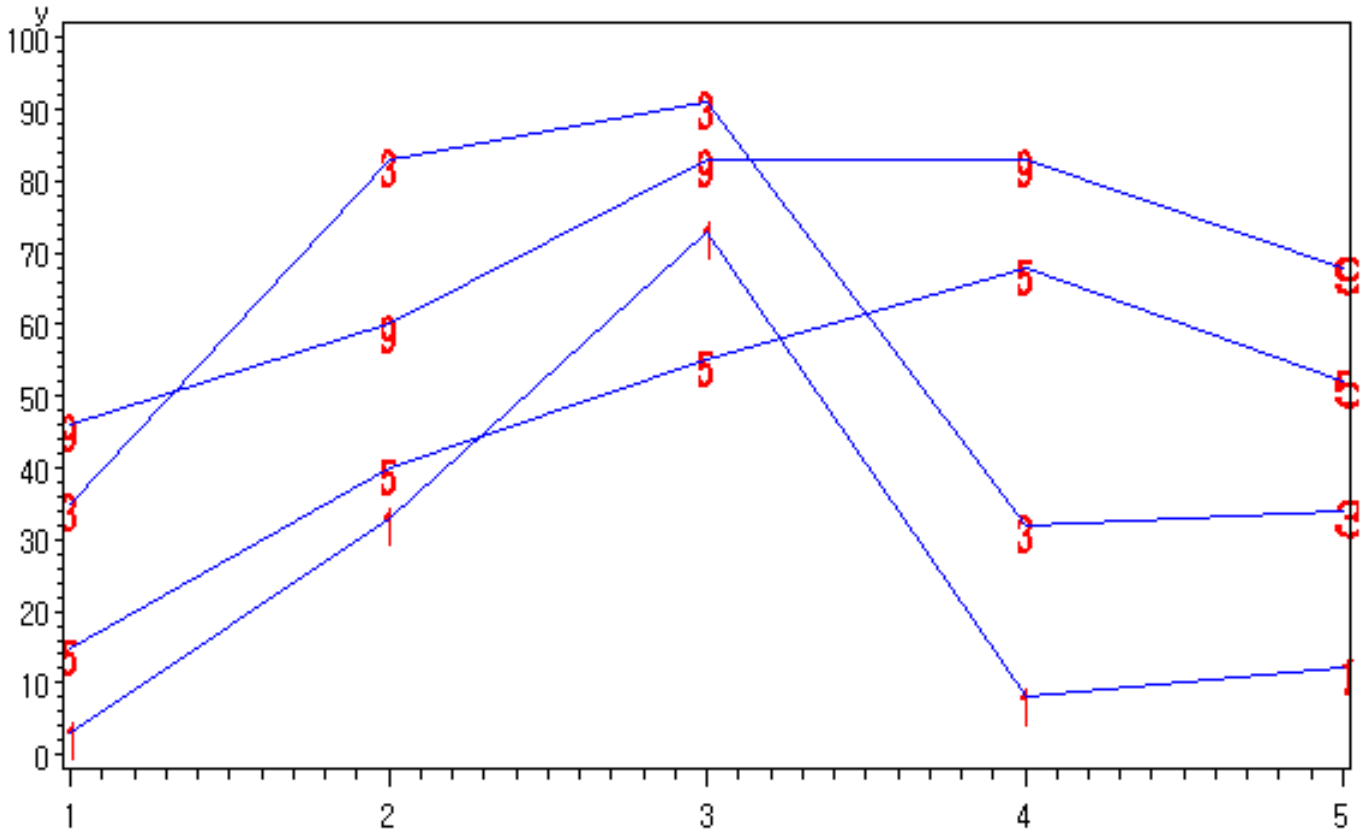
2005 National League: Chernoff Faces

체르노프의 얼굴 (Chernoff's face)

```
> library(aplpack)
> faces(iris[1:16,1:4])
 # first 16 observations
```

