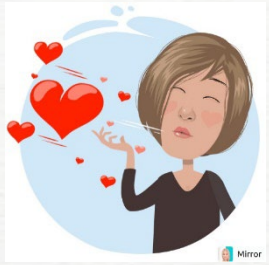
Chapter11. 상관분석&단순선형 회귀분석



목차

I. 상관분석

- 병아리의 성장(몸무게)에 영향을 미치는 요소는 무엇인가?

II. 단순선형 회귀분석

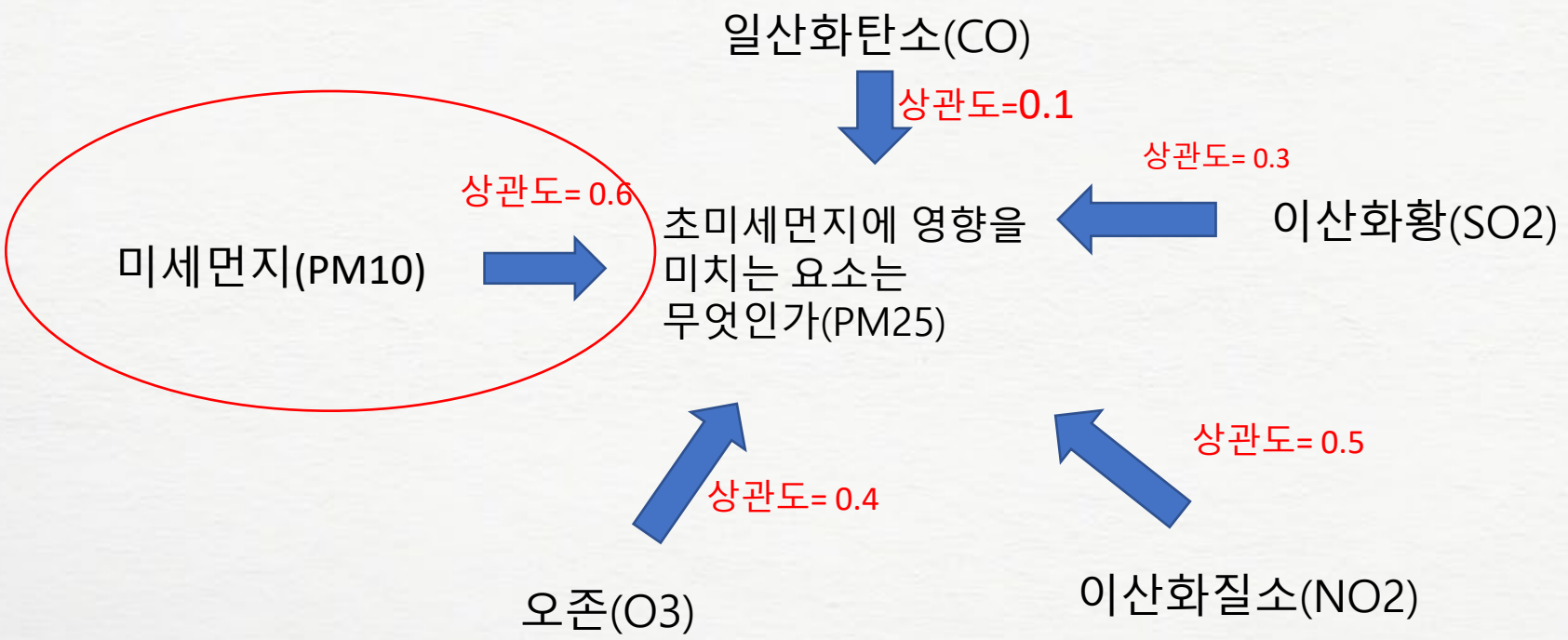
- 주행속도에 대한 제동거리 예측

III. 응용

- 초미세먼지에 영향을 미치는 요소는 무엇인가?
- 초미세먼지량 예측

요약1. 상관분석이란?

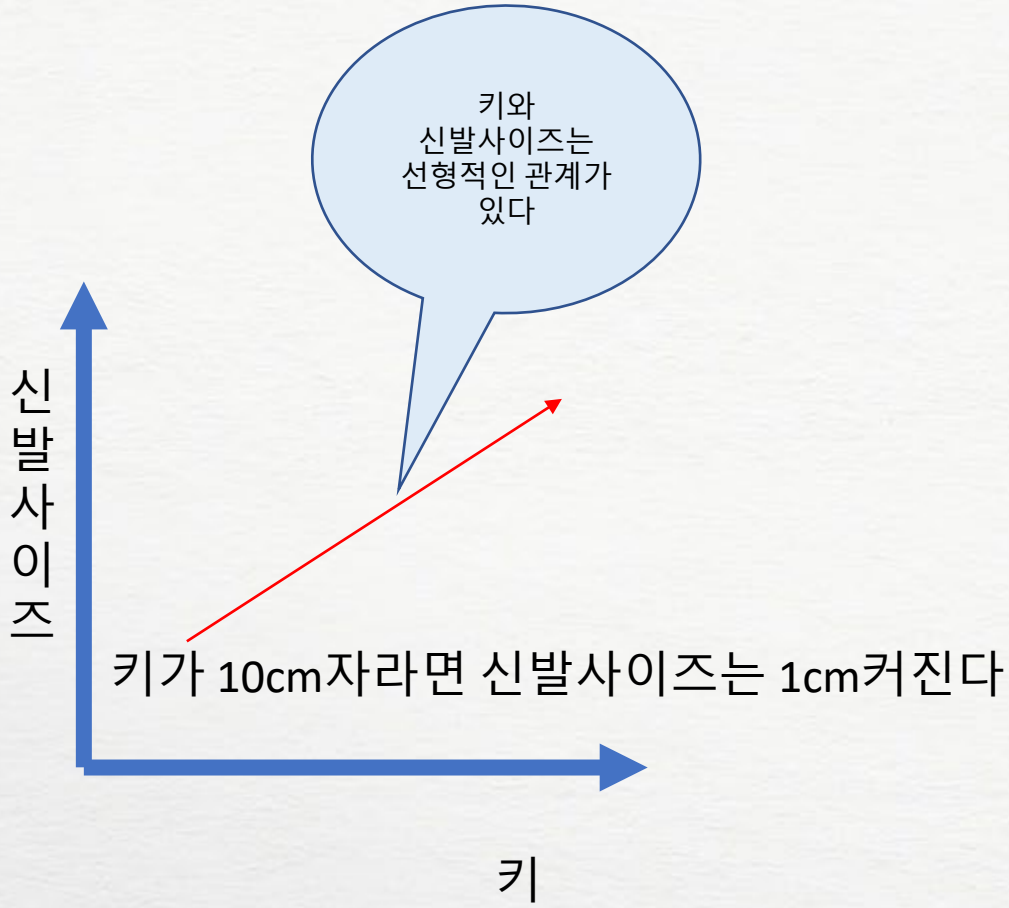
두 요소간 상관이 있는가를 계산
=> (A라는 요소의 값이 커지면 B라는 요소의 값이 커질 때 상관이 있다라고 결정)

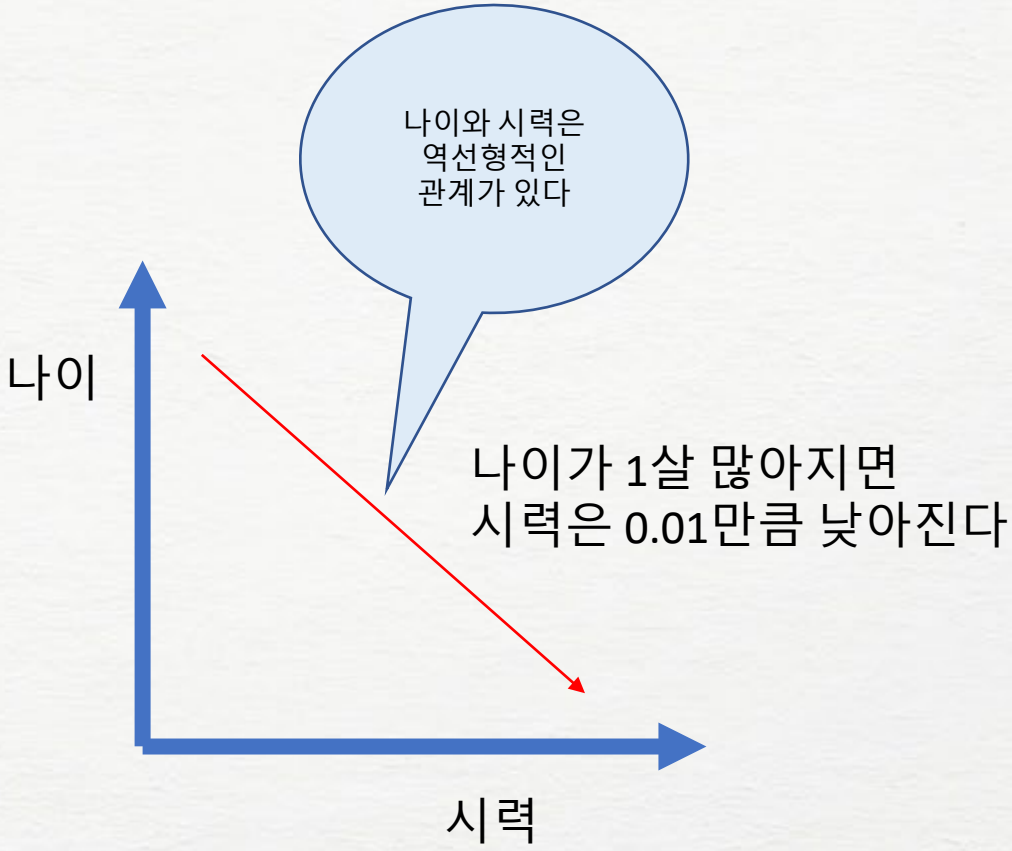


요약1. 상관분석이란?

상관분석(Correlation analysis)

연속형인 두 변수 간에 어떤 선형적인(Linear) 또는
비선형적인(Non-linear) 관계를 갖고 있는지 분석하는 방법

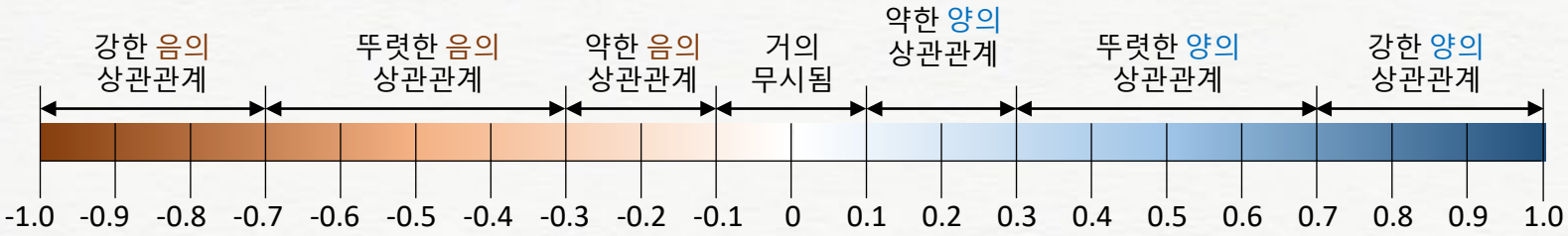




요약2. 상관계수란?

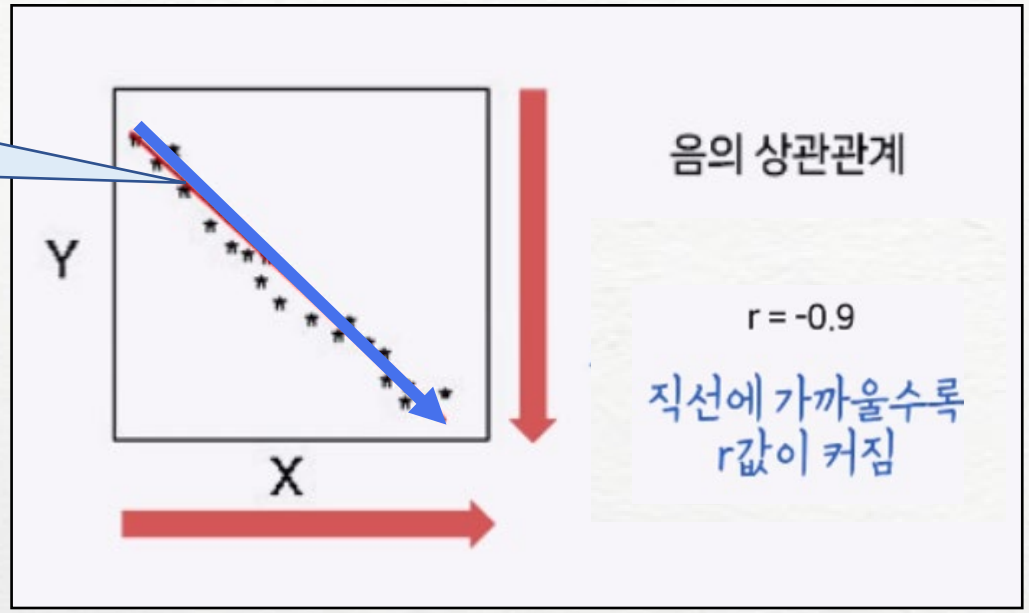
상관계수(Correlation coefficient)

-1~1 사이의 **값**



일반적인 판정기준

r은 기울기



다이어트하는 날짜(X)가 늘어나면 몸무게(Y)는 점점 작아진다.

요약3. cor()함수-상관분석 실시

cor()함수는 각 요소간 서로 서로 상관도를 계산하여 출력한다.

