Random Vector

Mean, Variance, Covariance & Graphs

H. Park

HUFS

확률벡터(_____)

- 변수(variable) 그리고 확률변수(random variable)
- 1개의 확률변수(random variable), 평균, 분산
- 2개의 확률변수, 평균, 분산, 공분산, 상관계수
- 여러 개의 확률변수 ... 그리고 확률벡터

공분산행렬(_____)

- x_{ij} = i-번째 개체의 j-번째 변수값
- x_{i1} = 키, x_{i2} =몸무게, x_{i3} =가슴둘레

$$X = \begin{pmatrix} 165 & 63 & 85 \\ 170 & 70 & 90 \\ 180 & 75 & 105 \\ 173 & 70 & 85 \end{pmatrix} \longrightarrow$$

$$\boldsymbol{x}_1 = \begin{pmatrix} 165 & 63 & 85 \end{pmatrix}^T = \begin{pmatrix} 165 \\ 63 \\ 85 \end{pmatrix} = \text{1st sub. vector}$$

$$\bar{\boldsymbol{x}} = \begin{pmatrix} \bar{x}_1 & \bar{x}_2 & \bar{x}_3 \end{pmatrix}^T = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{pmatrix} = \text{sample mean vectorr}$$

sample covariance/correlation

$$S = \text{sample covariance matrix}$$

$$= \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{pmatrix}$$

$$D = diag \begin{pmatrix} s_{11} & s_{22} & s_{33} \end{pmatrix}$$

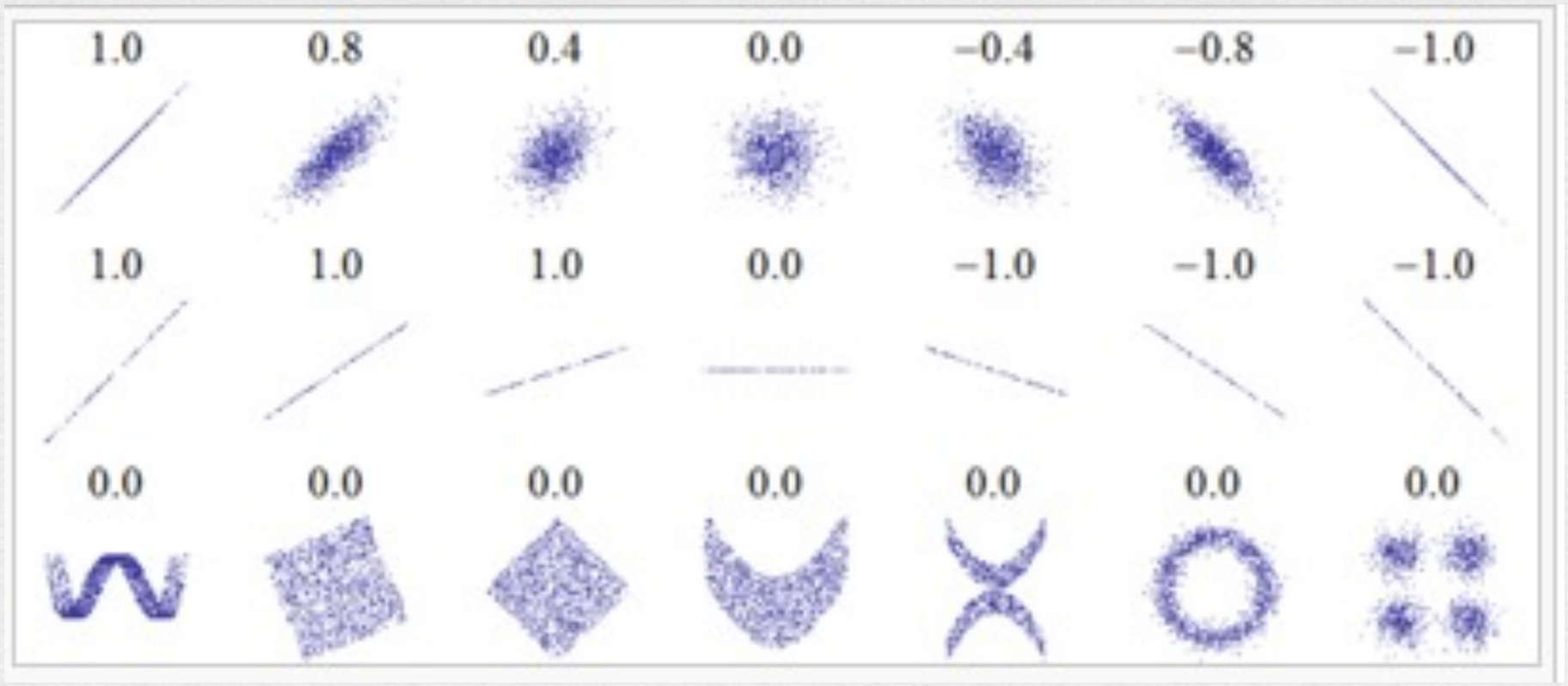
$R =$ sample correlation matrix

$$= D^{-1/2} S D^{-1/2}$$

$$= \begin{pmatrix} \sqrt{s_{11}} & 0 & 0 \\ 0 & \sqrt{s_{22}} & 0 \\ 0 & 0 & \sqrt{s_{33}} \end{pmatrix}^{-1} \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{23} & s_{33} \end{pmatrix} \begin{pmatrix} \sqrt{s_{11}} & 0 & 0 \\ 0 & \sqrt{s_{22}} & 0 \\ 0 & 0 & \sqrt{s_{33}} \end{pmatrix}^{-1}$$

상관계수의 헛점

- $r_{xy} = \pm 1$
- $r_{xy} = 0$
- $r_{xy} > 0$
- $r_{xy} < 0$



exercise (R 실습)

- Find the sample mean vector (키, 몸무게, 가슴둘레)
- Find the sample covariance matrix
- Find the sample correlation matrix
- Verify $R = D^{-1/2} S D^{-1/2}$
- Verify $$S = \frac{1}{4-1} \sum_{i=1}^{4} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$$