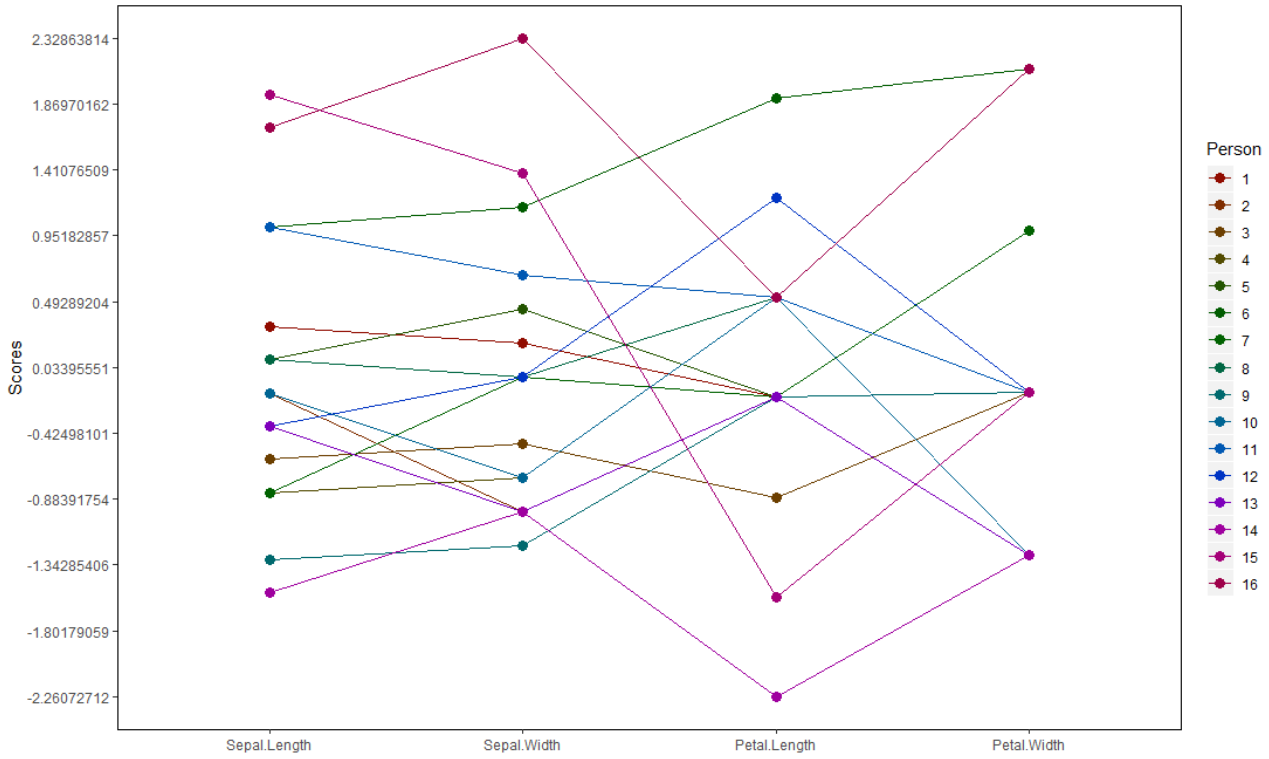


프로파일도표 (Profile Chart)