- cor() 함수를 이용해 두 변수간의 상관도를 계산할 수 있음

```
> w_cor <- cor(w_n)     # w_n 데이터 셋으로 상관분석한 결과를 w_cor 변수에 넣음
> w_cor                 # w_cor 상관분석 결과 확인
```

	weight	egg_weight	movement
weight	1.0000000	0.9571693	0.3807186
egg_weight	0.9571693	1.0000000	0.4282457
movement	0.3807186	0.4282457	1.0000000

병아리의 몸무게와 가장 상관도가 높은 요소는 종란무게이다.

	weight (병아리의 몸무게)
egg_weight(종란 무게)	0.9571693
movement(이동거리)	0.3807186

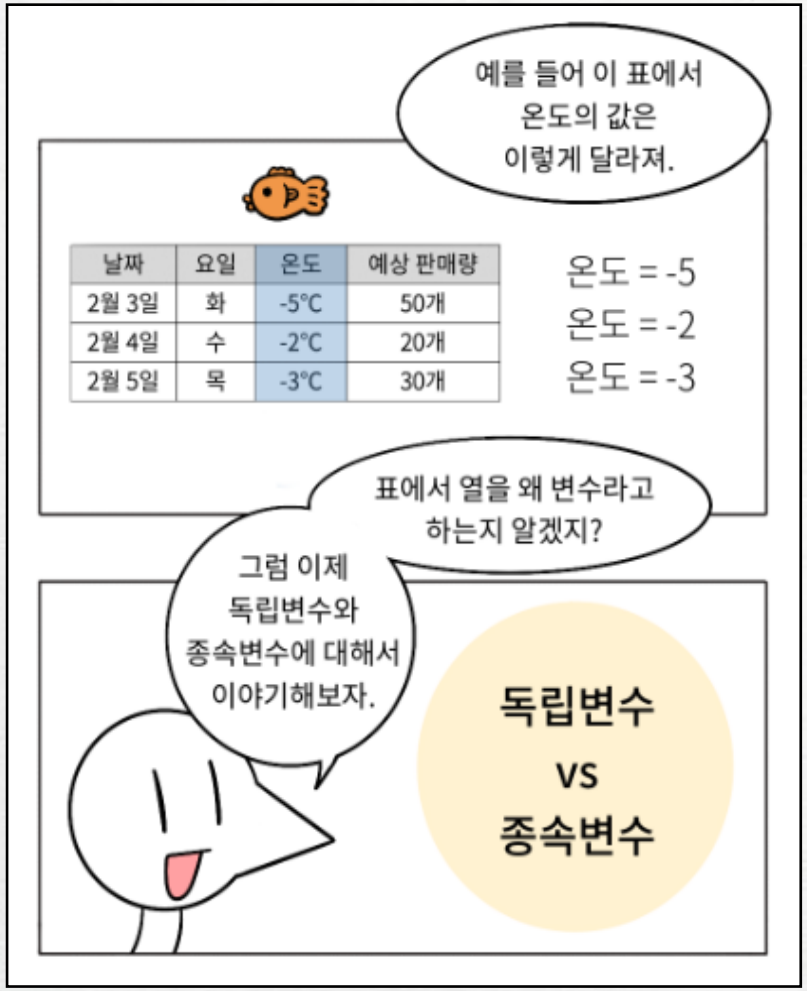
➡ 병아리의 몸무게와 종란 무게는 0.957의 상관도를 나타낸다.

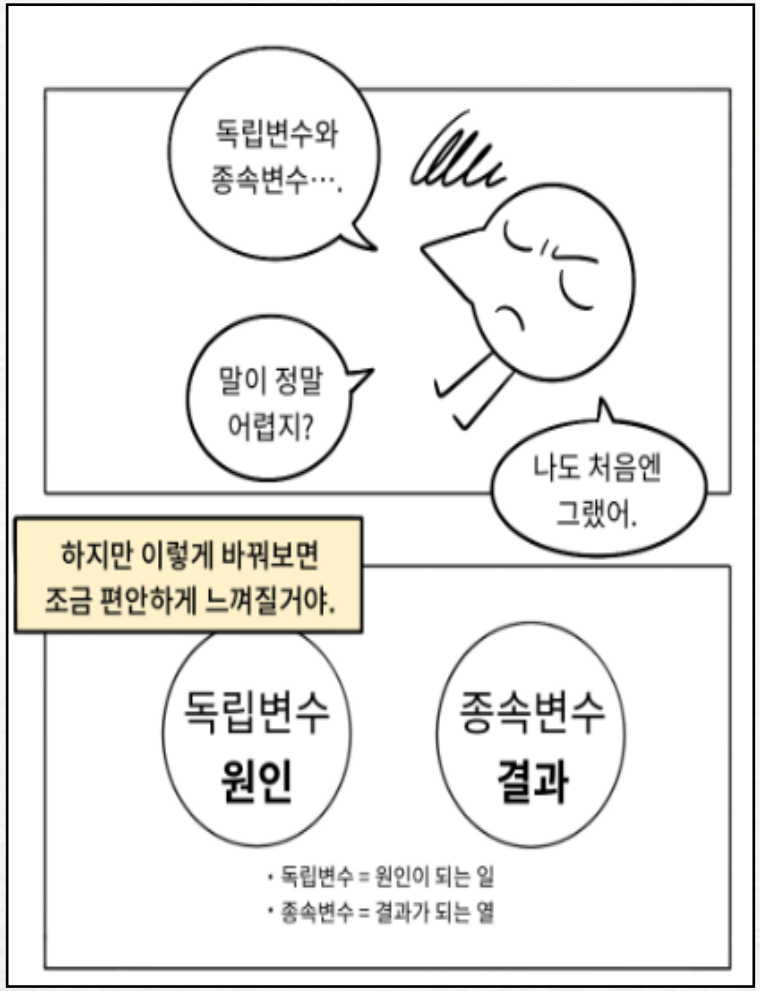
➡ 병아리의 몸무게와 이동거리는 0.38의 상관도를 나타낸다.

요약4. 독립변수와 종속변수

독립변수-> 원인

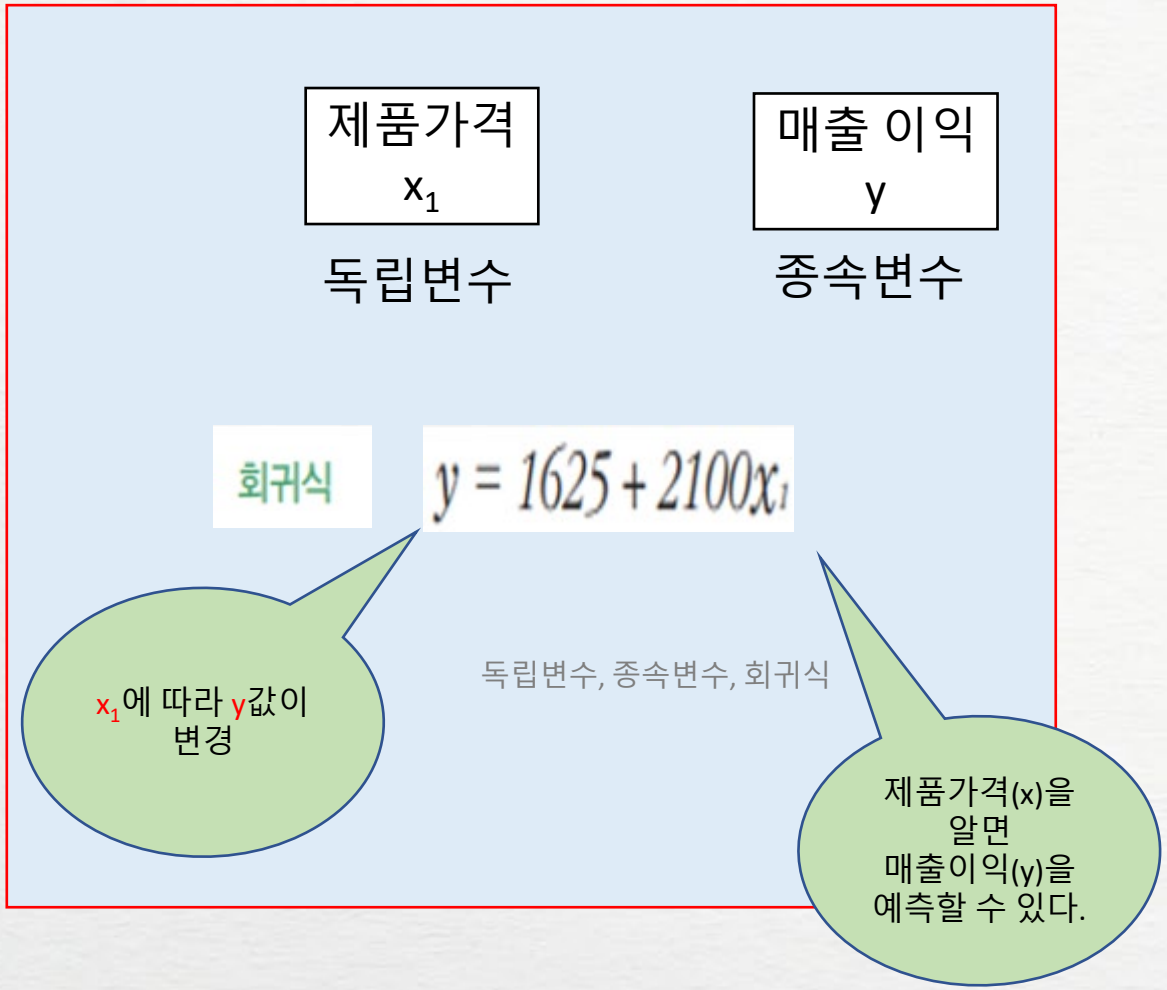
종석변수-> 원인으로 인해 발생한 결과





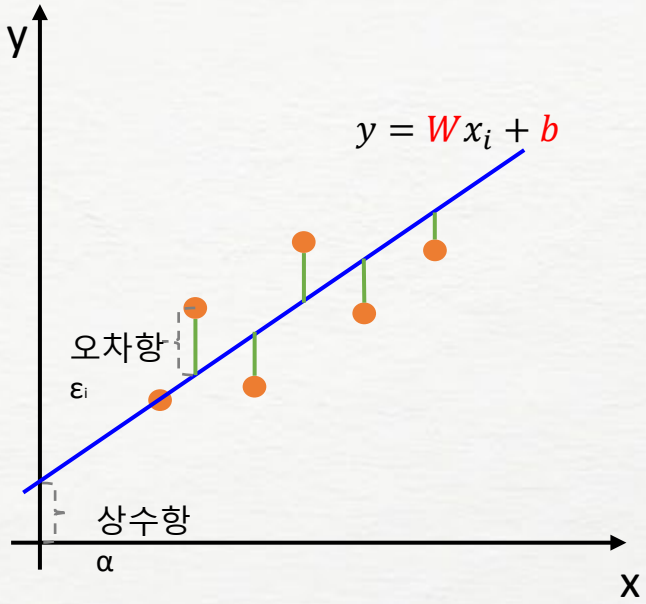
요약5. 단순선형 회귀분석이란?

- **회귀분석(regression analysis)**: 독립변수가 종속변수에 미치는 영향을 파악하여 예측값 도출(W와 b값 도출)



회귀식 $y = Wx_i + b$

x_i 독립변수, y 종속변수

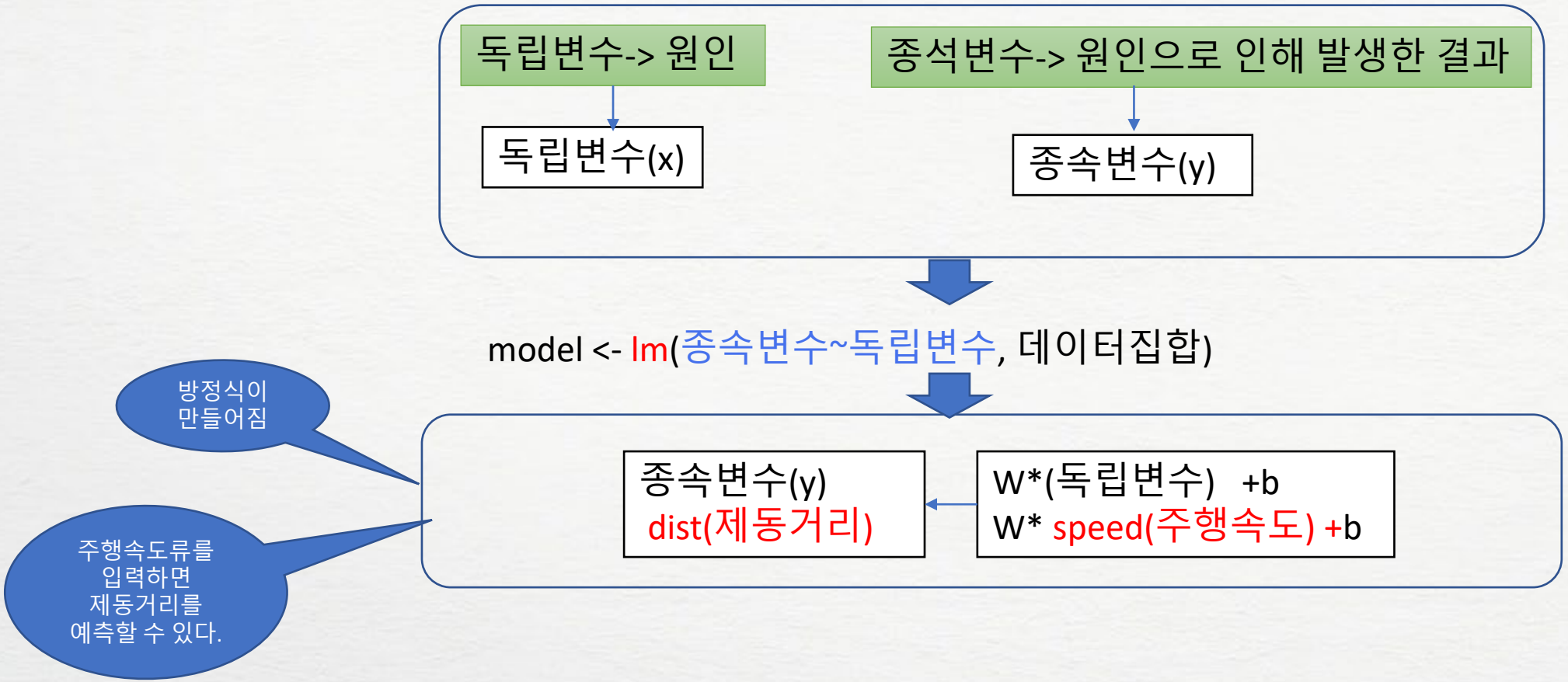


<회귀모형 개념>

요약6. lm()함수-회귀식만들기

회귀식을 구하는 함수-> lm( )

```
model <- lm(종속변수~독립변수, 데이터집합)
model <- lm(dist~speed, cars)
```