키	sub1	sub2	sub3	sub4
$x_{i1} - \bar{x}_1$				
$(x_{i1} - \bar{x}_1)^2$				

몸무게	sub1	sub2	sub3	sub4
$x_{i2} - \bar{x}_2$				
$(x_{i2} - \bar{x}_2)^2$				

키*몸무게	sub1	sub2	sub3	sub4
$x_{i1} - \bar{x}_1$				
$x_{i2} - \bar{x}_2$				
$(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$				

Population Mean/Variance

- sample mean vs. (population) mean

$$\bar{x} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{pmatrix} \quad \text{vs.} \quad E(X) = \begin{pmatrix} E(x_1) \\ E(x_2) \\ E(x_3) \end{pmatrix}$$

- sample covariance matrix

vs. (population) covariance matrix

$$S = \frac{1}{4-1} \sum_{i=1}^{4} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$$

$$\Sigma = E\left\{ (\boldsymbol{x} - E(\boldsymbol{x}))\,(\boldsymbol{x} - E(\boldsymbol{x}))^T \right\}$$

Correlation Coefficient
상관계수

- 아래 상관계수의 정의를 기록하고, 어떤 분야에 어떤 계수를 사용할지 설명해 보아라
(관련 자료출처 기록 필수)

1. 피어슨 상관계수 (Pearson Correlation Coefficient)

2. 스피어만 상관계수 (Spearman Correlation Coefficient)

3. 크론바 알파계수 (Cronbach Alpha Coefficient)

Multiple Correlation Coefficient
다중상관계수

"An estimate of the combined influence of two or more variables on the observed (dependent) variable"

" 변수 와 나머지 변수들,
의 선형결합 의 상관관계를
나타내는 상관계수 가운데 최대값 "

Multiple Correlation Coefficient (다중상관계수)

y vs. $x_1, x_2, \cdots, x_q$

$max\ Corr(y, \boldsymbol{\beta}^T \boldsymbol{x})$ 

S_q = sample covariance matrix for q-variables

$\boldsymbol{s}_{oq}$ = sample covariance vector, $\mathrm{cov}(y, x_i)$

$$= \begin{pmatrix} \mathrm{cov}(y, x_1) \\ \mathrm{cov}(y, x_2) \\ \cdots \\ \mathrm{cov}(y, x_q) \end{pmatrix}$$

$$s_o = \sqrt{\mathrm{var}(y)}$$

다중상관계수 쉽게 구하는 방법

Regress y *on* $x_1,\ x_2,\ \cdots,\ x_q$ → [ ]

R^2

→ [ ]

-다중상관계수 (_____)

-결정계수 (_____)

다중상관계수의 계산 (R 실습)

1) $\boldsymbol{S}_2 = \begin{pmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) \\ \mathrm{cov}(x_2, x_1) & \mathrm{var}(x_2) \end{pmatrix}$ 계산

2) $s_o = \sqrt{\mathrm{var}(y)}$ 계산

3) $\boldsymbol{s}_{o2} = \begin{pmatrix} \mathrm{cov}(y, x_1) \\ \mathrm{cov}(y, x_2) \end{pmatrix}$ 계산

4) $r_{max} = \dfrac{\sqrt{\boldsymbol{s}_{02}^T S_2^{-1} \boldsymbol{s}_{02}}}{s_0}$ 계산

Table 3.1: Data on holiday cottages

Price	Age	Area
DDK 1000	years	m^2
(y)	(x_1)	(x_2)
745	36	66
895	37	68
442	47	64
440	32	53
1598	1	101

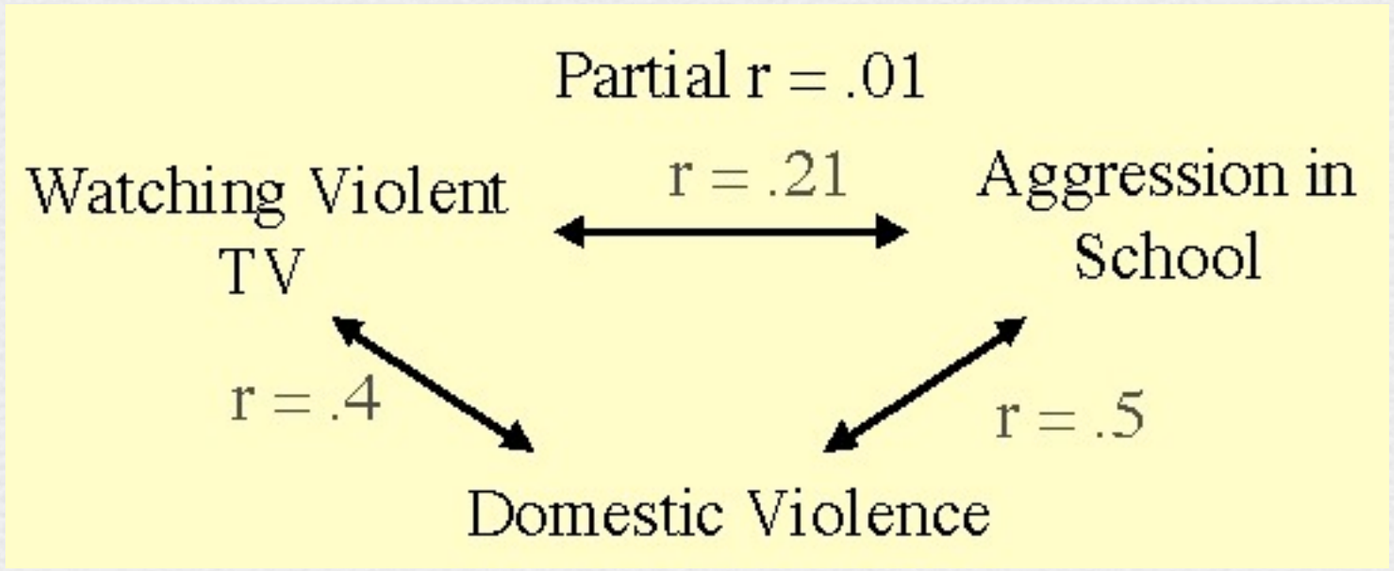
6) $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$ 의 R^2 와 비교

Partial Correlation Coefficient

- A measure of the strength of association between a dependent variable and one independent variable when the effect of all other independent variable's effect is removed;