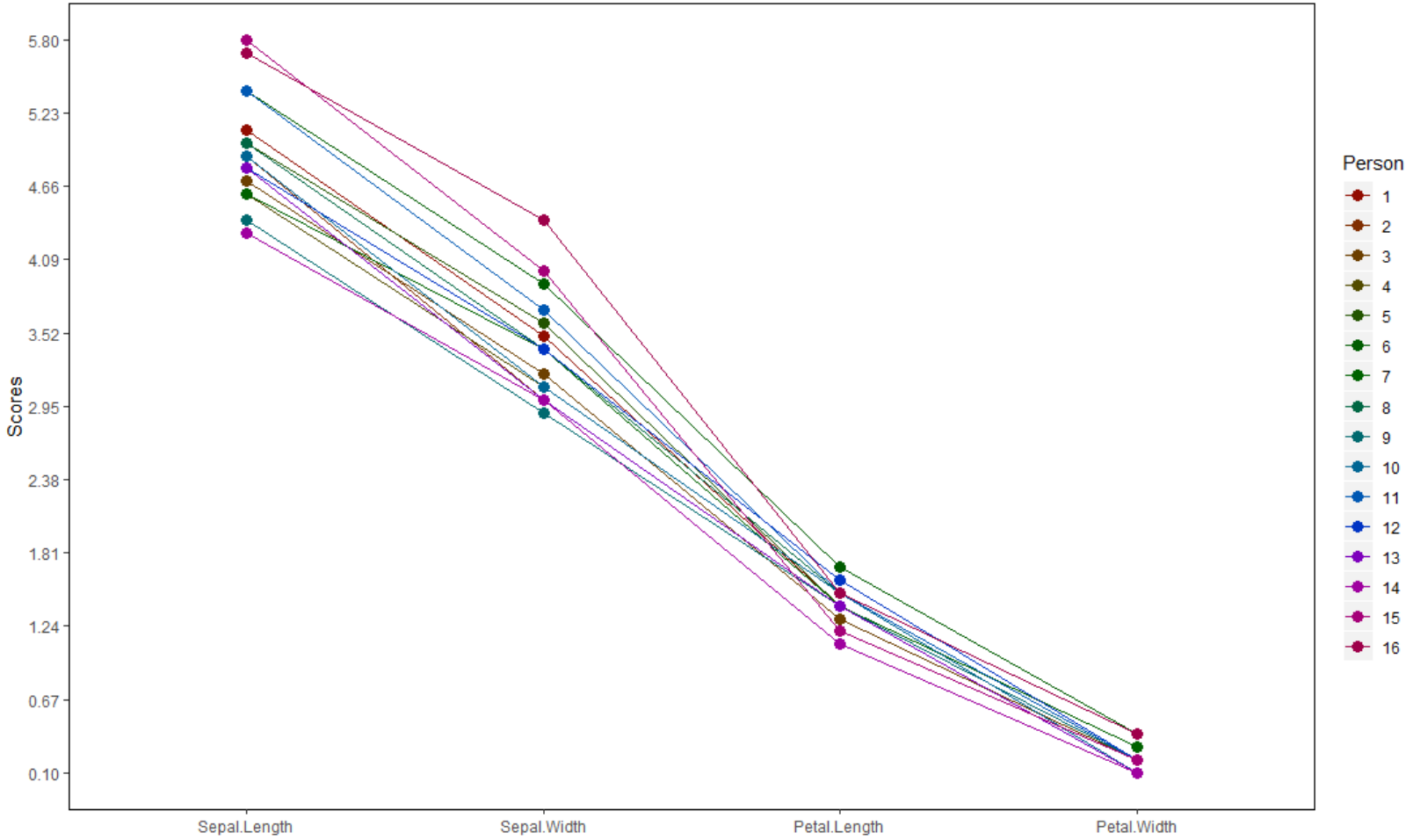
프로파일도표 (Profile Chart)

```
> library(profileR)
> profileplot(iris[1:16,1:4])
> profileplot(iris[1:5,1:4],standardize=FALSE)
```