요약6. lm()함수-회귀식만들기

종속변수(y): dist(제동거리) | 독립변수(x): speed(주행속도)

```
model <- lm(종속변수~독립변수, 데이터집합)
model <- lm(dist~speed, cars)
```

model

```
Call:
lm(formula = dist ~ speed, data = cars)

Coefficients:
(Intercept)        speed
    -17.579        3.932
```

lm이 방정식을 만드는 방법

주행속도(speed)를 알면 제동거리(dist)를 예측할 수 있다.

y(dist) = 3.932 x(speed) + -17.579

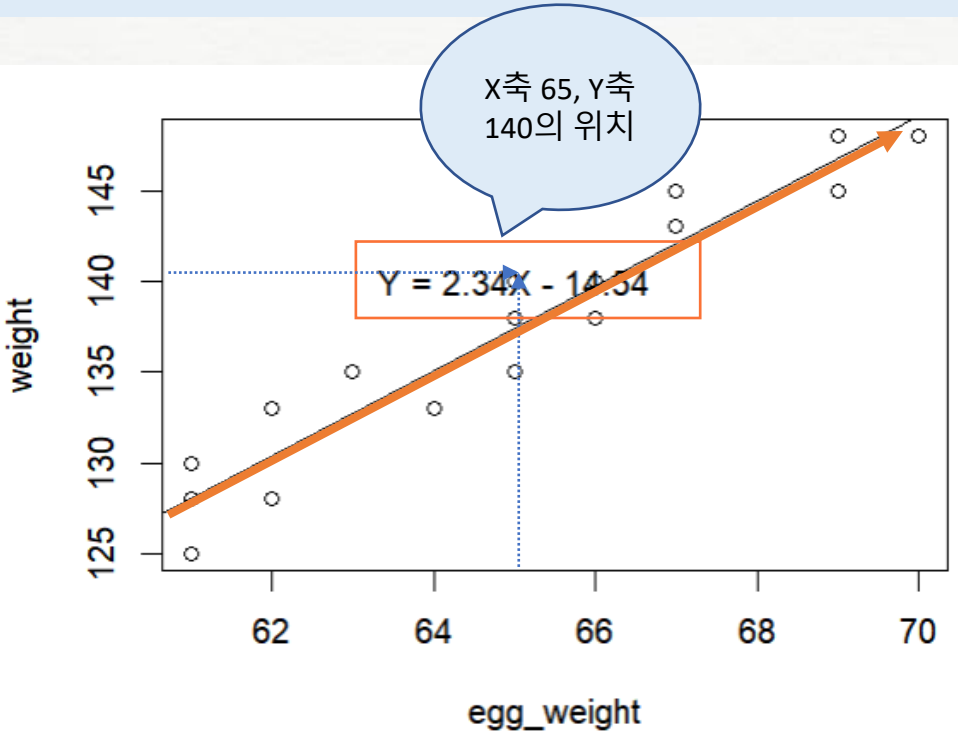
y = W * x + b

W:3.932
b: -17.579

요약7. 회귀식 표기

- 회귀선을 그리고 회귀식을 그래프상에 표기

```
> abline(회귀식)                                    # 회귀선 그리기
> text(x = 65, y = 140, label = 'Y = 2.34X − 14.54') # 회귀직선 텍스트로 표시, x와 y값은 그래프 상의 빈 공간에 임의의 점을 선정했음
```



y = Wx + b (W, b 는 상수)

y = 2.34x -14.54 (W, b 는 상수)

요약8. 단순선형 회귀분석을 이용한 예측

Coef()는 회귀식의 상수를 추출하는 함수

회귀식에 종란 무게(x)를 입력하면 병아리의 무게(y)를 예측할 수 있음

y = Wx + b (W, b 는 상수)

```
b<- coef(회귀식)[1]          # 회귀식의 가장 오른쪽상수,b=-14.54
W <- coef(회귀식)[2]         #회귀식의 x상수, W=2.34

egg_weight <- 71             # 종란무게(egg_weight )=71
weight <- W*egg_weight+ b    #weight<-2.34*71 -14.54
weight                       # 병아리의 무게
```

y = 2.34x -14.54

y = 2.34x -14.54

y = 2.34x -14.54

y = 2.34*71 -14.54

```
egg_weight
  151.3891
```

종란무게가 71이면 병아리의 무게는 151.3891이 될 것이다.

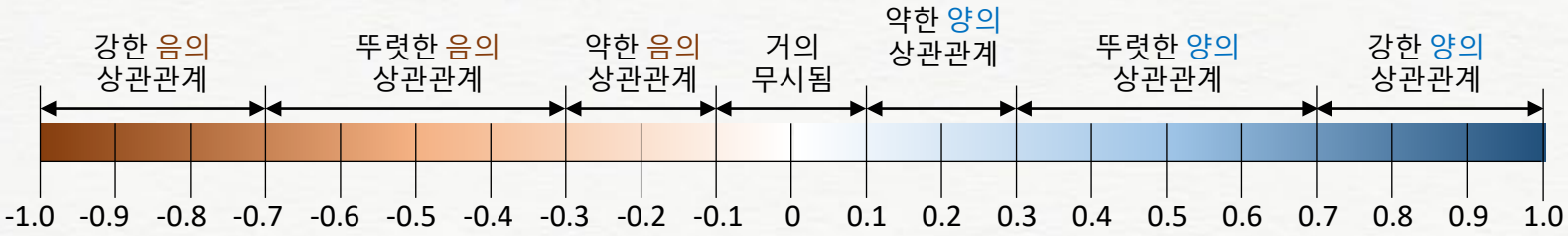
0. 소스이해문제

문제1. 상관계수란?

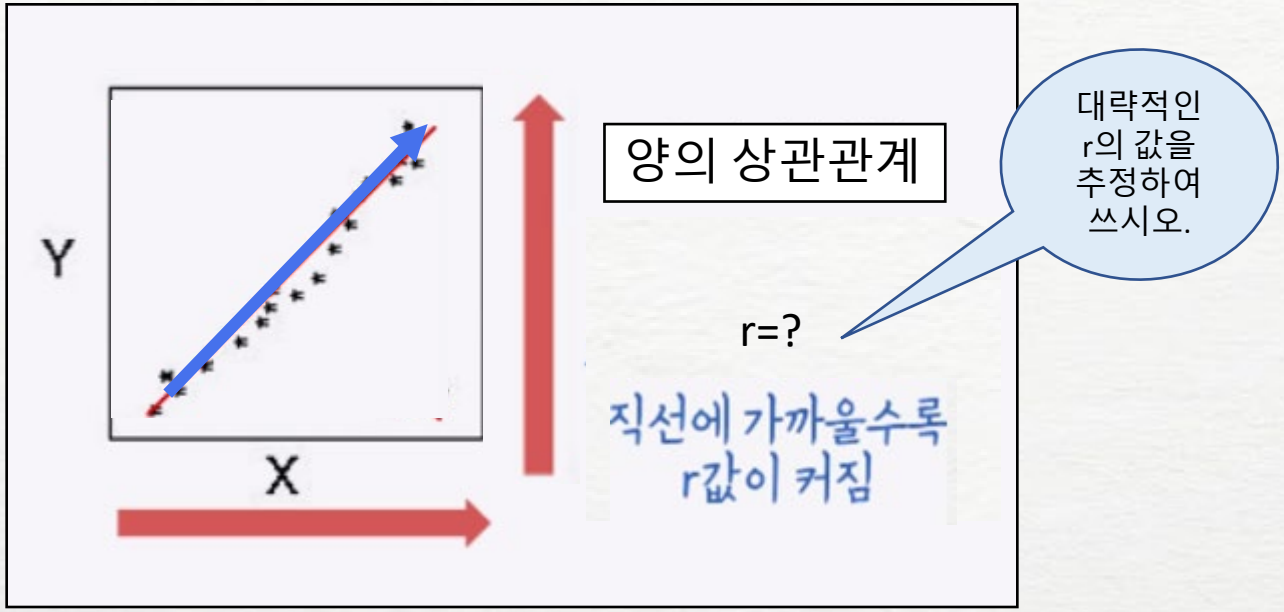
문제1_r값 쓰기

상관계수(Correlation coefficient)

-1~1 사이의 **값**



일반적인 판정기준



문제2. 상관분석 실시

문제2. ?를 쓰시오.

- cor() 함수를 이용해 두 변수간의 상관도를 계산할 수 있음

```
> w_cor <- cor(w_n)     # w_n 데이터 셋으로 상관분석한 결과를 w_cor 변수에 넣음
> w_cor                 # w_cor 상관분석 결과 확인
              weight egg_weight  movement      food
weight     1.0000000  0.9571693 0.3807186 0.8775735
egg_weight 0.3471693  1.0000000 0.4282457 0.8081467
movement   0.3807186  0.4282457 1.0000000 0.3190107
food       0.8775735  0.8081467 0.3190107 1.0000000
```

병아리의 몸무게와 가장 연관도가 높은 요소는 ?이다.

weight (병아리의 몸무게)

egg_weight(종란 무게)	0.3471693
Movement(이동거리)	0.3807186
food(섭취량)	0.8775735

➡ 병아리의 몸무게와 종란 무게는 0.3471693의 상관도를 나타낸다.

➡ 병아리의 몸무게와 이동거리는 0.38의 상관도를 나타낸다.

➡ 병아리의 몸무게와 섭취량은 0.87의 상관도를 나타낸다.

문제3. 회귀식 구하기

종속변수(y): weight(병아리무게)

독립변수(x): egg_weight(종란무게)

문제3.
(가)~(마)를
쓰시오.

```
>W_nmodel <- lm(  (가)  , w_n)    # 회귀식 구하기
>W_nmodel
```

```
Call:
lm(formula = weight ~ egg_weight, data = w_n)

Coefficients:
(Intercept)   egg_weight
    -14.548        2.337
```

y= (나) *x + (다)

y = W *x + b

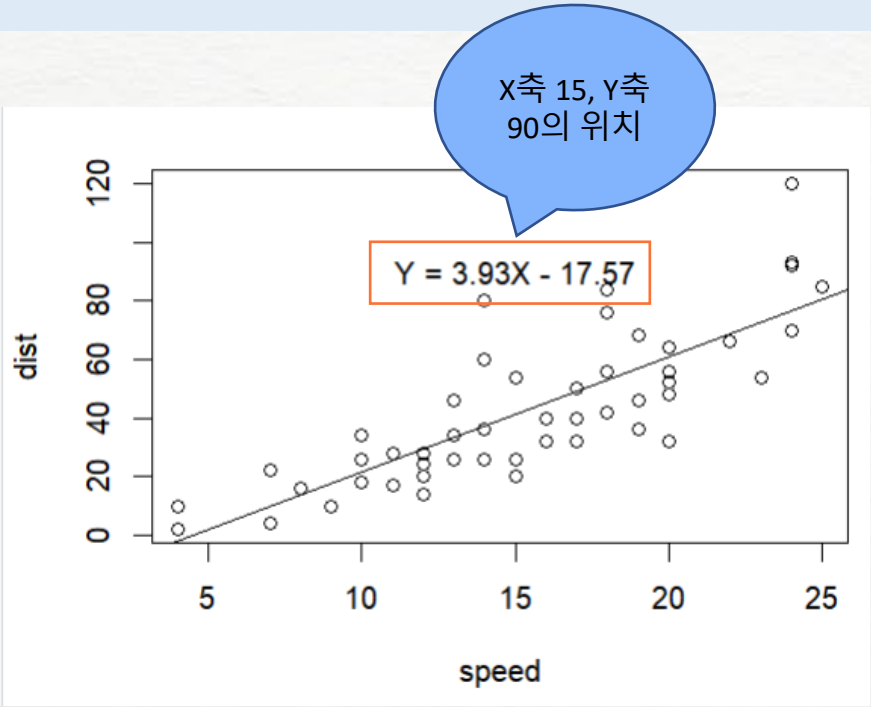
W=(라)
b=(마)

문제4. 회귀식 표기

문제4.(가)~(다)를 쓰시오.