Read more: http://www.answers.com/topic/partial-correlation-coefficient#ixzz1WQV0BYbh

A second example has to do with the covariation between economic growth rate and social conflict. Let us hypothesize that they have a negative correlation due to economic growth increasing opportunities, multiple group membership, and cross-pressures, thus draining off conflict. Let us find, however, that the actual correlation is near zero. Before rushing out into the streets to proclaim that economic growth is independent of conflict, however, we might consider whether exogenous influences are dampening the real correlation. In this case, we could argue that the educational growth rate is the depressant. Increasing education creates new interests, broadens expectations, and generates a consciousness of deprivations. Thus, if education increases faster than opportunities, social conflict would increase. To assess the correlation between economic growth rate and conflict, therefore, we should hold constant the educational growth rate.



$$\underbrace{x_1,\ x_2}_{(1)},\ \underbrace{x_3,\ x_4}_{(2)} \longrightarrow \quad \text{or}$$

Partial correlation coefficients are very useful because it allows to directly estimate the proportion of unexplained variation of y that becomes explained with the addition of variable x [to the model]

$$\underbrace{y_1,\ x_3}_{(1)},\ \underbrace{x_1,\ x_2}_{(2)} \longrightarrow \quad \text{or}$$

$$R_{y_1 x_3 . x_1 x_2} = \text{Partial } R^2 \text{ for reg. of } y_1 \text{ on } x_3 \text{ given } x_1,\ x_2$$

Partial Correlation Coefficient 공식

- Divide $x_1,\ x_2,\ \cdots,\ x_p$ into two groups $\boldsymbol{x}_{(1)},\ \boldsymbol{x}_{(2)}$
- Select two variables $x_1,\ x_2$ among $\boldsymbol{x}_{(1)}$
- correlation coeff. between $x_1,\ x_2$ given $\boldsymbol{x}_{(2)}$

$$
\begin{aligned}
S &= \text{sample covariance matrix} \\
&= \left(\begin{array}{cc|c} s_{11} & s_{12} & s_{1(2)} \\ s_{21} & s_{22} & s_{2(2)} \\ \hline s_{(2)1} & s_{(2)2} & s_{(2)(2)} \end{array}\right) \\
&= \left(\begin{array}{c|c} S_{11} & S_{1(2)} \\ \hline S_{(2)1} & S_{(2)(2)} \end{array}\right)
\end{aligned}
$$

부분상관계수의 계산 (R 실습)

$$\boldsymbol{S} = \begin{pmatrix} \mathrm{var}(y) & \mathrm{cov}(y, x_1) & \mathrm{cov}(y, x_2) \\ \mathrm{cov}(x_1, y) & \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) \\ \mathrm{cov}(x_2, y) & \mathrm{cov}(x_2, x_1) & \mathrm{var}(x_2) \end{pmatrix}$$

$\phi_{yx_1.(x_2)}$

$R^2_{yx_1.x_2}$

r_{yx_1}

$R^2_{yx_1}$

Table 3.1: Data on holiday cottages

Price	Age	Area
DDK 1000	years	m^2
(y)	(x_1)	(x_2)
745	36	66
895	37	68
442	47	64
440	32	53
1598	1	101

부분결정계수 쉽게 구하는 방법 (R 실습)

$R^2_{y_1 x_2 . x_1}$

1) Residual_1 from the regression of y_1 on x_1

2) Residual_2 from the regression of x_2 on x_1

3) R^2 from the regression of Residual_2 on Residual_1

Canonical Correlation Coefficient

- Divide $x_1,\ x_2,\ \cdots,\ x_p$ into two groups $\boldsymbol{x}_{(1)},\ \boldsymbol{x}_{(2)}$
- Find the maximum correlation coefficient between $\beta_{(1)}^T \boldsymbol{x}_{(1)}$ and $\beta_{(2)}^T \boldsymbol{x}_{(2)}$

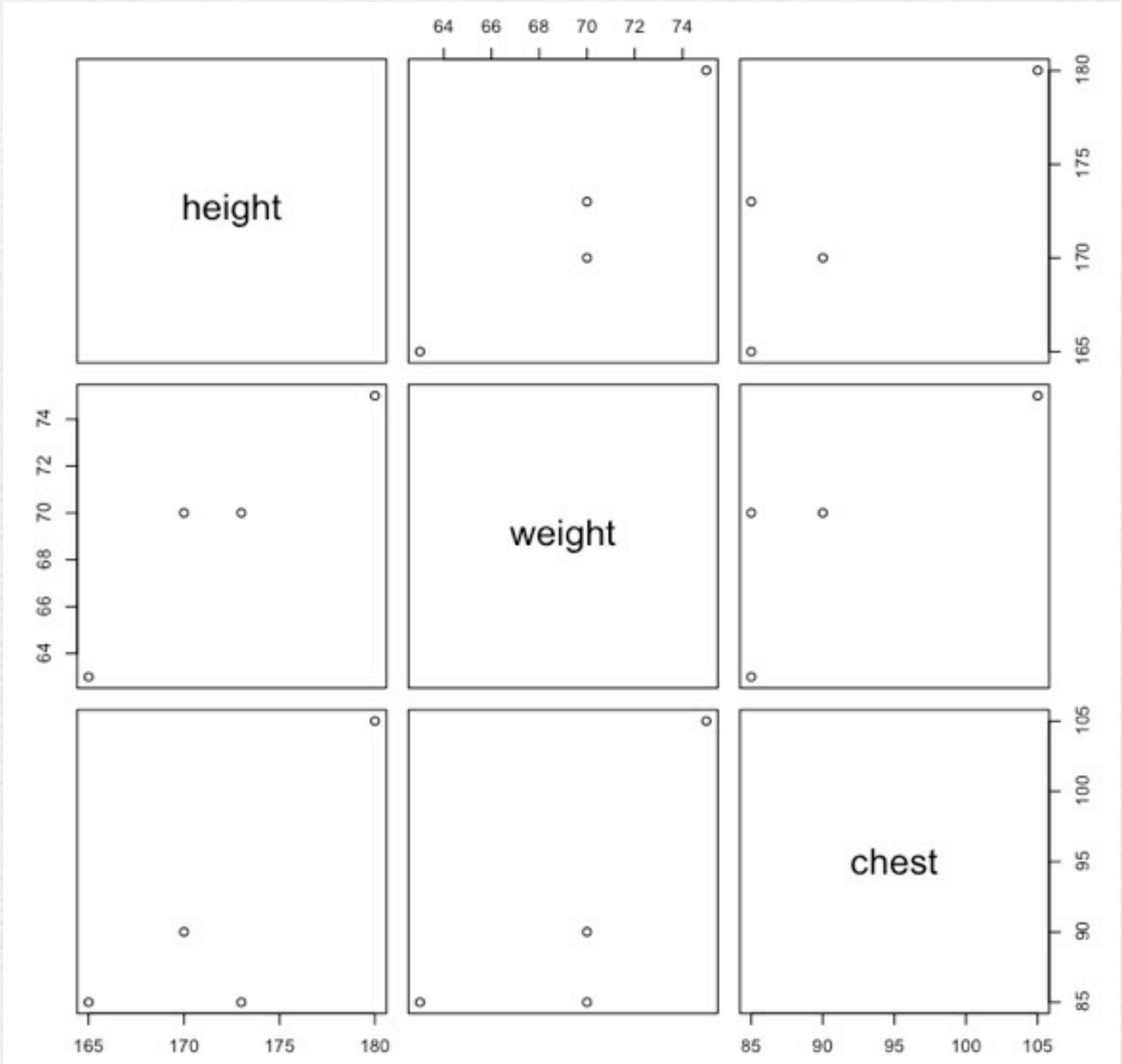
>>>> later ...