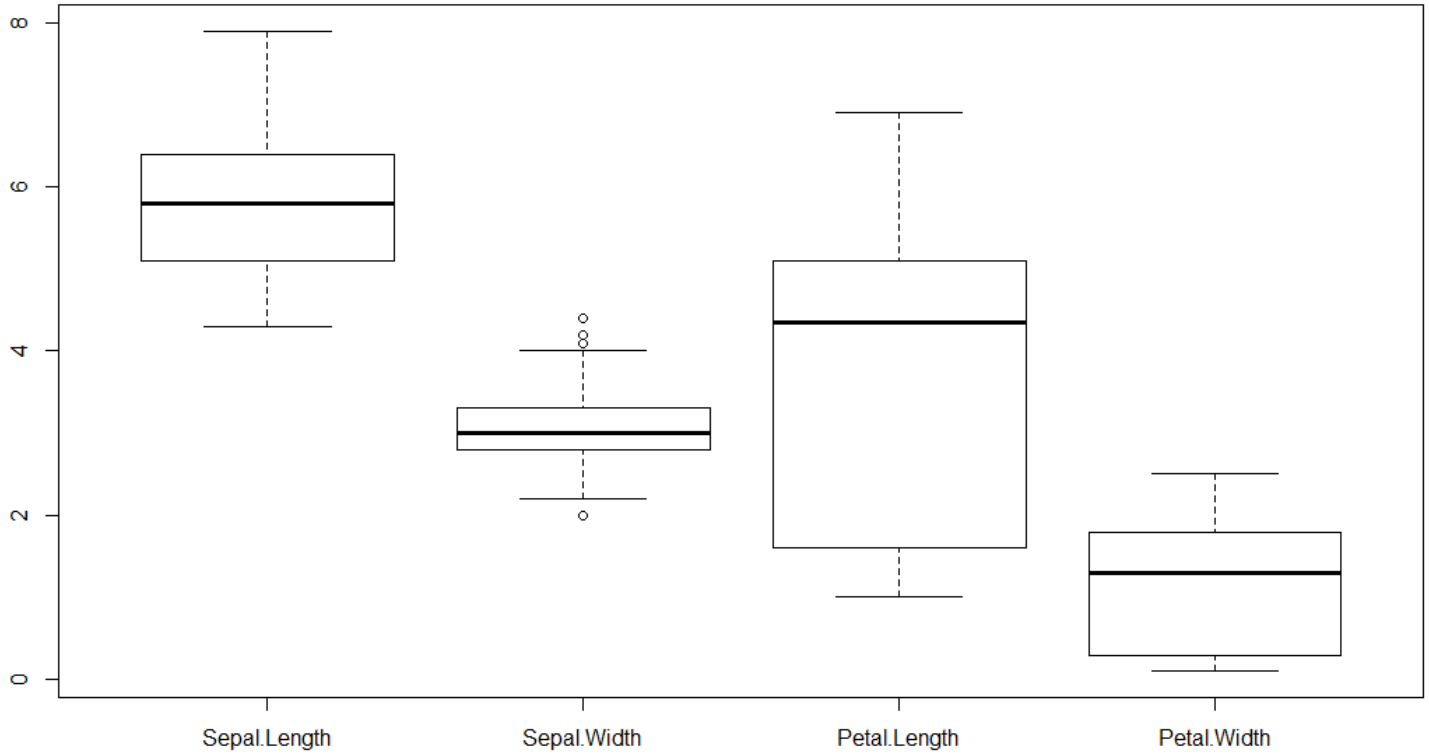
프로파일도표 (Profile Chart)

```
> profileplot(iris[1:16,1:4],standardize=FALSE)
```



상자 그림 (Boxplot)

```
> boxplot(iris[,1:4])
```



1.2 자료의 대수적 요약

- 평균 (Mean): $\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n = (x_{1j} + x_{2j} + \cdots + x_{ij} + \cdots + x_{nj})/n$

  평균벡터(Mean vector): $\bar{\mathbf{x}} = (\bar{x}_1,\ \bar{x}_2,\ \cdots,\ \bar{x}_j,\ \cdots,\ \bar{x}_p)' = \frac{1}{n}\mathbf{X}'\mathbf{1}.$

- 분산 (Variance): $s_{jj} = s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2, \quad j = 1,\ 2,\ \cdots,\ p.$

- 표준편차 (Standard Deviation): $s_j = \sqrt{s_j^2}, \quad j = 1,\ 2,\ \cdots,\ p.$

- 공분산 (Covariance): $s_{jk} = s_{kj} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \quad j,\ k = 1,\ 2,\ \cdots,\ p$

- 상관계수 (Correlation Coefficient):

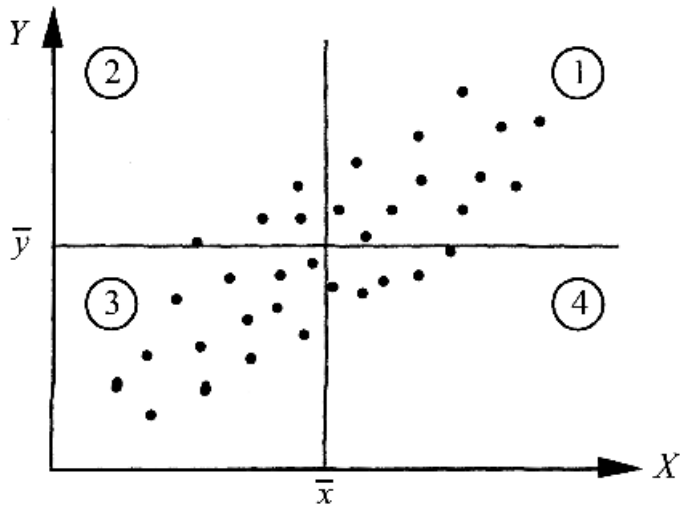
$$r_{jk} = r_{kj} = \frac{s_{jk}}{s_j\ s_k} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}$$

