데이터과학을 위한 R프로그래밍

7주차. 상관분석과 회귀모형

이혜선 교수
포항공과대학교 산업경영공학과

목차

7주차. 상관분석과 회귀모형

7주차

1차시

상관분석

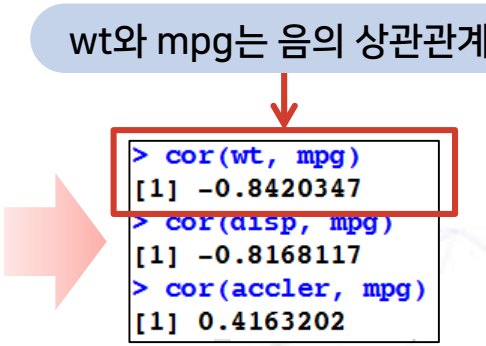
POSTECH

상관분석 : 상관계수

상관계수 : cor(변수1, 변수2)

```
# Correlation coefficent

library(dplyr)

# set working directory
setwd("D:/tempstore/moocr")

# autompg data
car<-read.csv("autompg.csv")
#head(car)
#dim(car)

# subset of car : cyl (4,6,8)
car1<-filter(car, cyl==4 | cyl==6 | cyl==8)
attach(car1)
```

```
#correlation
cor(wt, mpg)
cor(disp, mpg)
cor(accler, mpg)
```



- pearson의 상관계수 : cor(옵션없는 경우)
- kendall 혹은 spearman의 상관계수 : cor(변수1, 변수2, method=c("spearman"))

POSTECH

상관분석 : 상관계수

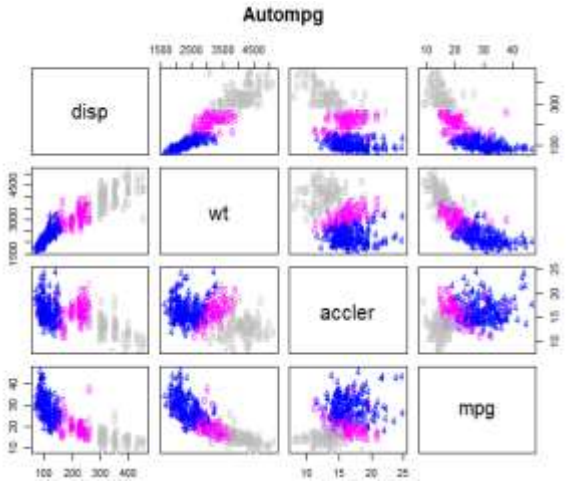
상관계수와 산점도

```
# pariwise plot
# new variable lists
vars1<-c("disp", "wt", "accler", "mpg")
# pariwise plot
pairs(car1[vars1], main ="Autompg",cex=1, col=as.integer(car1$cyl),p
```

(1)차량 무게와 배기량과는 정비례 관계
(양의 상관계수)

(2)MPG(연비)와 (wt,disp)는 상관성이 높다
(반비례 음의 상관계수)

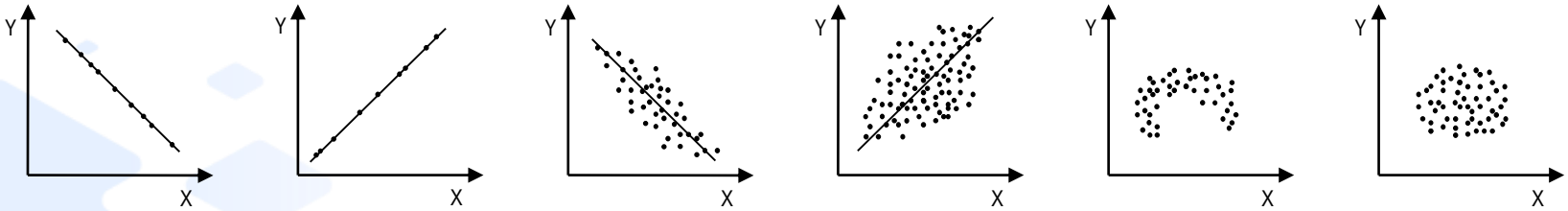
(3)Cylinder별로 색으로 표시
(파란색:4, 핑크:6, 회색:8)