Summary

- Correlation Coefficient (상관계수)
- 피어슨 상관계수, 스피어만 상관계수, 크론바 알파
- Multiple Correlation Coefficient (다중상관계수)
- Coefficient of Determination (결정계수)
- Partial Correlation Coefficient (부분상관계수)
- Partial Coefficient of Determination (부분결정계수)
- Canonical Correlation Coefficient (정중상관계수)

Scatter Plot / Matrix Plot

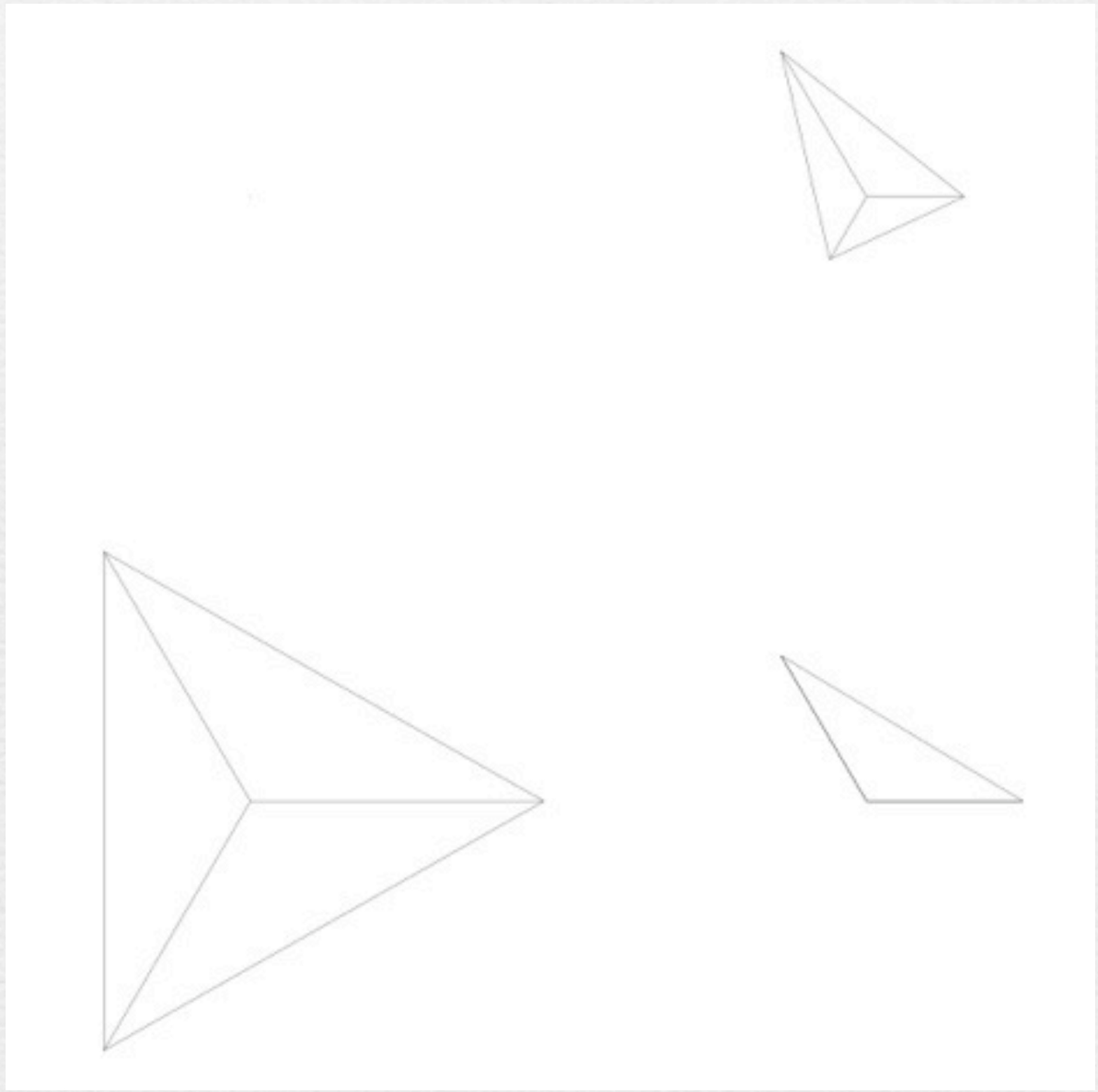
- 두 변수간의 관계를 설명
- (R) _____
- (R) _____



Star Plot

- 각 사람(개체)에 대한 여러 변수의 크기를 표현함으로써 여러 개체를 비슷한 군(group)으로 구분할 수 있다.