공분산 (Covariance)

$(y_i - \bar{y})$와 $(x_i - \bar{x})$의 값에 대한 대수적 부호

사분면	$y_i - y$	$x_i - \bar{x}$	$(y_i - y)(x_i - \bar{x})$
1	+	+	+
2	+	−	−
3	−	−	+
4	−	+	−

$$\mathrm{Cov}(Y, X) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1}$$



- Cov(*Y*, *X*) > 0 → *Y*와 *X* 사이에 양의 관계가 있다.
- Cov(*Y*, *X*) < 0 → *Y*와 *X* 사이에 음의 관계가 있다.
- 공분산은 측정단위의 변화에 영향을 받기 때문에, Cov(*Y*, *X*)는 그러한 관계의 강도(strength)가 얼마나 되는지를 알려주지는 않는다. $-\infty < \mathrm{Cov}(Y, X) < \infty$

공분산행렬 (Covariance Matrix)

$$\underset{p\times p}{\mathbf{S}} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1k} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2k} & \cdots & s_{2p} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ s_{j1} & s_{j2} & \cdots & s_{jk} & \cdots & s_{jp} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pk} & \cdots & s_{pp} \end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix} 996.3 & 544.6 & 266.9 & 504.0 & 480.4 \\ 544.6 & 484.5 & 234.4 & 133.1 & 169.8 \\ 266.9 & 234.4 & 153.2 & 48.8 & 86.2 \\ 504.0 & 133.1 & 48.8 & 815.4 & 693.1 \\ 480.4 & 169.8 & 86.2 & 693.1 & 622.0 \end{pmatrix}$$

상관계수 (Pearson's Correlation Coefficient)

$$\mathrm{Cor}(Y,X) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i-\bar{y}}{s_y}\right)\left(\frac{x_i-\bar{x}}{s_x}\right)$$

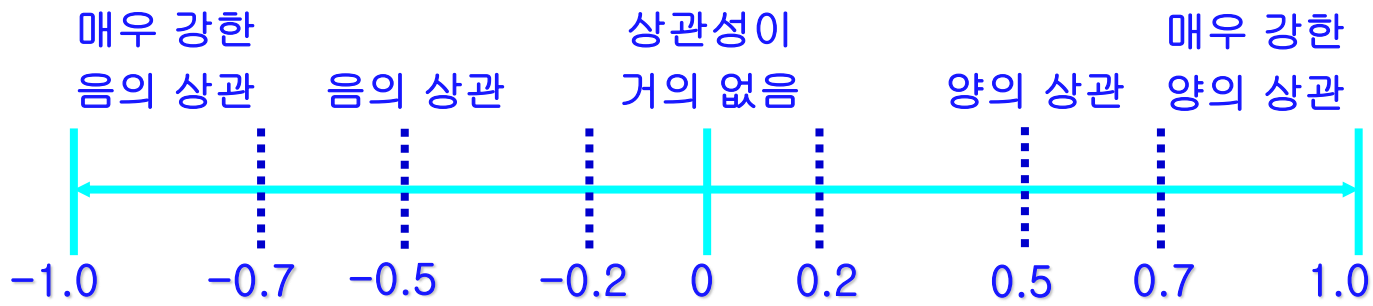
$$\mathrm{Cor}(Y,X) = \frac{\mathrm{Cov}(Y,X)}{s_y s_x}$$