- 회귀식을 그래프상에 표기

```
>text(x = (가), y = (나), label = '  (다)  ') # 회귀직선 텍스트로 표시, x와 y값은 그래프 상의 빈 공간에 임의의 점을 선정했음
```



y = Wx + b (W, b 는 상수)

y = 3.93x -17.57 (W, b 는 상수)

문제5. 단순선형 회귀분석을 이용한 예측

- 회귀식에 의하여 주행속도를 입력하면 제동거리를 예측할 수 있음

```
>b <- coef(model)[1]
>W <- coef(model)[2]

>speed <- 100
>dist <- W*speed + b
>dist
 (가)
```

```
#b=-17
#W=3.93

# 주행속도(speed=100)
#거리(dist)=W100+b
# 제동거리
```

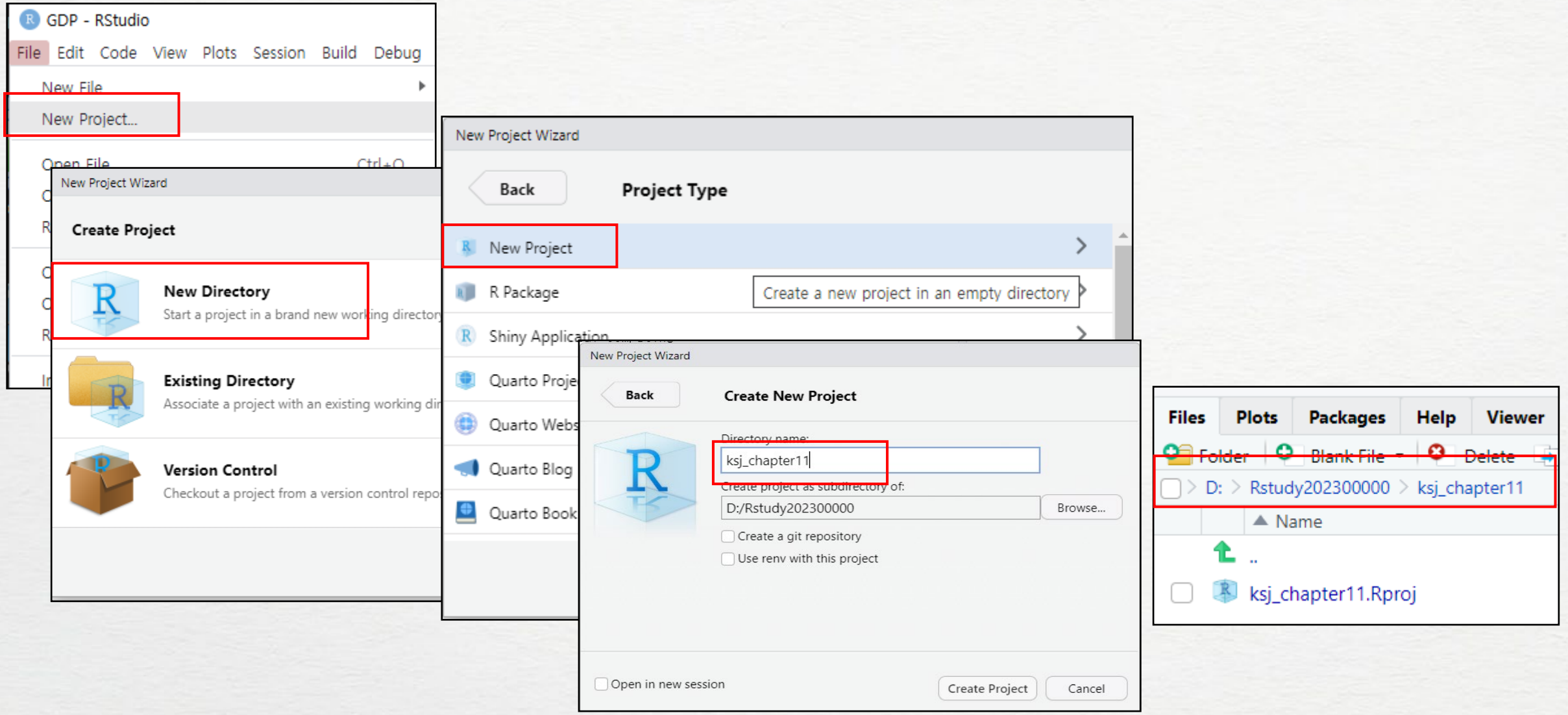
y = 3.93x - 17 (W, b 는 상수)
y = 3.93x - 17 (W, b 는 상수)

y = 3.93x - 17(W, b 는 상수)

y = 3.93*100 - 17(W, b 는 상수)

문제5. 속도가 100인 경우의 제동거리 (가)를 쓰시오.

* 프로젝트 시작


GDP - RStudio
File Edit Code View Plots Session Build Debug
New File
New Project...
Open File Ctrl+O
New Project Wizard
Create Project
New Directory
Start a project in a brand new working directory
Existing Directory
Associate a project with an existing working dir
Version Control
Checkout a project from a version control repo
New Project Wizard
Back
Project Type
New Project
R Package
Create a new project in an empty directory
Shiny Application
Quarto Proje
Quarto Webs
Quarto Blog
Quarto Book
New Project Wizard
Back
Create New Project
Directory name:
ksj_chapter11
Create project as subdirectory of:
D:/Rstudy202300000
Browse...
Create a git repository
Use renv with this project
Open in new session
Create Project
Cancel
Files Plots Packages Help Viewer
Folder
Blank File
Delete
D: > Rstudy202300000 > ksj_chapter11
Name
..
ksj_chapter11.Rproj

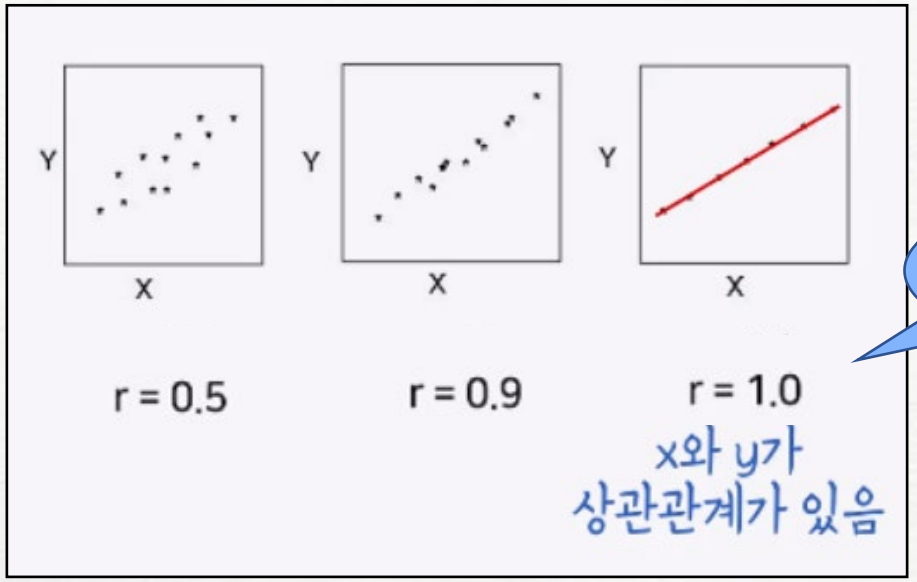
I. 상관분석

20

1. 상관분석이란?

상관계수(Correlation coefficient) → 두 변수 간의 연관된 정도를 나타내는 값

-1~1 사이의 **값** → **(-)부호일 경우 반비례 관계인 음의 상관관계**
(+)부호일 경우 비례 관계인 양의 상관관계



r은 상관계수

2. 병아리의 성장(몸무게)에 영향을 미치는 요소 확인을 위한 상관분석

2.1 상관분석 개요

- 분석 목적 : 병아리의 성장(몸무게)에 영향을 미치는 요소 확인
- ch5-1.csv 데이터 셋의 경우 부화한 지 1주일 된 병아리 몸무게(weight), 종란 무게(egg_weight), 하루 평균 이동거리(movement), 하루 평균 사료 섭취량(food) 데이터가 포함되어 있으며 총 5개의 열(변수)과 30개의 행으로 구성되어있음

ch5-1.csv

chick_nm	weight	egg_weigh	movemen	food
a01	140	65	146	14
a02	128	62	153	12
a03	140	65	118	13
a04	135	65	157	13
a05	145	69	157	13
a06	138	65	143	13
a07	125	61	110	11

병아리 몸무게(weight)

종란 무게(egg_weight)

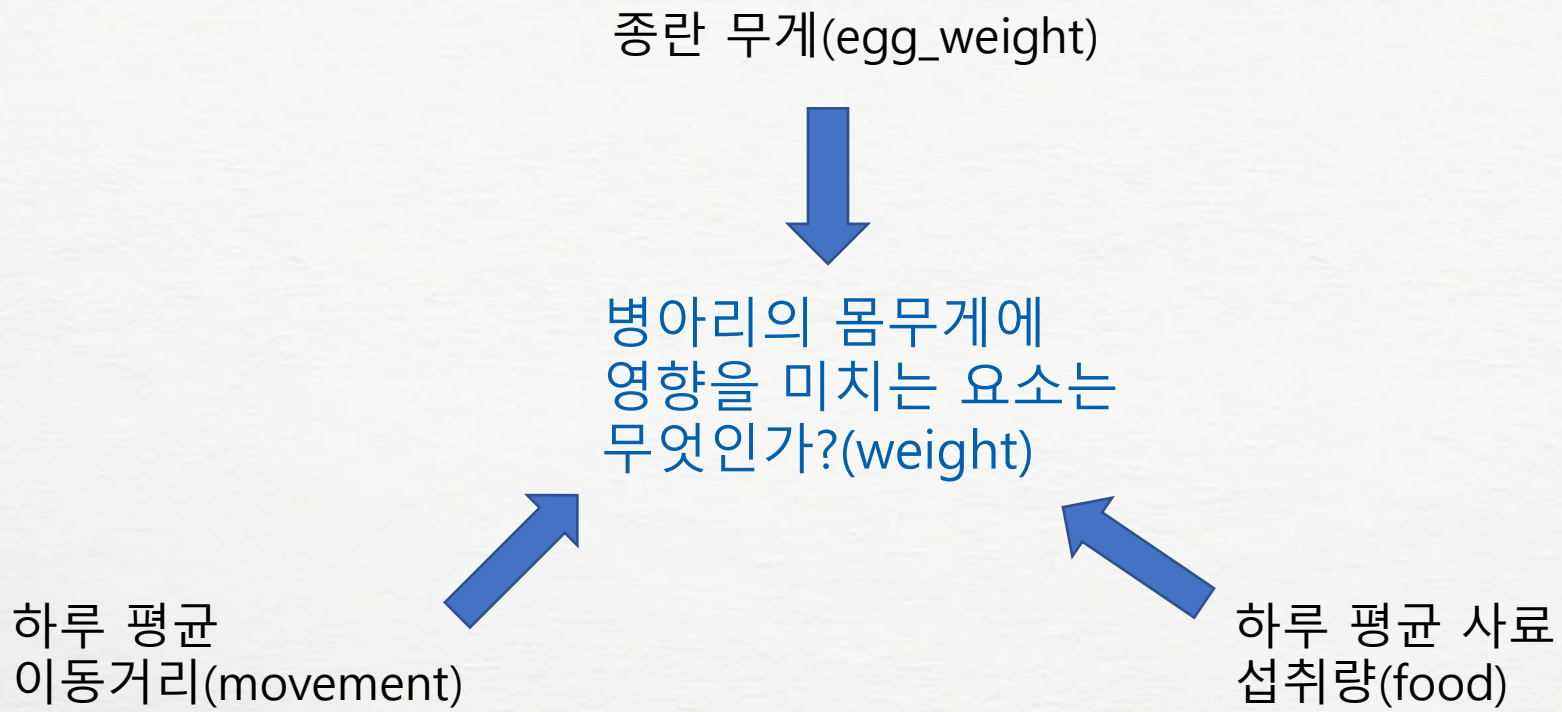
하루 평균 이동거리(movement)

하루 평균 사료 섭취량 (food)

2. 병아리의 성장(몸무게)에 영향을 미치는 요소 확인을 위한 상관분석

2.1 상관분석 개요

- 분석 목적 : 병아리의 성장(몸무게)에 영향을 미치는 요소 확인
- ch5-1.csv 데이터 셋의 경우 부화한 지 1주일 된 병아리 몸무게(weight), 종란 무게(egg_weight), 하루 평균 이동거리(movement), 하루 평균 사료 섭취량(food) 데이터가 포함되어 있으며 총 5개의 열(변수)과 30개의 행으로 구성되어있음



2. 병아리의 성장(몸무게)에 영향을 미치는 요소 확인을 위한 상관분석

2.1 상관분석 개요

```
w_n <- w[,2:5]
```

chick_nm	weight	egg_weigh	movemen	food
a01	140	65	146	14
a02	128	62	153	12
a03	140	65	118	13
a04	135	65	157	13
a05	145	69	157	13
a06	138	65	143	13
a07	125	61	110	11

```
w_cor <- cor(w_n)  #상관분석
```

w_cor

	weight	egg_weight	movement	food
weight	1.0000000	0.9571693	0.3807186	0.8775735
egg_weight	0.9571693	1.0000000	0.4282457	0.8081467
movement	0.3807186	0.4282457	1.0000000	0.3190107
food	0.8775735	0.8081467	0.3190107	1.0000000

상관분석 결과
몸무게와 상관관계가
가장 높은 변수는 종란
무게

종란 무게(egg_weight)

0.9571693

병아리의 몸무게에
영향을 미치는 요소는
무엇인가?(weight)

0.875735

하루 평균 사료
섭취량(food)

0.3807186

하루 평균
이동거리(movement)

2. 병아리의 성장(몸무게)에 영향을 미치는 요소 확인을 위한 상관분석

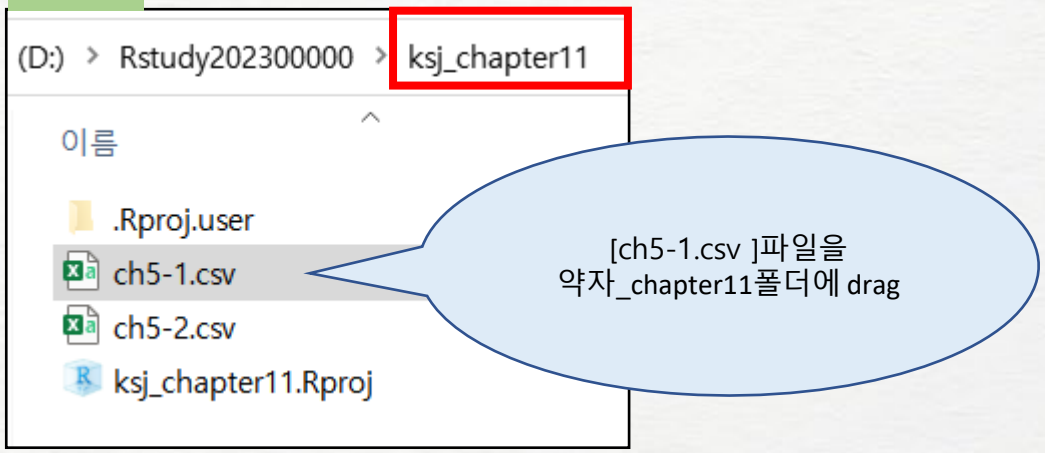
2.2 상관분석 데이터 준비

- ch5-1.csv 데이터 셋의 경우 부화한 지 1주일 된 병아리 몸무게(weight), 종란 무게(egg_weight), 하루 평균 이동거리(movement), 하루 평균 사료 섭취량(food) 데이터가 포함되어 있으며 총 5개의 열(변수)과 30개의 행으로 구성되어있음

ch5-1.csv

chick_nm	weight	egg_weigh	movemen	food
a01	140	65	146	14
a02	128	62	153	12
a03	140	65	118	13
a04	135	65	157	13
a05	145	69	157	13
a06	138	65	143	13
a07	125	61	110	11