- (R) _____



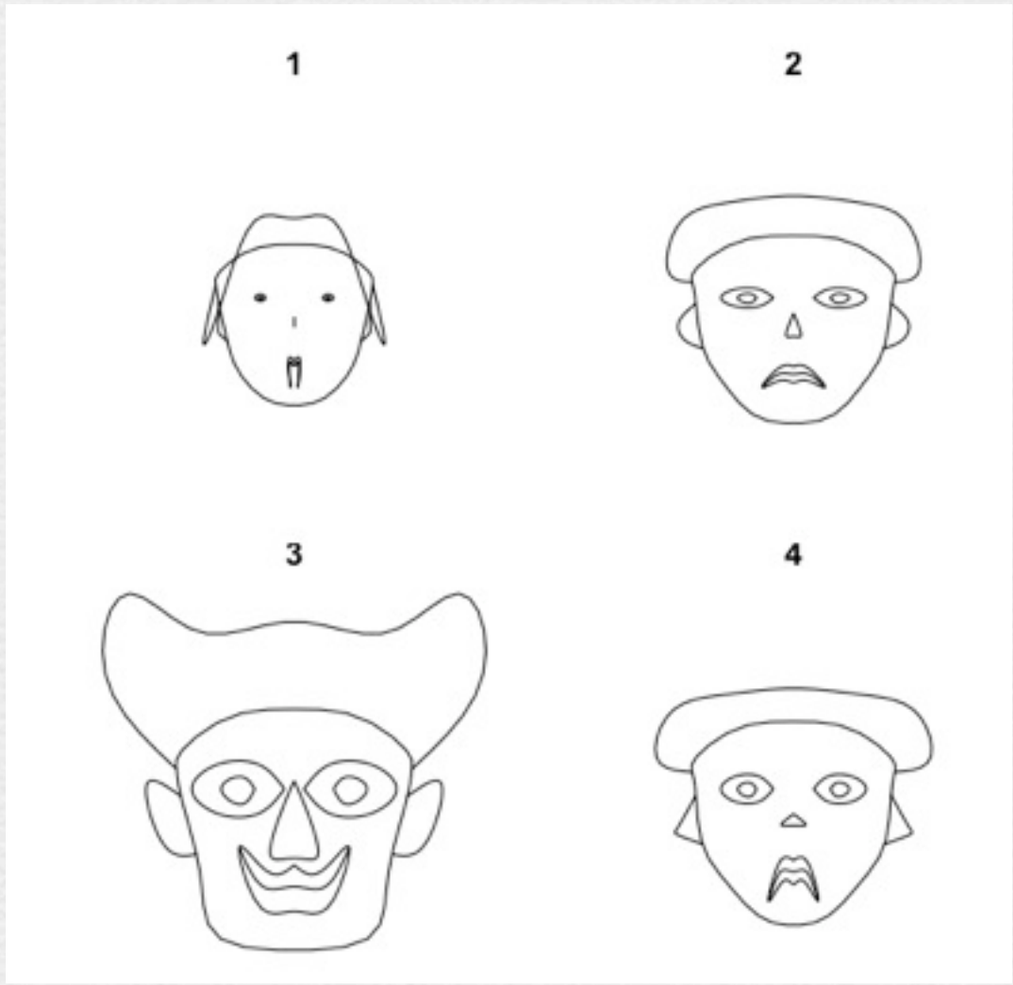
Chernoff Face Plot

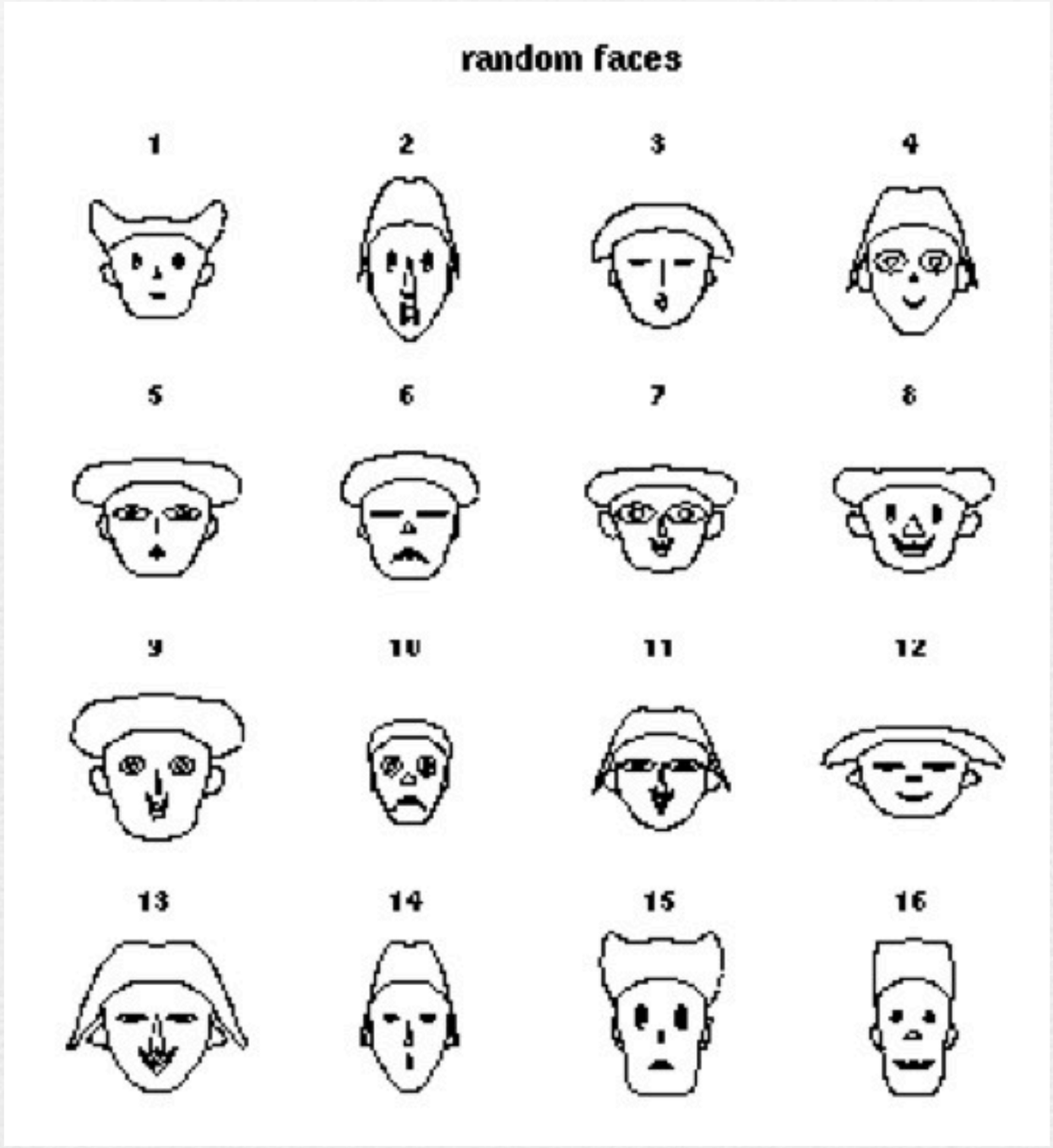
- 사람의 얼굴 특징을 이용하여 각 사람(개체)의 다변량 자료의 값을 동시에 표현한 그림. 사람의 얼굴에 특징에 민감한 인간의 속성을 이용함.