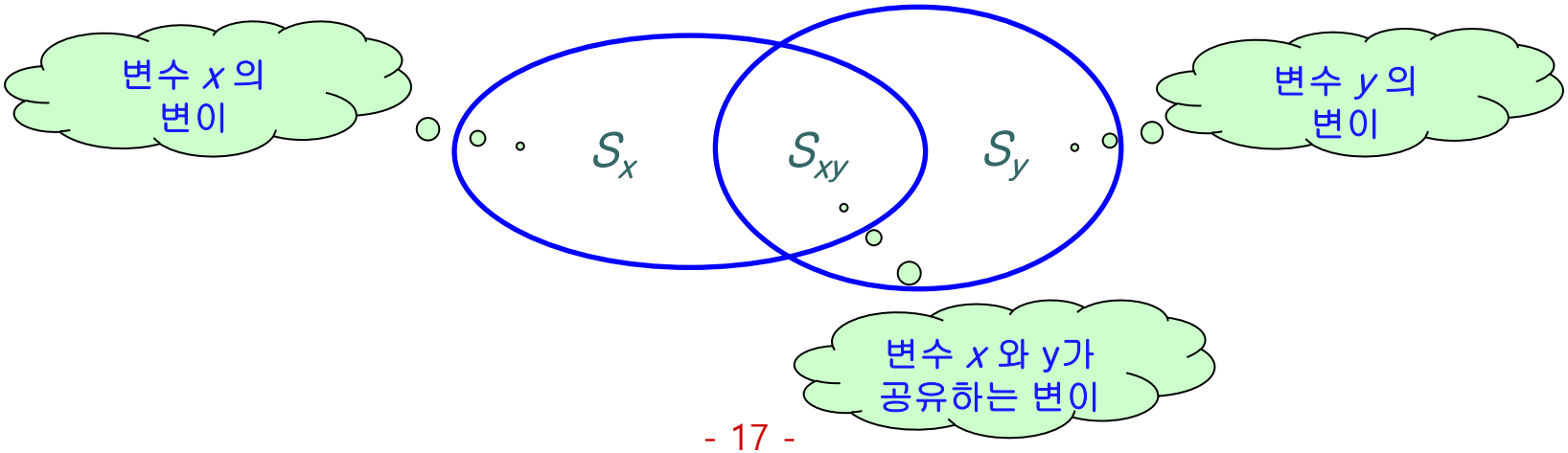
$$= \frac{\sum(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\sum(y_i-\bar{y})^2\sum(x_i-\bar{x})^2}}$$

- 공분산을 두 변수의 표준편차의 곱으로 나눈 것.
- 두 변수가 가지는 변이에 비하여 공유하는 변이의 양이 어느 정도인지를 나타냄.
- $-1 < \mathrm{Cor}(Y, X) < 1$
- $\mathrm{Cor}(Y, X) \fallingdotseq 1$: 강한 양(+)의 상관.
- $\mathrm{Cor}(Y, X) \fallingdotseq -1$: 강한 음(−)의 상관.
- $\mathrm{Cor}(Y, X) \fallingdotseq 0$: 무상관 (선형적인 증감의 관계가 없음).

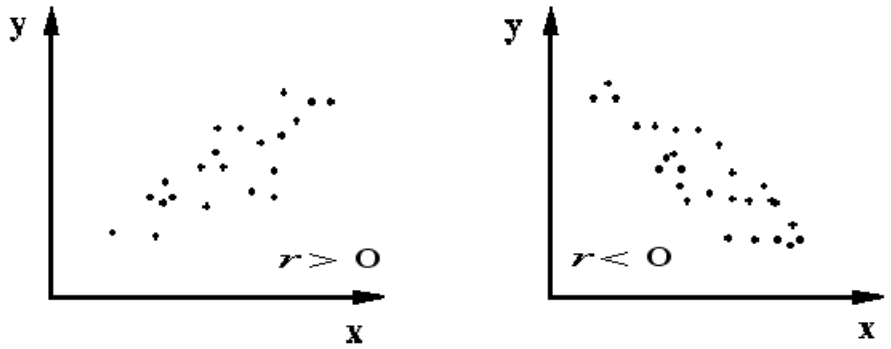
상관계수 (Pearson's Correlation Coefficient)

$$r = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{\dfrac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}}{\sqrt{\dfrac{\sum (X_i - \bar{X})^2}{n-1}} \cdot \sqrt{\dfrac{\sum (Y_i - \bar{Y})^2}{n-1}}} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \cdot \sqrt{\sum (Y_i - \bar{Y})^2}}$$