코드



2. 병아리의 성장(몸무게)에 영향을 미치는 요소 확인을 위한 상관분석

2.2 상관분석 데이터 준비

- ch5-1.csv 데이터 셋의 경우 부화한 지 1주일 된 병아리 몸무게(weight), 종란 무게(egg_weight), 하루 평균 이동거리(movement), 하루 평균 사료 섭취량(food) 데이터가 포함되어 있으며 총 5개의 열(변수)과 30개의 행으로 구성되어있음

코드

```
> w <- read.csv("ch5-1.csv", header = TRUE)   # 엑셀파일을 불러와서 w에 저장
> head(w)
```

	chick_nm	weight	egg_weight	movement	food
1	a01	140	65	146	14
2	a02	128	62	153	12
3	a03	140	65	118	13
4	a04	135	65	157	13
5	a05	145	69	157	13
6	a06	138	65	143	13

```
> str(w)                                          # w의 변수명 확인
$ chick_nm  : Factor w/ 30 levels "a01","a02","a03",..: 1 2 3 4 5 6 7 8 9 10 ...
$ weight    : int  140 128 140 135 145 138 125 148 133 145 ...
$ egg_weight: int  65 62 65 65 69 65 61 69 64 69 ...
$ movement  : int  146 153 118 157 157 143 110 159 133 174 ...
$ food      : int  14 12 13 13 13 13 11 15 11 13 ...
```

2. 병아리의 성장(몸무게)에 영향을 미치는 요소 확인을 위한 상관분석

2.3 상관분석 실시

- 첫 번째 열(chick_nm)이 문자라 상관분석이 되지 않기 때문에 첫 번째 열을 제외하고 별도의 데이터 셋 구성

코드

```
> w_n <- w[,2:5]  # w 데이터 셋에서 2~5열 데이터만 가져오기
> head(w_n)       # 첫번째 열은 제외됨
  weight egg_weight movement food
1    140         65      146   14
2    128         62      153   12
3    140         65      118   13
4    135         65      157   13
5    145         69      157   13
6    138         65      143   13
```

w_n <- w[,2:5]

chick_nm	weight	egg_weigh	movemen	food
a01	140	65	146	14
a02	128	62	153	12
a03	140	65	118	13
a04	135	65	157	13
a05	145	69	157	13
a06	138	65	143	13
a07	125	61	110	11

2. 병아리의 성장(몸무게)에 영향을 미치는 요소 확인을 위한 상관분석

2.3 상관분석 실시

- cor() 함수를 이용해 상관분석을 실시할 수 있음

코드

```
> w_cor <- cor(w_n)     # w_n 데이터 셋으로 상관분석한 결과를 w_cor 변수에 넣음
> w_cor                 # w_cor 상관분석 결과 확인
```

	weight	egg_weight	movement	food
weight	1.0000000	0.9571693	0.3807186	0.8775735
egg_weight	0.9571693	1.0000000	0.4282457	0.8081467
movement	0.3807186	0.4282457	1.0000000	0.3190107
food	0.8775735	0.8081467	0.3190107	1.0000000

	weight	
egg_weight	0.9571693	→ 병아리의 몸무게와 종란 무게는 0.957의 상관도를 나타낸다.
movement	0.3807186	→ 병아리의 몸무게와 이동거리는 0.38의 상관도를 나타낸다.
food	0.8775735	→ 병아리의 몸무게와 섭취량은 0.87의 상관도를 나타낸다.

2. 병아리의 성장(몸무게)에 영향을 미치는 요소 확인을 위한 상관분석

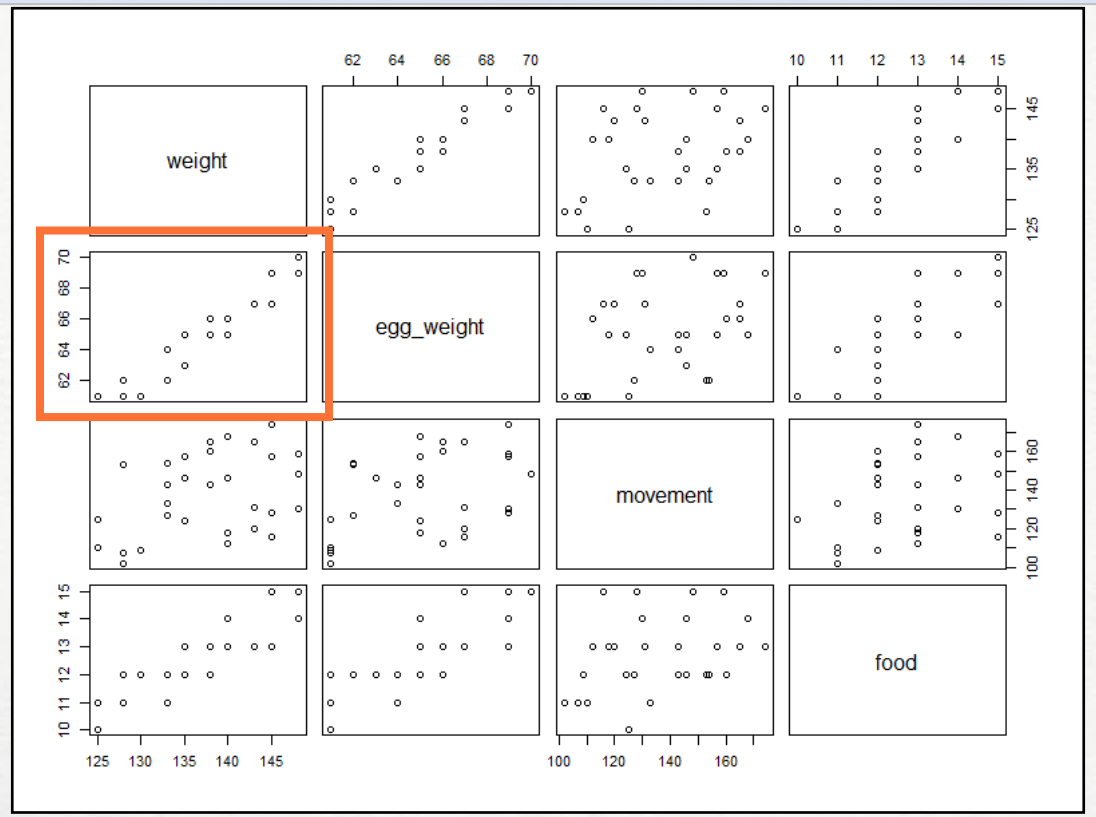
2.4 상관분석 결과표현

- 상관분석의 결과는 주로 산점도(Scatter plot)로 나타냄(corrplot패키지 이용)

코드

```
> plot(w_n)                    # w_n 데이터 셋을 산점도로 표현
```

병아리 몸무게(weight)와 가장 상관관계가 큰 변수는 종란 무게(egg_weight)로 0.957의 상관도



<기본 상관관계 산점도>

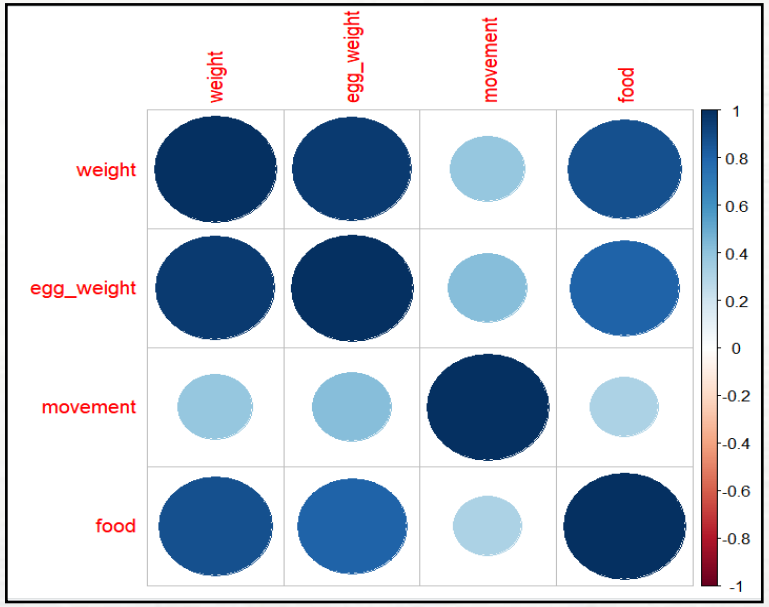
2. 병아리의 성장(몸무게)에 영향을 미치는 요소 확인을 위한 상관분석

2.4 상관분석 결과표현

- 상관분석의 결과는 주로 산점도(Scatter plot)로 나타냄(corrplot패키지 이용)

코드

```
> install.packages("corrplot")          # corrplot 패키지 설치
> library(corrplot)                     # corrplot 패키지 불러오기
> corrplot(w_cor)                       # 상관분석 결과인 w_cor을 corrplot 패키지로 실행
```



<corrplot 패키지 활용 상관분석 결과1>

30/16

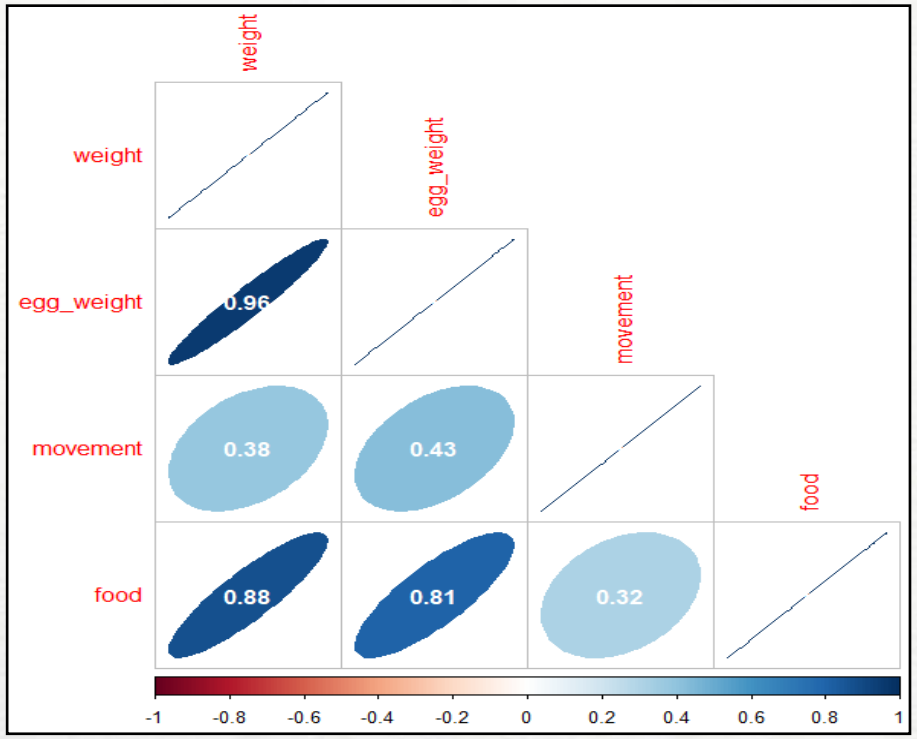
2. 병아리의 성장(몸무게)에 영향을 미치는 요소 확인을 위한 상관분석

2.4 상관분석 결과표현

- 상관분석의 결과는 주로 산점도(Scatter plot)로 나타냄(corrplot패키지 이용)

코드

```
> corrplot(w_cor, method = "ellipse",
+          type = "lower",
+          addCoef.col = "white")
```