- (R) http://www.wiwi.uni-bielefeld.de/~wolf/software/R-wtools/faces/faces.R

- 윗 함수를 실행시킨 후,

_____ 수행함





1-height of face

2-width of face

3-shape of face

4-height of mouth

5-width of mouth

6-curve of smile

7-height of eyes

8-width of eyes

9-height of hair

10-width of hair

11-styling of hair

12-height of nose

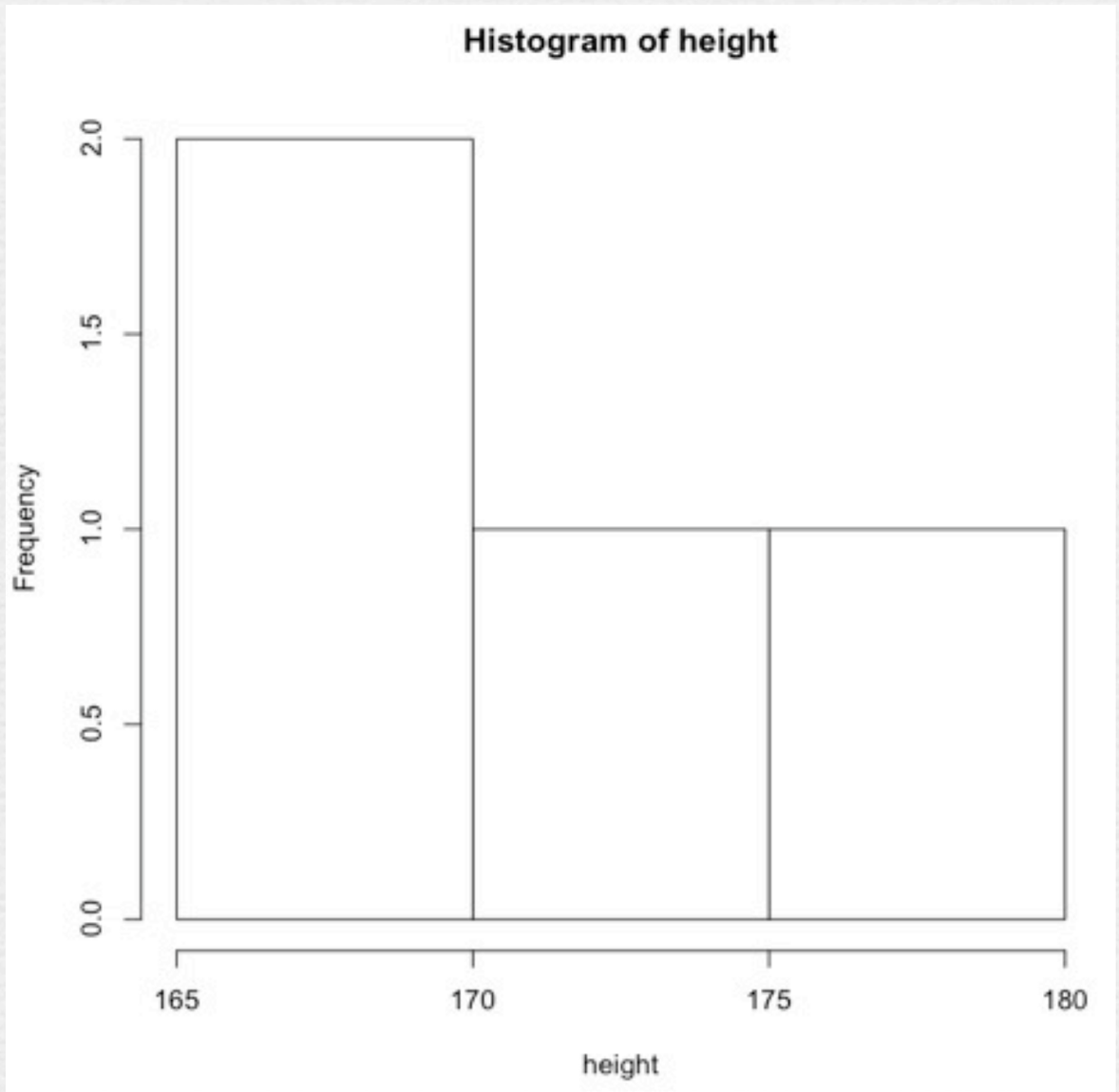
13-width of nose

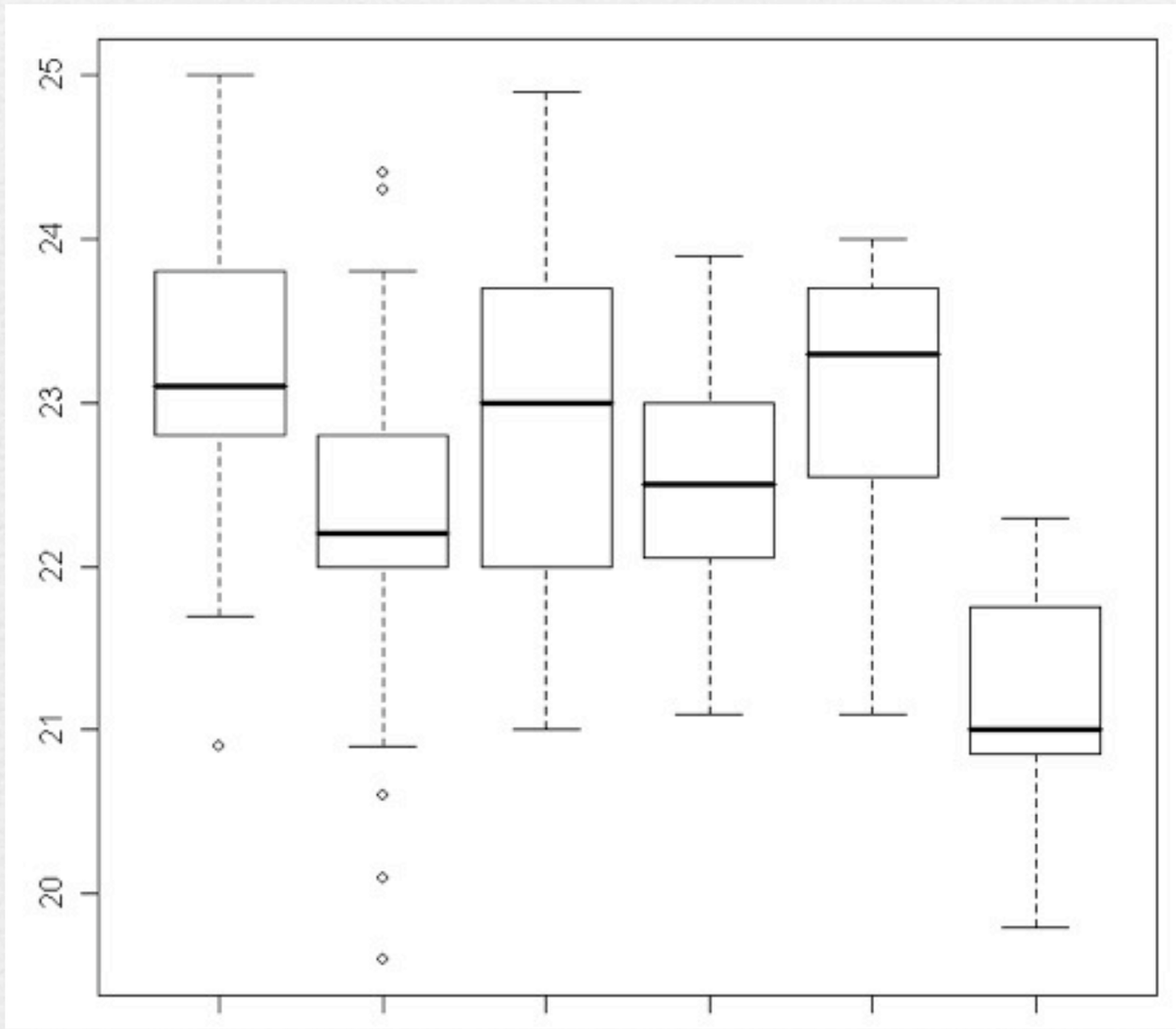
14-width of ears

15-height of ears.