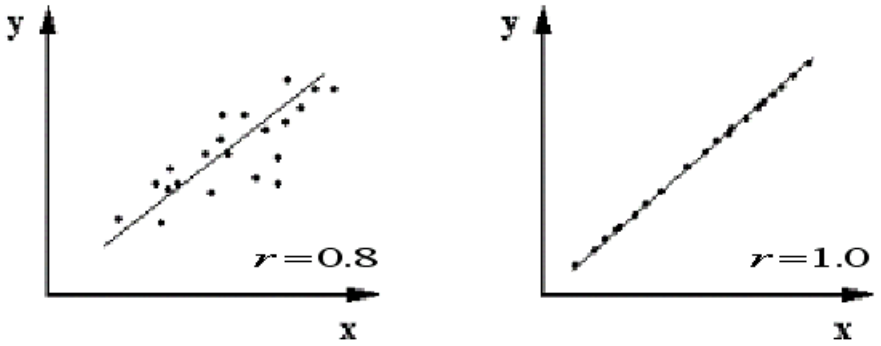




상관계수의 부호와 크기



(a) 상관계수의 부호

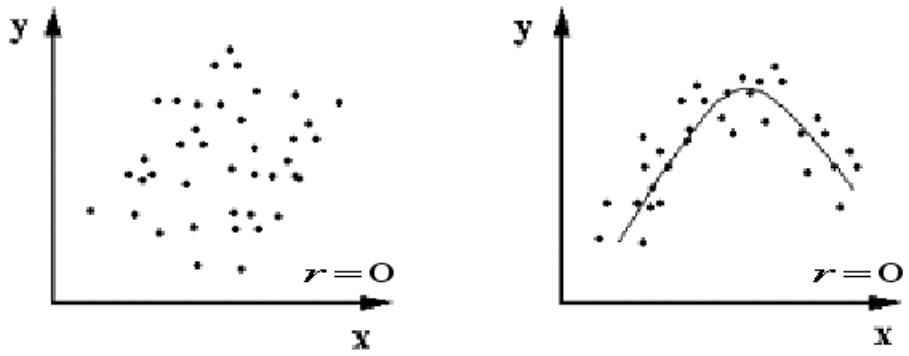


(b) 상관계수의 크기

- 상관계수의 부호는 증감의 방향성을 나타낸다.
- 상관계수의 절대값의 크기는 직선의 주변에 자료가 어느 정도 집중되어 있는지를 나타낸다.

상관계수 해석의 제약성: 선형성 (Linearity)

- 상관계수는 두 변수의 '선형집중성'만을 을 재는 측도로서 비선형 연관관계를 반영하지 못함.



상관계수가 0인 경우

Y와 X 사이에 완벽한 비선형관계를 가지지만 $Cor(Y, X) = 0$인 데이터 세트

Y	X	Y	X	Y	X
1	−7	46	−2	41	3
14	−6	49	−1	34	4
25	−5	50	0	25	5
34	−4	49	1	14	6
41	−3	46	2	1	7

상관행렬 (Correlation Matrix)

$$\underset{p\times p}{\mathbf{R}} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2k} & \cdots & r_{2p} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ r_{j1} & r_{j2} & \cdots & r_{jk} & \cdots & r_{jp} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pk} & \cdots & 1 \end{pmatrix}$$

$$\mathbf{R} = \begin{pmatrix} 1 & 0.784 & 0.683 & 0.559 & 0.610 \\ 0.784 & 1 & 0.860 & 0.212 & 0.309 \\ 0.683 & 0.860 & 1 & 0.138 & 0.279 \\ 0.559 & 0.212 & 0.138 & 1 & 0.973 \\ 0.610 & 0.309 & 0.279 & 0.973 & 1 \end{pmatrix}$$

예) 평균, 분산 및 공분산행렬과 상관행렬의 계산

```
> mean(iris[,1])
> sd(iris[,1])
> var(iris[,1])

> cov(iris[,1],iris[,2])
> cov(iris[,3],iris[,4])
> cov(iris[,1:4])

> cor(iris[,1],iris[,2])
> cor(iris[,3],iris[,4])
> cor(iris[,1:4])
```

1.2.5 중심화와 표준화

- 표준화 (Standardization)

$$z_i = \frac{y_i - \bar{y}}{s_y}$$

중심화 (Centering)
평균이 0이 되도록 함.
중심으로부터의 편차에 관심을 가짐.