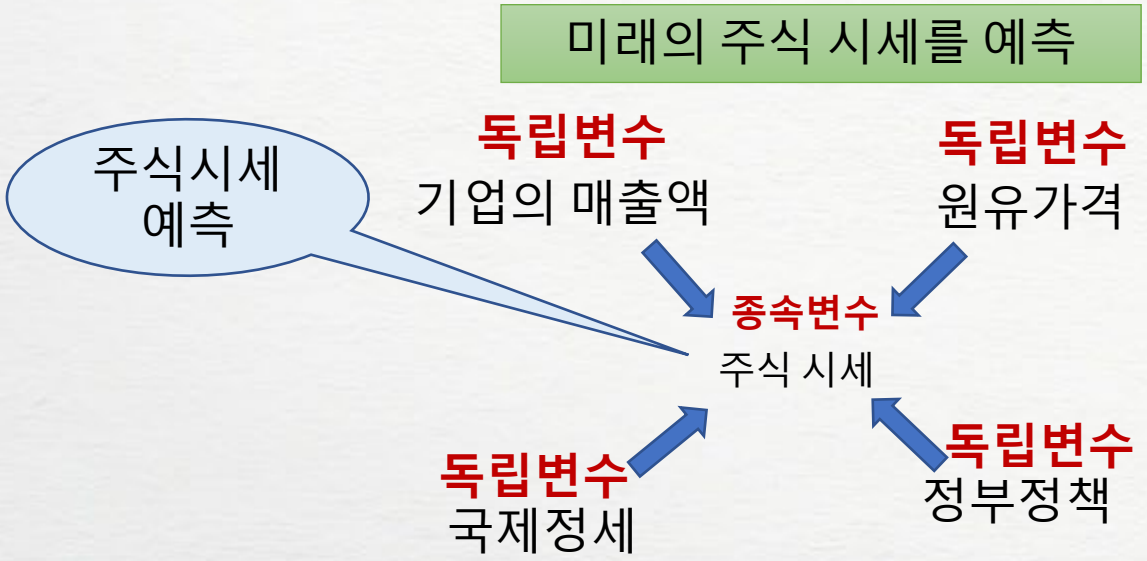
<corrplot 패키지 활용 상관분석 결과2>

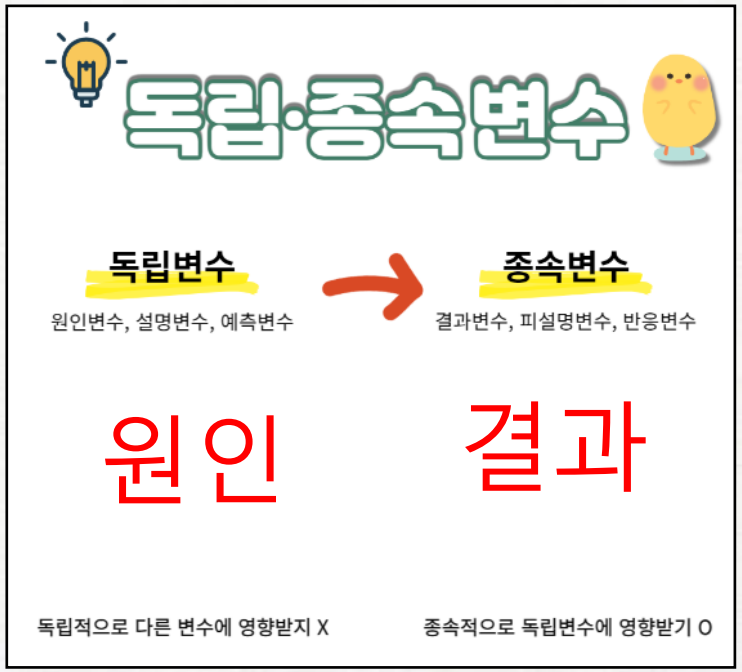
31/
16

Ⅱ. 단순선형 회귀분석

1. 독립변수와 종속변수

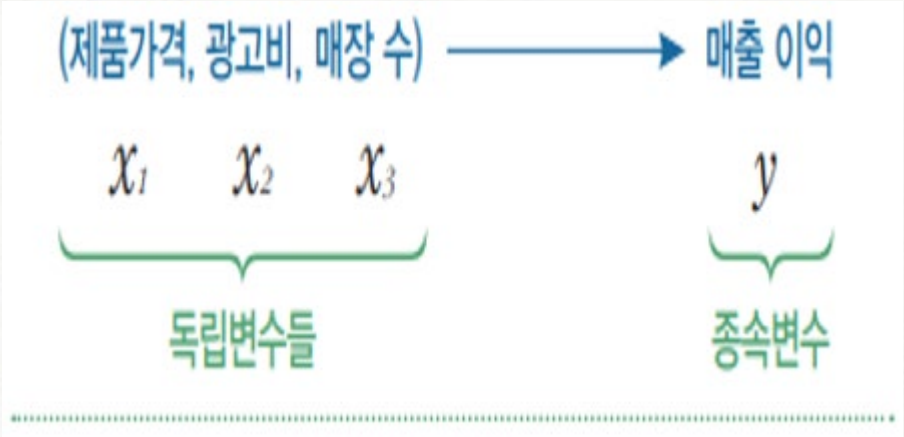
주식 시세 ➡ 기업의 매출액, 원유가격, 국제정세, 정부정책 발표 등 매우 많은 요인들에 의해 영향 받음





2. 회귀분석이란?

- **회귀분석(regression analysis)**: 독립변수가 종속변수에 미치는 영향을 파악하여 예측값 도출

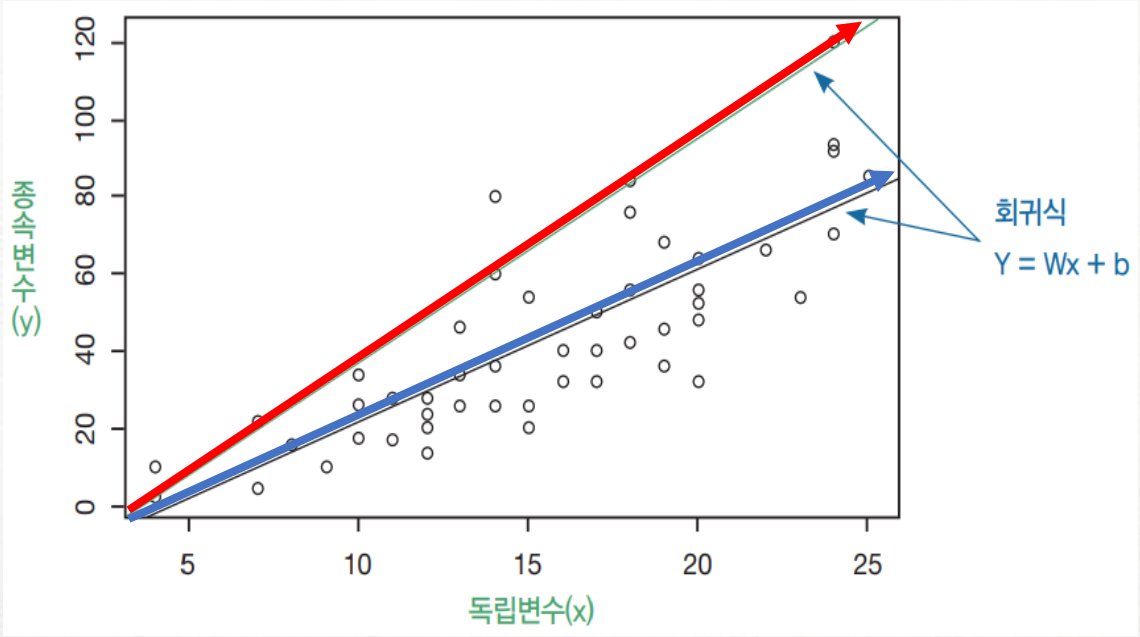


독립변수, 종속변수, 회귀식

3. 단순선형 회귀분석

- 독립변수(x) 자료를 가지고 앞으로의 일, 종속변수 y를 예측하는 문제
- 단순선형회귀식: 1차식

y = Wx + b (W, b 는 상수)



산점도와 단순선형회귀식

4. 주행속도에 대한 제동거리 예측

4.1 데이터 준비

- cars: 내장된 데이터셋

  내장된 데이터셋으로 속도에 대한 제동거리를 나타내는 데이터

코드

```
>head(cars)
```

	speed	dist
1	4	2
2	4	10
3	7	4
4	7	22
5	8	16
6	9	10

4. 주행속도에 대한 제동거리 예측

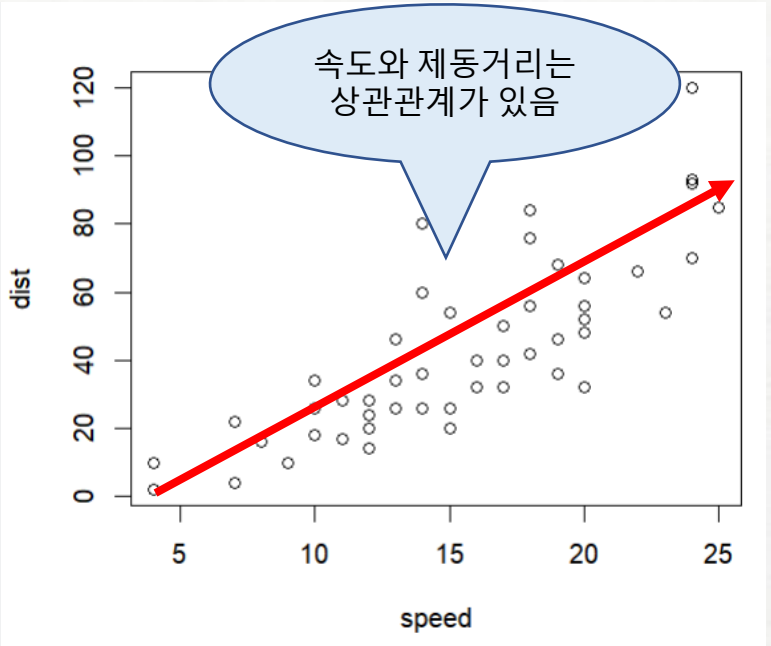
4.2 선형관계확인

- 속도와 제동거리의 선형관계를 확인함(산점도)

	speed	dist
1	4	2
2	4	10
3	7	4
4	7	22
5	8	16
6	9	10

코드

```
>head(cars)
>plot(dist~speed, data=cars)        # 산점도를 통해 선형 관계 확인(x축:speed, y축:dist)
```



4. 주행속도에 대한 제동거리 예측

4.3 회귀모델 구하기

- 회귀모델을 구하기

코드

```
>model <- lm(dist~speed, cars)    # 회귀모델 구하기(lm)
>model
```

```
Call:
lm(formula = dist ~ speed, data = cars)

Coefficients:
(Intercept)        speed
    -17.579        3.932
```

- **dist~speed**

  회귀모델에서 독립변수와 종속변수를 지정하는 것으로, ~를 기준으로 '종속변수~독립변수'의 순서로 지정해야 한다. 여기서 순서가 바뀌면 안 된다.

- **cars**

  회귀모델을 만드는 데 사용할 데이터셋이다. 여기에서는 dist와 speed가 cars의 열이어야 한다.

회귀식 y(dist) = x(speed)*3.932 -17.579

종속변수: dist(제동거리)

회귀식 y(dist) = x(speed)*W + b

독립변수: speed(주행속도)

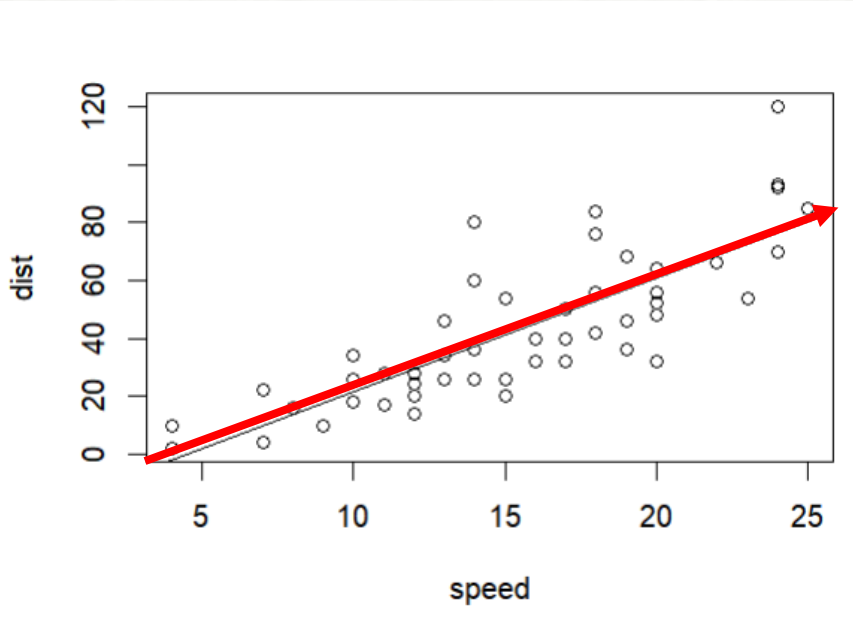
4. 주행속도에 대한 제동거리 예측

4.3 회귀모델 구하기

- 회귀선을 표기

코드

```
>abline(model)                    # 회귀선을 산점도 위에 표시
```



4. 주행속도에 대한 제동거리 예측

4.4 회귀식 구하기

코드

```
>coef(model)[1]
```

```
(Intercept)
  -17.57909
```

b 값 출력

y = Wx + b (a, b 는 상수)

y = Wx -17.579 (W, b 는 상수)

```
>coef(model)[2]
```

```
    speed
3.932409
```

W 값 출력

y = Wx -17.579 (a, b 는 상수)

y = 3.932x -17.579 (W, b 는 상수)

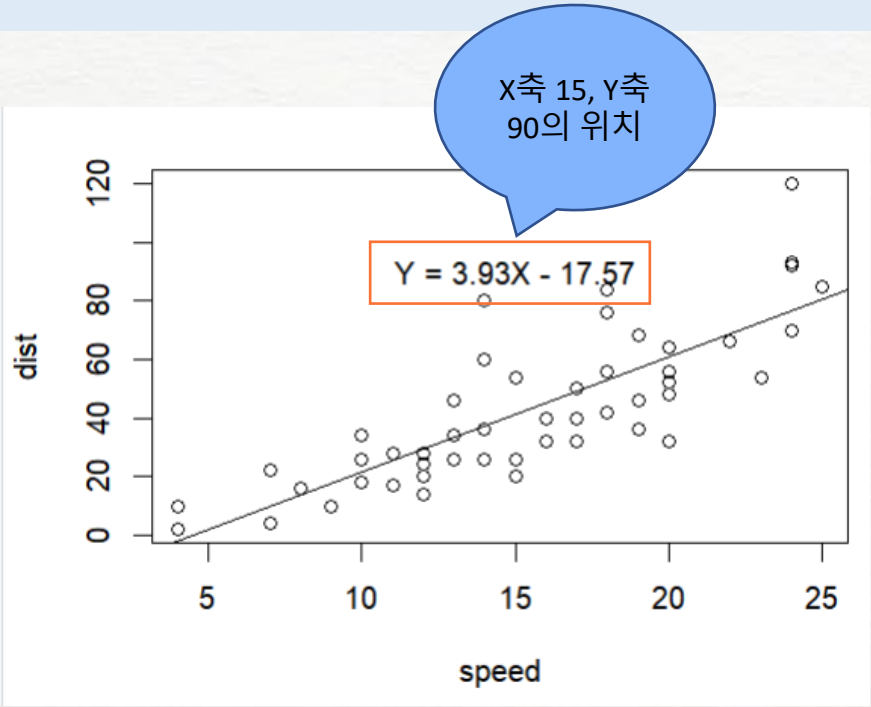
4. 주행속도에 대한 제동거리 예측

4.5 회귀식 표기

- 회귀식을 그래프상에 표기

코드

```
>text(x = 15, y = 90, label = 'Y = 3.93X - 17.57') # 회귀직선 텍스트로 표시, x와 y값은 그래프 상의 빈 공간에 임의의 점을 선정했음
```



y = Wx + b (W, b 는 상수)

y = 3.93x -17.57 (W, b 는 상수)

4. 주행속도에 대한 제동거리 예측

4.6 주행속도에 따른 제동거리 예측

- 회귀식에 의하여 주행속도를 입력하면 제동거리를 예측할 수 있음

코드

```
>b <- coef(model)[1]                #b=-17.57
>W <- coef(model)[2]                #W=3.93

>speed <- 50                        # 주행속도(speed=50)
>dist <- W*speed + b                #거리(dist)=W50+b
>dist                               # 제동거리
```

```
speed
179.0413
```

속도가 50인 경우의 제동거리는 179.0413

y = Wx + b (W, b 는 상수)

