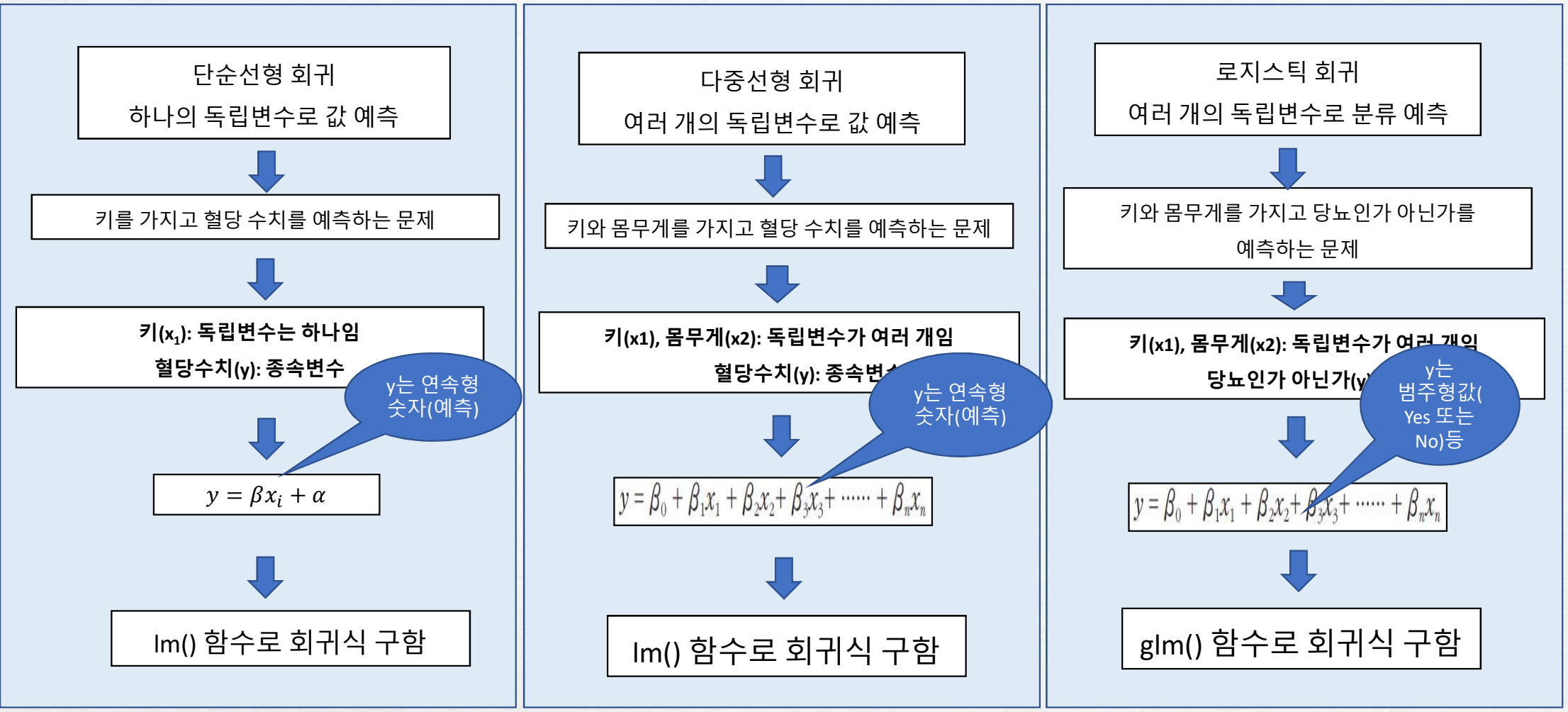
Chapter12. 다중선형 회귀분석&로지스틱 회귀분석

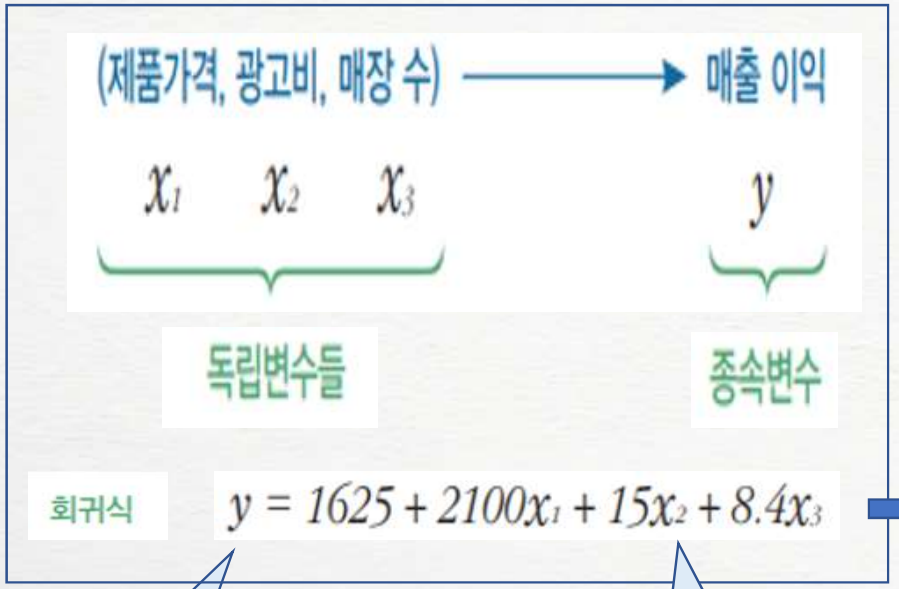
목차

요약1. 회귀모델(예측)


단순선형 회귀
하나의 독립변수로 값 예측
키를 가지고 혈당 수치를 예측하는 문제
키(x_1): 독립변수는 하나임
혈당수치(y): 종속변수
y는 연속형 숫자(예측)
$y = \beta x_i + \alpha$
lm() 함수로 회귀식 구함
다중선형 회귀
여러 개의 독립변수로 값 예측
키와 몸무게를 가지고 혈당 수치를 예측하는 문제
키(x1), 몸무게(x2): 독립변수가 여러 개임
혈당수치(y): 종속변
y는 연속형 숫자(예측)
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots\cdots + \beta_n x_n$
lm() 함수로 회귀식 구함
로지스틱 회귀
여러 개의 독립변수로 분류 예측
키와 몸무게를 가지고 당뇨인가 아닌가를 예측하는 문제
키(x1), 몸무게(x2): 독립변수가 여러 개임
당뇨인가 아닌가(y
y는 범주형값(Yes 또는 No)등
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots\cdots + \beta_n x_n$
glm() 함수로 회귀식 구함

요약2. 다중선형 회귀분석이란?

- **다중선형 회귀분석(regression analysis)**: 독립변수 여러 개가 종속변수에 미치는 영향을 파악하여 예측값 도출 (w_1, w_2, w_3와 b값 도출)



회귀식

Y(매출이익) = W_1 * x_1 + W_2 * x_2 + W_3 * x_3 + b (**W_1, W_2, W_3, b 는 상수**)

x_i 독립변수, y 종속변수

<독립변수, 종속변수, 회귀식>

W_1 <- 2100 W_2 <- 15 W_3 <- 8.4 b <- 1625

x_1,x_2,x_3에 따라 y값이 변경

제품가격(x_1), 광고비(x_2), 매장수(x_3)를 알면 매출이익(y)을 예측할 수 있다.