척도화 (Scaling)
표준편차가 1이 되도록 함.
측정단위 자체를 없앰.

- 관측치의 상대적 위치의 척도로 사용됨.
- 관측치간 상대적인 크기를 비교할 수 있음.
- 단위가 없는 순수한 수치. (평균=0, 표준편차=1)
- 관측치 전체 데이터 내에서의 위치를 나타내는 데 효율적으로 사용됨.
- 표준점수가 ±2.0(또는 ±3.0)을 벗어나면 특이값으로 볼 수 있음.

중심화된 자료행렬

$$\mathbf{C} = \{c_{ij}\} = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1j} - \bar{x}_j & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2j} - \bar{x}_j & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{i1} - \bar{x}_1 & x_{i2} - \bar{x}_2 & \cdots & x_{ij} - \bar{x}_j & \cdots & x_{ip} - \bar{x}_p \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nj} - \bar{x}_j & \cdots & x_{np} - \bar{x}_p \end{pmatrix}$$

$$\mathbf{C} = \begin{pmatrix} -29.1 & -26.5 & -4.5 & -38.6 & -30.4 \\ -29.1 & -29.5 & -18.5 & -18.6 & -22.4 \\ 2.9 & 23.5 & 13.5 & -14.6 & -8.4 \\ 2.9 & 23.5 & 7.5 & -13.6 & -10.4 \\ -17.1 & -19.5 & -22.5 & 21.4 & 9.6 \\ -29.1 & -6.5 & -1.5 & -36.6 & -34.4 \\ 35.9 & 23.5 & 7.5 & 1.4 & 7.6 \\ -17.1 & -12.5 & -0.5 & 29.4 & 33.6 \\ 13.9 & 0.5 & 5.5 & 36.4 & 25.6 \\ 65.9 & 23.5 & 13.5 & 33.4 & 29.6 \end{pmatrix}$$

$$\mathbf{S} = \mathbf{C}'\mathbf{C}/(n-1)$$

표준화된 자료행렬

$$\mathbf{Z} = \{z_{ij}\} = \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{s_1} & \cdots & \frac{x_{1j}-\bar{x}_j}{s_j} & \cdots & \frac{x_{1p}-\bar{x}_p}{s_p} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ \frac{x_{i1}-\bar{x}_1}{s_1} & \cdots & \frac{x_{ij}-\bar{x}_j}{s_j} & \cdots & \frac{x_{ip}-\bar{x}_p}{s_p} \\ \vdots & \cdots & \vdots & \cdots & \cdots \\ \frac{x_{n1}-\bar{x}_1}{s_1} & \cdots & \frac{x_{nj}-\bar{x}_j}{s_j} & \cdots & \frac{x_{np}-\bar{x}_p}{s_p} \end{pmatrix}$$

$$\mathbf{Z} = \begin{pmatrix} -0.922 & -1.204 & -0.364 & -1.352 & -1.219 \\ -0.922 & -1.340 & -1.495 & -0.651 & -0.898 \\ 0.092 & 1.068 & 1.091 & -0.511 & -0.337 \\ 0.092 & 1.068 & 0.606 & -0.476 & -0.417 \\ -0.542 & -0.886 & -1.818 & 0.749 & 0.385 \\ -0.922 & -0.295 & -0.121 & -1.282 & -1.379 \\ 1.137 & 1.068 & 0.606 & 0.049 & 0.305 \\ -0.542 & -0.568 & -0.040 & 1.030 & 1.347 \\ 0.440 & 0.023 & 0.444 & 1.275 & 1.026 \\ 2.088 & 1.068 & 1.091 & 1.170 & 1.187 \end{pmatrix}$$

$$\mathbf{R}_x = \mathbf{S}_z = \mathbf{Z}'\mathbf{Z}/(n-1)$$

예) 중심화와 표준화

- 중심화

```
> iris1.center<-scale(iris[,1],center=TRUE,scale=FALSE)
> mean(iris1.center)
[1] -3.666902e-16
> sd(iris1.center)
[1] 0.8280661
```

- 표준화

```
> iris1.std<-scale(iris[,1],center=TRUE,scale=TRUE)

> mean(iris1.std)
[1] -4.484318e-16
> sd(iris1.std)
[1] 1
```