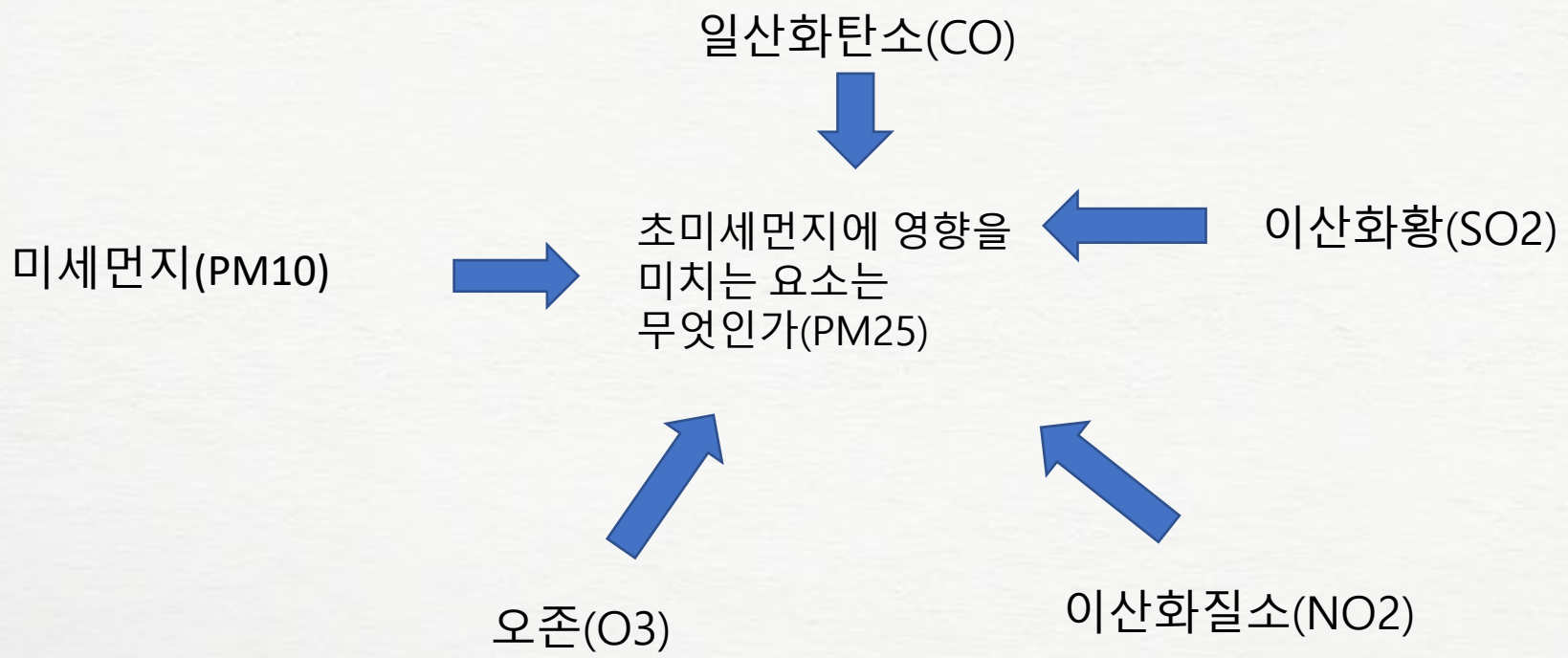
y = 3.93x -17.57 (W, b 는 상수)

III. 응용

1. 초미세먼지에 영향을 미치는 요소는 무엇인가?

1.1 상관분석 개요

- 분석 목적 : 초미세먼지에 영향을 미치는 요소는 무엇인가?
- [2019Atmosphere.csv] 데이터 셋: 2019년 대기 정보를 나타내는 데이터



1. 초미세먼지에 영향을 미치는 요소는 무엇인가?

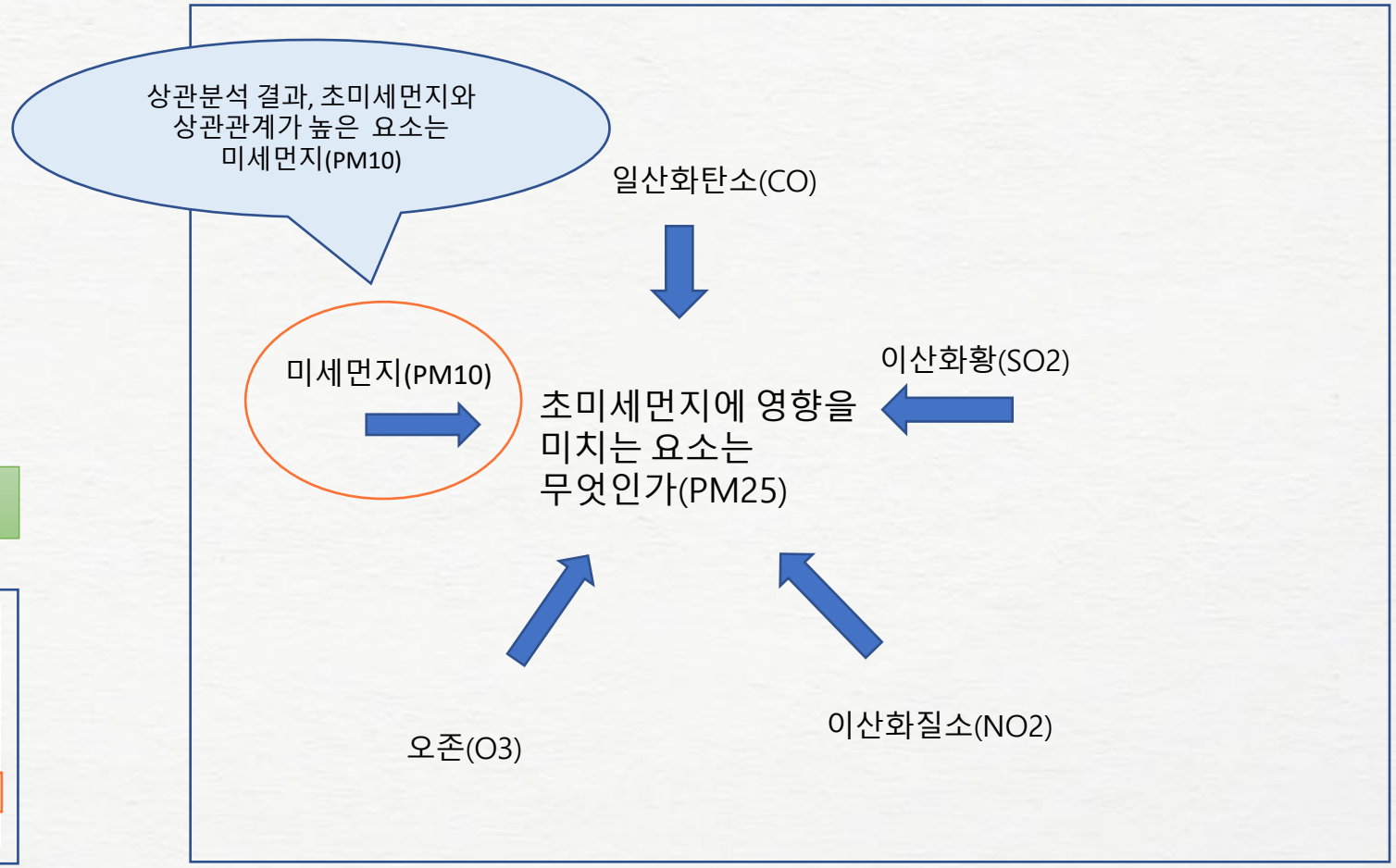
1.1 상관분석 개요

```
w_nA <- wA[ , 3:8]
```

A	B	C	D	E	F	G	H
loc	mdate	SO2	CO	O3	NO2	PM10	PM25
111123	2019100101	0.00	0.8	0.008	0.066	53	43
111123	2019100102	0.00	0.8	0.008	0.062	52	40
111123	2019100103	0.00	0.8	0.008	0.055	51	39
111123	2019100104	0.00	0.7	0.01	0.045	54	40
111123	2019100105	0.00	0.7	0.009	0.041	56	43
111123	2019100106	0.00	0.6	0.021	0.029	50	39
111123	2019100107	0.00	0.6	0.01	0.039	48	37
111123	2019100108	0.00	0.8	0.005	0.048	49	38

```
w_corA <- cor(w_nA)  #상관분석
```

	SO2	CO	O3	NO2	PM10	PM25
SO2	1.0000000	0.6213381	-0.15174257	0.4582891	0.1790310	0.33831184
CO	0.6213381	1.0000000	-0.37398843	0.6875159	0.2948801	0.53314922
O3	-0.1517426	-0.3739884	1.00000000	-0.7338416	0.1879649	-0.04419511
NO2	0.4582891	0.6875159	-0.73384159	1.0000000	0.1093246	0.39192641
PM10	0.1790310	0.2948801	0.18796486	0.1093246	1.0000000	0.67756693
PM25	0.3383118	0.5331492	-0.04419511	0.3919264	0.6775669	1.00000000



1. 초미세먼지에 영향을 미치는 요소는 무엇인가?

1.2 상관분석 데이터 준비

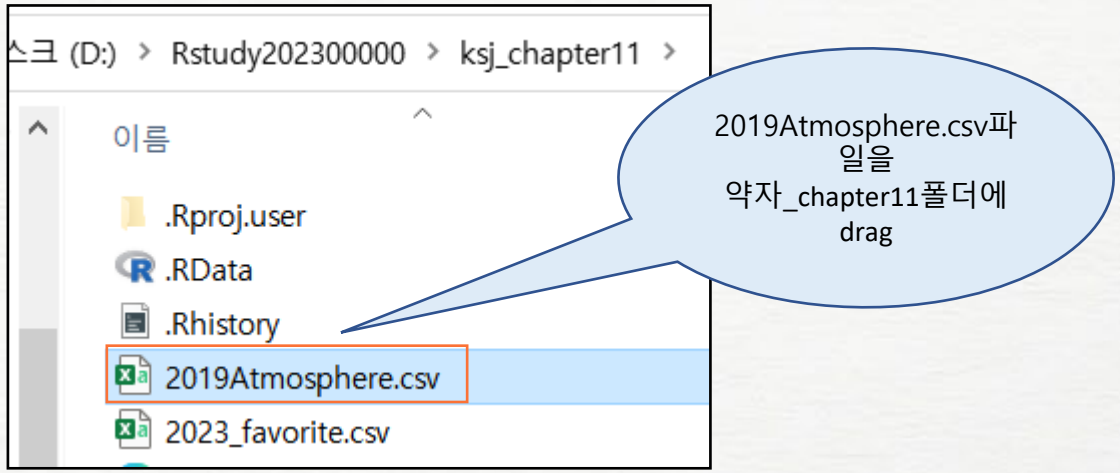
- 2019Atmosphere.csv데이터 셋: 2019년도 대기정보를 저장한 셋

2019Atmosphere.csv

A	B	C	D	E	F	G	H
loc	mdate	SO2	CO	O3	NO2	PM10	PM25
111123	2019100101	0.003	0.8	0.008	0.066	53	43
111123	2019100102	0.004	0.8	0.008	0.062	52	40
111123	2019100103	0.003	0.8	0.008	0.055	51	39
111123	2019100104	0.003	0.7	0.01	0.045	54	40
111123	2019100105	0.004	0.7	0.009	0.041	56	43
111123	2019100106	0.003	0.6	0.021	0.029	50	39
111123	2019100107	0.003	0.6	0.01	0.039	48	37
111123	2019100108	0.004	0.8	0.005	0.048	49	38

코드

약자_chapter11



1. 초미세먼지에 영향을 미치는 요소는 무엇인가?

1.2 상관분석 데이터 준비

코드

```
> wA <- read.csv("2019Atmosphere.csv", header = TRUE)          # 2019Atmosphere.csv를 데이터셋으로 불러옴
> head(wA)
```

	loc	mdate	SO2	CO	O3	NO2	PM10	PM25
1	111123	2019100101	0.003	0.8	0.008	0.066	53	43
2	111123	2019100102	0.004	0.8	0.008	0.062	52	40
3	111123	2019100103	0.003	0.8	0.008	0.055	51	39
4	111123	2019100104	0.003	0.7	0.010	0.045	54	40
5	111123	2019100105	0.004	0.7	0.009	0.041	56	43
6	111123	2019100106	0.003	0.6	0.021	0.029	50	39

```
> str(wA)
'data.frame':  3316 obs. of  8 variables:
 $ loc  : int  111123 111123 111123 111123 111123 111123 111123 111123 111123 111123 ...
 $ mdate: int  2019100101 2019100102 2019100103 2019100104 2019100105 2019100106 2019100107 20191
00108 2019100109 2019100110 ...
 $ SO2  : num  0.003 0.004 0.003 0.003 0.004 0.003 0.003 0.004 0.004 0.004 ...
 $ CO   : num  0.8 0.8 0.8 0.7 0.7 0.6 0.6 0.8 0.8 0.7 ...
 $ O3   : num  0.008 0.008 0.008 0.01 0.009 0.021 0.01 0.005 0.01 0.02 ...
 $ NO2  : num  0.066 0.062 0.055 0.045 0.041 0.029 0.039 0.048 0.049 0.045 ...
 $ PM10 : int  53 52 51 54 56 50 48 49 49 45 ...
 $ PM25 : int  43 40 39 40 43 39 37 38 37 32 ...
```

1. 초미세먼지에 영향을 미치는 요소는 무엇인가?

1.3 상관분석 실시

응용1. (가)를 쓰시오.

- 첫번째 열을 제외하고 3~8번째 열까지 별도의 데이터 셋 구성

코드

```
> w_nA <- (가)      # w 데이터 셋에서 3~8열 데이터만 가져오기
> head(w_nA)
```

	SO2	CO	O3	NO2	PM10	PM25
1	0.003	0.8	0.008	0.066	53	43
2	0.004	0.8	0.008	0.062	52	40
3	0.003	0.8	0.008	0.055	51	39
4	0.003	0.7	0.010	0.045	54	40
5	0.004	0.7	0.009	0.041	56	43
6	0.003	0.6	0.021	0.029	50	39

1. 초미세먼지에 영향을 미치는 요소는 무엇인가?

1.3 상관분석 실시

응용2. (가)를 쓰시오.