POSTECH

상관분석 : 상관계수와 산점도

- 상관계수(r)은 절대값이 0-1사이 값을 갖는다
  - 절대값이 0에 가까울수록 상관관계가 없다
  - 절대값이 1에 가까울수록 강한 상관성이 있다



(a) r = -1 (b) r = 1 (c) r = -0.7 (d) r = 0.5 (e) r = 0 (f) r = 0

POSTECH

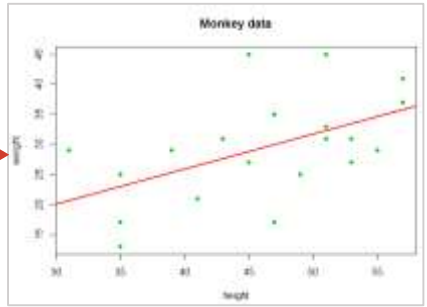
통계치와 그래프 : 주의!!

통계치와 그래프 – Monkey 데이터 + King Kong 한 마리

```
# correlation coefficients
cor(height, weight)
```

```
> cor(height, weight)
[1] 0.5267801
```

```
# scatterplot for weight and height
par(mfrow=c(1, 1))
plot(height, weight, pch=16, col=3,main="Monkey data")

# add the best fit linear line (lec4_2.r)
abline(lm(weight~height), col="red", lwd=2, lty=1)
```



weight와 height간 상관계수는 0.53으로 높지 않다

ID	height	weight
1	55	29
2	45	27
3	35	17
4	39	29
5	53	31
6	41	21
7	51	31
8	35	13
9	57	37
10	57	41
11	45	45
12	47	35
13	35	25
14	49	25
15	43	31
16	51	33
17	31	29
18	53	27
19	47	17
20	51	45

출처 Https://www.gettyimagesbank.com/176668388
Https://www.gettyimagesbank.com/509351302

POSTECH

통계치와 그래프 : 주의!!

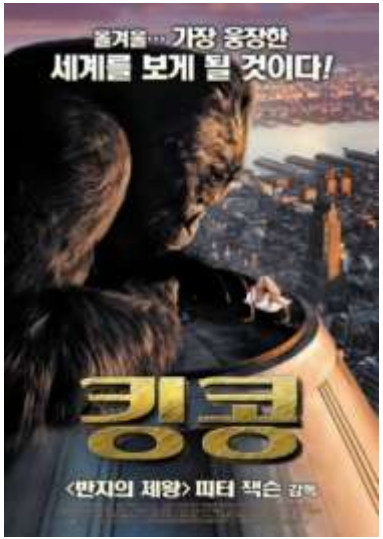
☑ 통계치와 그래프 – Monkey 데이터 + King Kong 한 마리

```
## Monkey data + Kingkong
monkey1<-read.csv("monkey_k.csv")
head(monkey1)
dim(monkey1)
attach(monkey1)
```

```
# correlation coefficients
cor(height, weight)
```

```
> cor(height, weight)
[1] 0.940375
```

상관계수 : 0.94



출처 https://ko.wikipedia.org/wiki/media/파일:킹콩_2005년_포스터.jpg

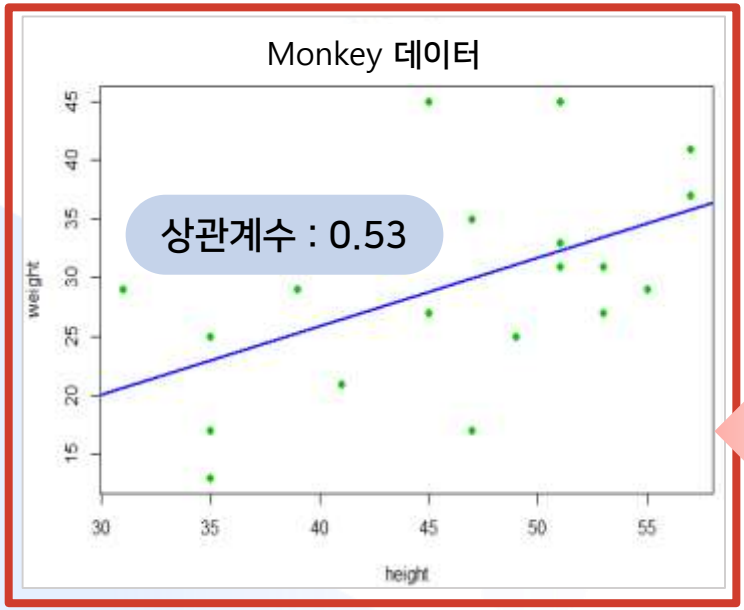
ID	height	weight
1	55	29
2	45	27
3	35	17
4	39	29
5	53	31
6	41	21
7	51	31
8	35	13
9	57	37
10	57	41
11	45	45
12	47	35
13	35	25
14	49	25
15	43	31
16	51	33
17	31	29
18	53	27
19	47	17
20	51	45
21	130	150