Histogram / Box plot

- 변수의 분포를 알아보거나 변수간의 값의 크기를 비교하는데 유용한 그림
- (R) _____

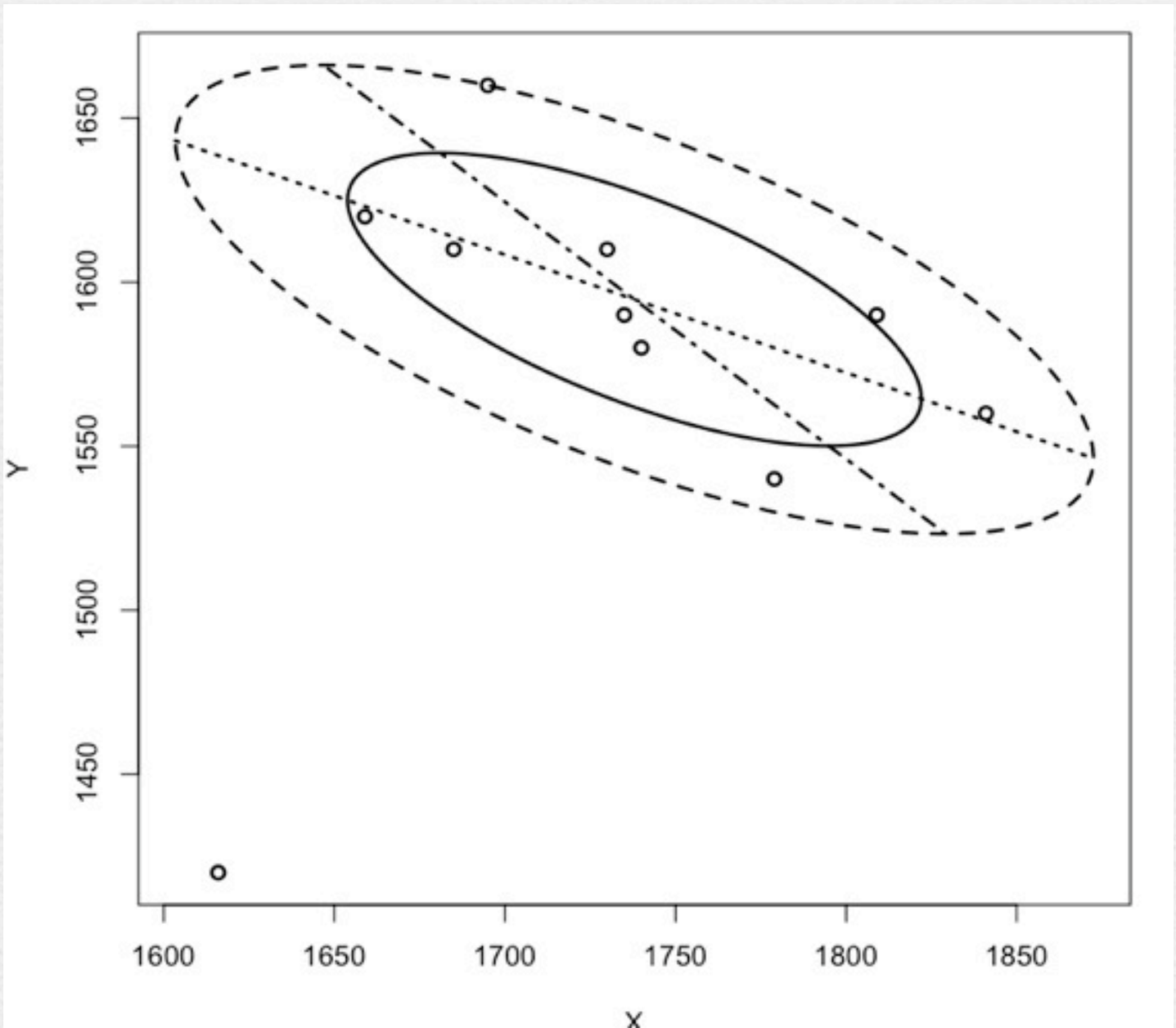




bivariate boxplot

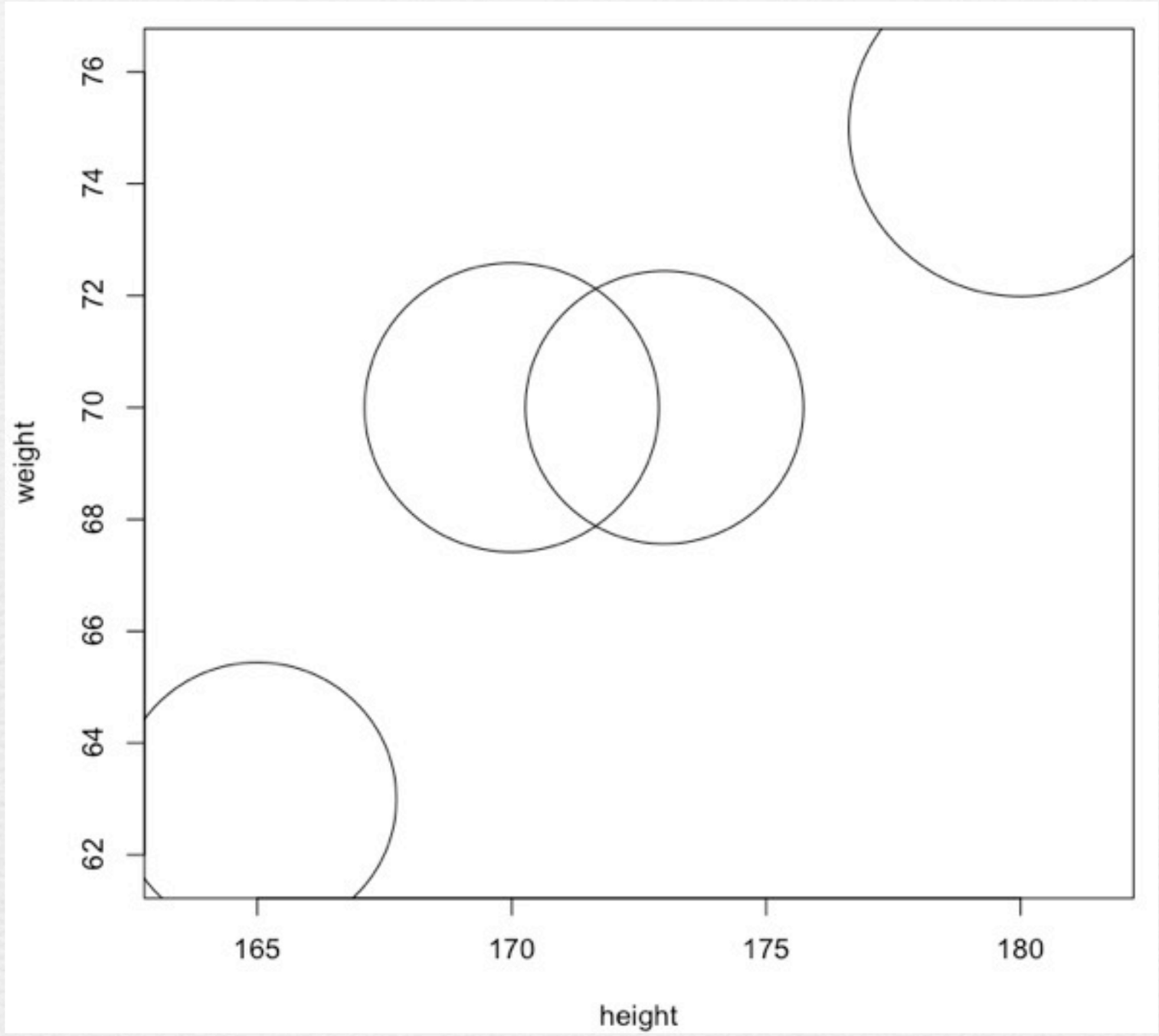
- 서로 다른 두 개의 변수에 대한 이변량 분포를 쉽게 확인하는 그래프

- (R) http://biostatistics.iop.kcl.ac.uk/publications/everitt/ 에서 zip 된 파일을 다운받아서 c:\mvar 디렉토리에 저장시킨 후, R 윈도우에서 function 을 수행한 다음에 bvbox(cbind(height,weight)) 라고 수행시킨다 (_____타원)



Bubble Plot

- 연관성있는 3개의 변수를 동시에 표시할 수 있는 그래프
- (R) _____



숙제

- 미국 40개 도시의 오염자료를 사용 (e-class)
- 각 변수간의 matrix plot 을 사용해 분석하라
- star plot, chernoff plot 을 사용해 분석하라
- 이변량 분포를 나타내는 변수쌍(pair) 가 있으면 적어도 한 쌍을 선택하고 그 근거를 기술하라