- cor() 함수를 이용해 상관분석을 실시

코드

```
> w_corA <- (가)          # w_nA 데이터 셋으로 상관분석한 결과를 w_corA 변수에 넣음
> w_corA                  # w_corA 상관분석 결과 확인
```

	SO2	CO	O3	NO2	PM10	PM25
SO2	1.0000000	0.6213381	-0.15174257	0.4582891	0.1790310	0.33831184
CO	0.6213381	1.0000000	-0.37398843	0.6875159	0.2948801	0.53314922
O3	-0.1517426	-0.3739884	1.00000000	-0.7338416	0.1879649	-0.04419511
NO2	0.4582891	0.6875159	-0.73384159	1.0000000	0.1093246	0.39192641
PM10	0.1790310	0.2948801	0.18796486	0.1093246	1.0000000	0.67756693
PM25	0.3383118	0.5331492	-0.04419511	0.3919264	0.6775669	1.00000000

=>초미세먼지(PM25)는 미세먼지(PM10)와 가장 상관도가 높음

1. 초미세먼지에 영향을 미치는 요소는 무엇인가?

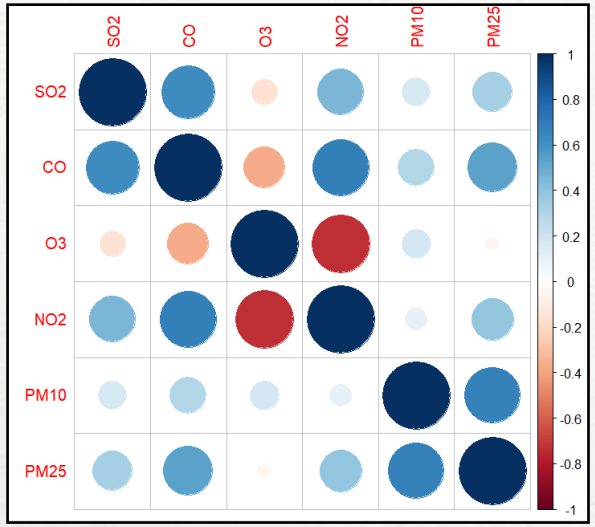
1.4 상관분석 결과표현

응용3. (가),(나)를 쓰시오.

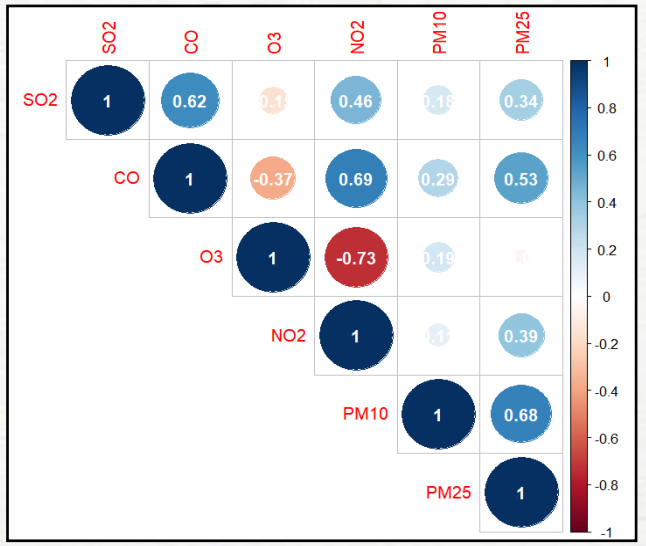
코드

```
> install.packages("corrplot")          # corrplot 패키지 설치
> library(corrplot)                     # corrplot 패키지 불러오기
> (가)                                  # 상관분석 결과인 w_corA을 corrplot 패키지로 실행해보기

# 동그란 원으로 표시하고, 상단에만 표시하고, 상관계수 표시
> (나)
```



<corrplot 패키지 활용 상관분석 결과1>



<corrplot 패키지 활용 상관분석 결과2>

50/16

2. 단순선형 회귀분석을 이용한 초미세먼지량 예측

2.1 데이터 준비

응용4. (가),(나)를 쓰시오.

초미세먼지량과 가장 상관도가 높았던 요소(PM25) ➡ 미세먼지(PM10)

코드

```
> w <- (가)
>w_nA <- (나)
>head(w_nA)
```

	SO2	CO	O3	NO2	PM10	PM25
1	0.003	0.8	0.008	0.066	53	43
2	0.004	0.8	0.008	0.062	52	40
3	0.003	0.8	0.008	0.055	51	39
4	0.003	0.7	0.010	0.045	54	40
5	0.004	0.7	0.009	0.041	56	43
6	0.003	0.6	0.021	0.029	50	39

2019Atmosphere.csv

loc	mdate	SO2	CO	O3	NO2	M10	PM25
111123	2019100101	0.003	0.8	0.00	0.066	53	43
111123	2019100102	0.004	0.8	0.00	0.062	52	40
111123	2019100103	0.003	0.8	0.00	0.055	51	39
111123	2019100104	0.003	0.7	0.0	0.045	54	40
111123	2019100105	0.004	0.7	0.00	0.041	56	43
111123	2019100106	0.003	0.6	0.02	0.029	50	39

2. 단순선형 회귀분석을 이용한 초미세먼지량 예측

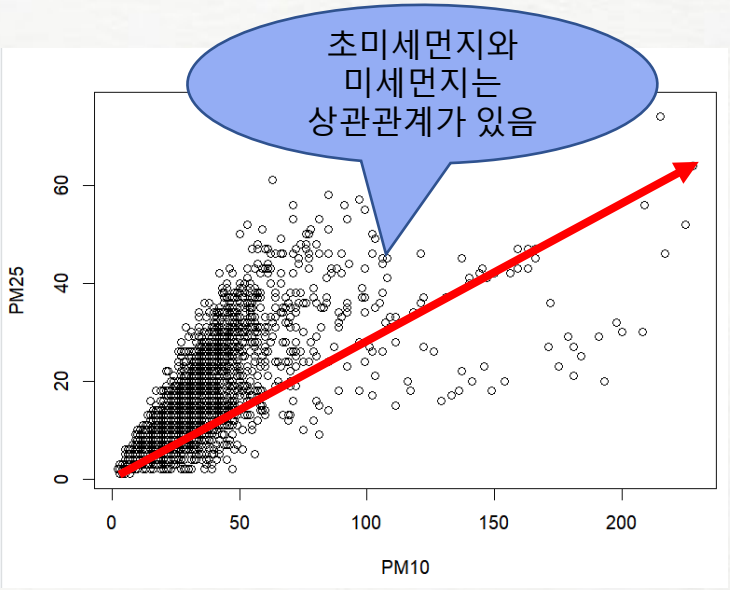
2.2 선형관계확인

- 초미세먼지와 미세먼지는 상관관계가 있음

응용5. (가)를 쓰시오.

코드

```
>head(w_nA)
>plot(  (가)  )     # 산점도를 통해 선형 관계 확인
```



2. 단순선형 회귀분석을 이용한 초미세먼지량 예측

2.3 회귀모델 구하기

- 회귀모델을 구하기

응용6. (가)를 쓰시오.

코드

```
>W_nAmodel <- (가)  # 회귀모델 구하기
>W_nAmodel

Call:
lm(formula = PM25 ~ PM10, data = w_nA)

Coefficients:
(Intercept)         PM10
     6.6377       0.2974
```

y(PM25) = x(PM10)*0.2974 +6.6377

y(PM25) = x(PM10)*W + b

2. 단순선형 회귀분석을 이용한 초미세먼지량 예측

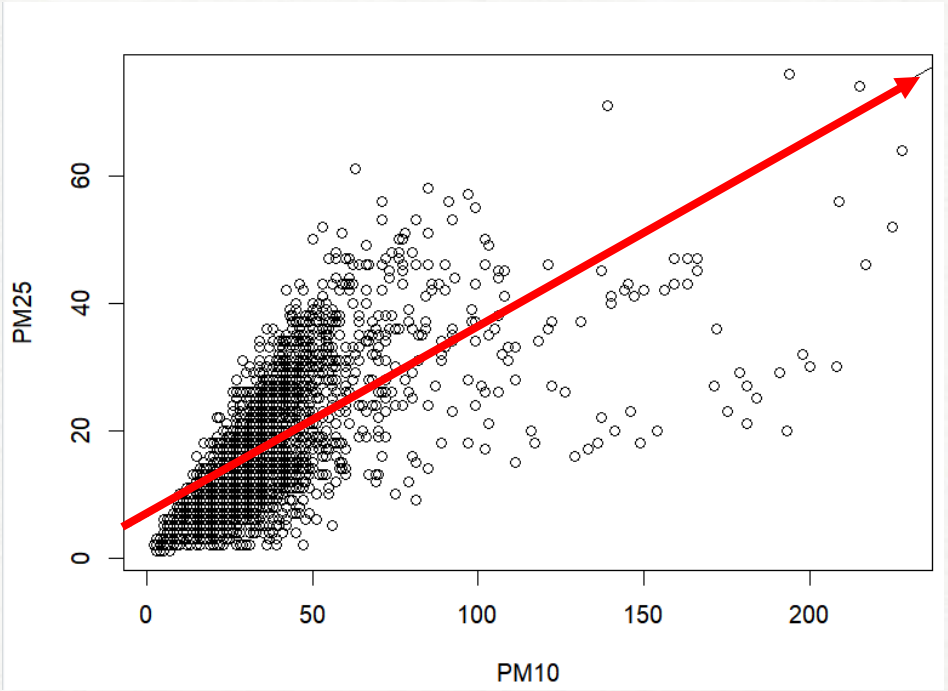
2.3 회귀모델 구하기

- 회귀선을 표기

응용7. (가)를 쓰시오.

코드

```
# 회귀선을 산점도 위에 표시
>(가)
```



y(PM25) = x(PM10)*0.2974 +6.6377

y(PM25) = x(PM10)*W + b

2. 단순선형 회귀분석을 이용한 초미세먼지량 예측

2.4 회귀식 구하기

- lm() 함수를 이용하여 쉽게 회귀식을 구할 수 있음

응용8. (가)를 쓰시오.

코드

```
>coef(W_nAmodel)[1]
(Intercept)
   6.637705
```

b 값 출력

y = Wx + b (a, b 는 상수)

y = Wx + 6.6377 (W, b 는 상수)

```
>(가)
      PM10
0.2973959
```

W 값 출력

y = Wx + 6.6377 (a, b 는 상수)

y = 0.2974 x + 6.6377 (W, b 는 상수)

y = Wx + b (W, b 는 상수)

y = 0.2974 x + 6.6377 (W, b 는 상수)

2. 단순선형 회귀분석을 이용한 초미세먼지량 예측

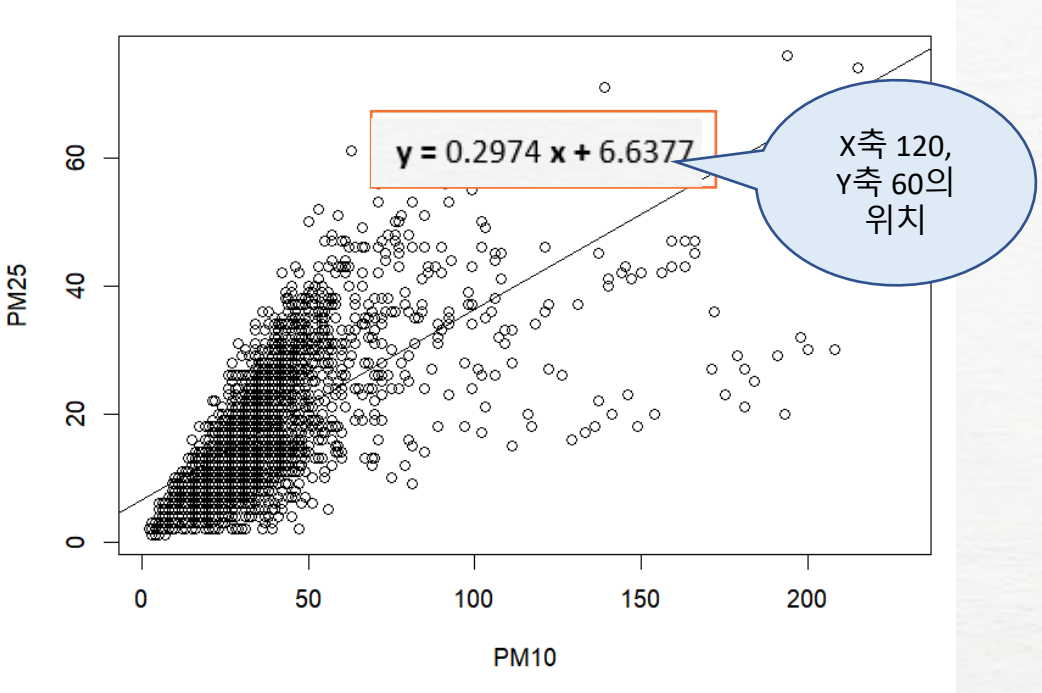
2.5 회귀식 표기

응용9. (가)를 쓰시오.

- 회귀식을 그래프상에 표기

코드

```
text( (가) , label = 'Y = 0.2974X +6.6377') # 회귀직선 텍스트로 표시, x와 y값은 그래프 상의 빈 공간에 임의의 점을 선정했음
```



y = Wx + b (W, b 는 상수)

y = 0.2974 x + 6.6377 (W, b 는 상수)

2. 단순선형 회귀분석을 이용한 초미세먼지량 예측

2.6 초미세먼지량 예측

응용10. (가)를 쓰시오.

- 회귀식에 의하여 미세먼지량을 입력하면 초미세먼지량을 예측할 수 있음

코드

```
b<- coef(W_nAmodel)[1]          # b=6.6377
W <- coef(W_nAmodel)[2]         #W=0.2974

pm10 <- 60                      # pm10=60
pm25<- (가)                     #PM25=W*pm10+b
pm25                            # 초미세먼지량
```

```
    PM10
24.48146
```

y = Wx + b (W, b 는 상수)

y = 0.2974 x + 6.6377 (W, b 는 상수)

미세먼지가 60이면 초미세먼지는 24.48일 것이다

오늘도 잘했어요