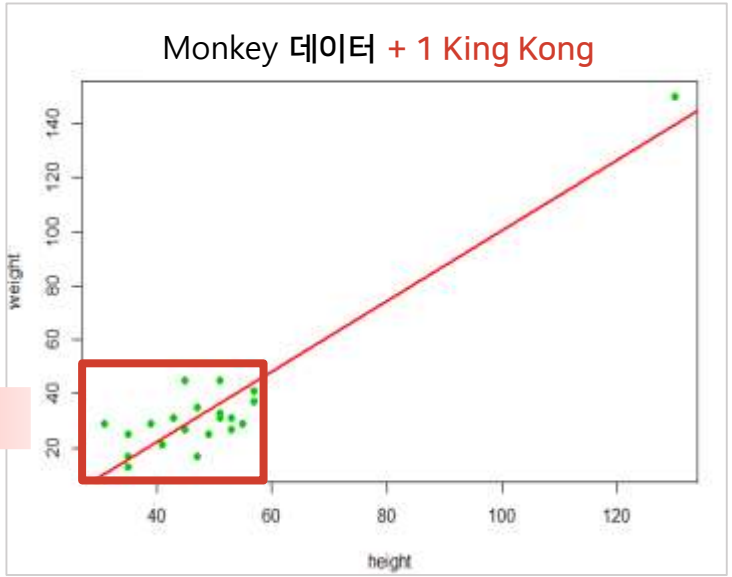
POSTECH

통계치와 그래프 : 주의!!

통계치와 그래프 - Monkey 데이터 + King Kong 한마리

- 한 마리의 킹콩 데이터가 몸무게와 신장의 상관관계에 대한 해석을 완전히 바꿔놓을 수 있다!!





POSTECH

통계치와 그래프 : 주의!!

선형회귀식 - Monkey 데이터

```
# linear model and summary of linear model
m1<-lm(weight~height)
summary(m1)
```

```
> m1<-lm(weight~height)
> summary(m1)

Call:
lm(formula = weight ~ height)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9797  -5.7186  -0.2983   3.9983  16.1797

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.7356    10.2815   0.266    0.793
height        0.5797     0.2205   2.629    0.017 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.573 on 18 degrees of freedom
Multiple R-squared:  0.2775,    Adjusted R-squared:  0.2374
F-statistic: 6.913 on 1 and 18 DF,  p-value: 0.01702
```

선형회귀식
Y(weight)=2.74+0.58X(height)

선형회귀식의 결정계수
R^2=0.27

POSTECH

통계치와 그래프 : 주의!!

선형회귀식 - Monkey 데이터 + King Kong 한마리

```
# linear model and summary of linear model
m2<-lm(weight~height)
summary(m2)
```

```
> m2<-lm(weight~height)
> summary(m2)

Call:
lm(formula = weight ~ height)

Residuals:
    Min      1Q  Median      3Q     Max
-14.219  -7.298  -2.372   8.243  18.706

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -30.2495     5.8203  -5.197 5.13e-05 ***
height        1.3078     0.1085  12.051 2.41e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.646 on 19 degrees of freedom
Multiple R-squared:  0.8843,    Adjusted R-squared:  0.8782
F-statistic: 145.2 on 1 and 19 DF,  p-value: 2.412e-10
```

선형회귀식
Y(weight)=-30.24+1.31X(height)

선형회귀식의 결정계수
R^2=0.88