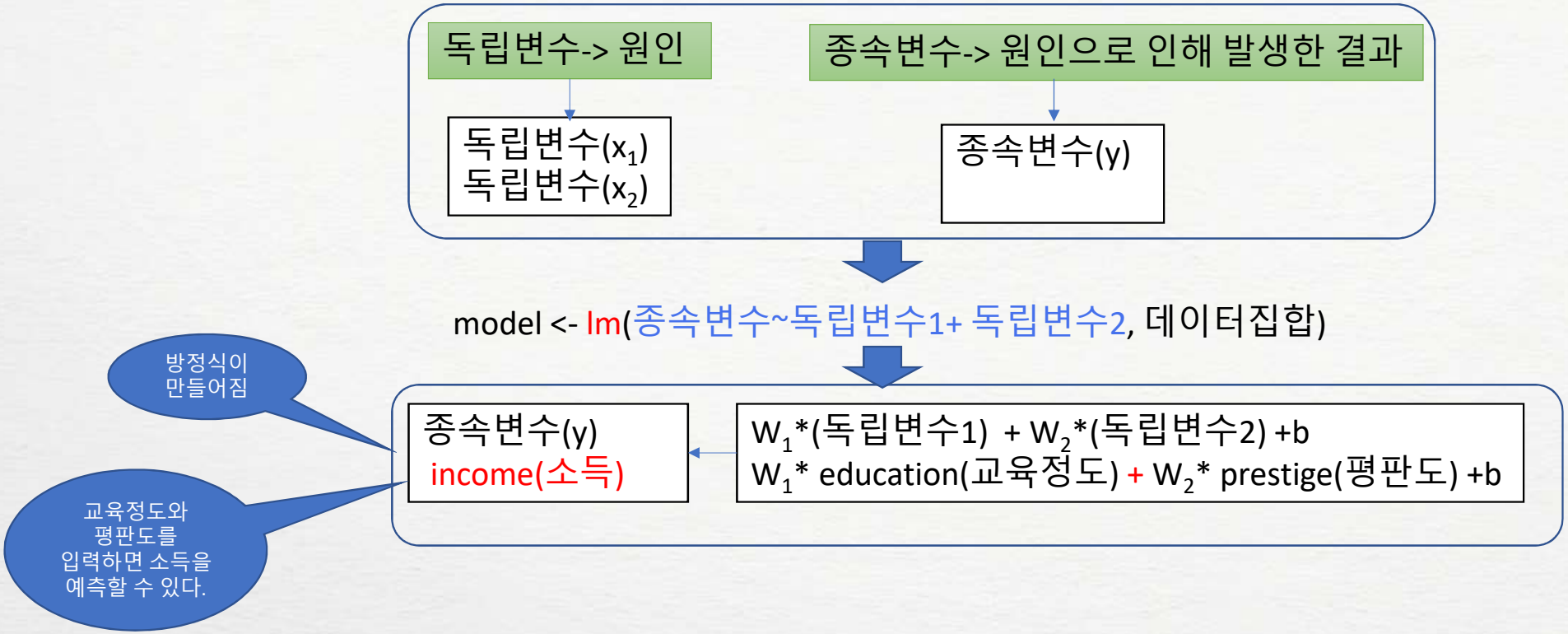
단순선형 회귀분석

제품가격(x_1)를 알면 매출이익(y)을 예측할 수 있다.

요약3. lm()함수- 회귀식 만들기

회귀식을 구하는 함수-> lm( )

model <- lm(종속변수~독립변수1+ 독립변수2, 데이터집합)
model <- lm(income~education+ prestige, newdata)



요약3. lm()함수- 회귀식 만들기

```
# 회귀식(lm->회귀식을 만들어 주는 함수)
> mod1 <- lm(income ~ education + prestige , data=newdata)
> mod1

Coefficients:
(Intercept)    education     prestige
    -253.8         177.2        141.4
```

lm이 방정식을 만드는 방법

회귀식

$y(income) = W_1 * x_1(education) + W_2 * x_2(prestige) + b$

회귀식

$y(income) = (177.2) * x_1(education) + (141.4) * x_2(prestige) + -253.8$

요약4. 다중선형회귀식 (W_1, W_2 값 추출)

-

y(income) = (177.2) * x_1(education)+ (141.4) * x_2(prestige) - 253.8

mod1

```
> coef(mod1)[1]
(Intercept)
  -253.8497
```

위 회귀식의 상수값 b 값 출력

$y = W_1x_1 + W_2x_2$ **-253.8**

```
> coef(mod1)[2]
education
  177.199
```

W_1 값 출력

$y =$ **177.2** $x_1 + W_2x_2 + b$

```
> coef(mod1)[3]
prestige
141.4354
```

W_2 값 출력

$y = W_1x_1 +$ **141.4** $x_2 + b$

coef(mod1)[1] coef(mod1)[2] coef(mod1)[3]

```
Coefficients:
(Intercept)    education     prestige
    -253.8        177.2        141.4
```

$y = W_1x_1 + W_2x_2 + b$ (W_1, W_2 ,b 는 상수)

$y = 177.2x_1 + 141.4x_2 - 253.8$

요약5. 소득을 예측하는 다중선형 회귀모델

여러 독립변수를 이용하여 예측값 계산

```
> b <- coef(mod1)[1]                    #y = 177.2x1+141.4x2 -253.8,    b<-253.8
> W1 <- coef(mod1)[2]                   #y = 177.2x1+141.4x2 -253.8,    W1<-177.2
> W2 <- coef(mod1)[3]                   #y = 177.2x1+141.4x2 -253.8,    W2<-141.4
> education  <- 12                      # 교육정도
> prestige <-  63                       # 평판

> income<- W1*education + W2*prestige + b
> income                                # 소득
  3823.2
```

교육연수 12, 평판도 63일 경우, 소득은 3823.2가 될 것이다.

⇒ x_1 ->education(12), x_2->prestige(63)를 입력

⇒ y->income의 값을 예측

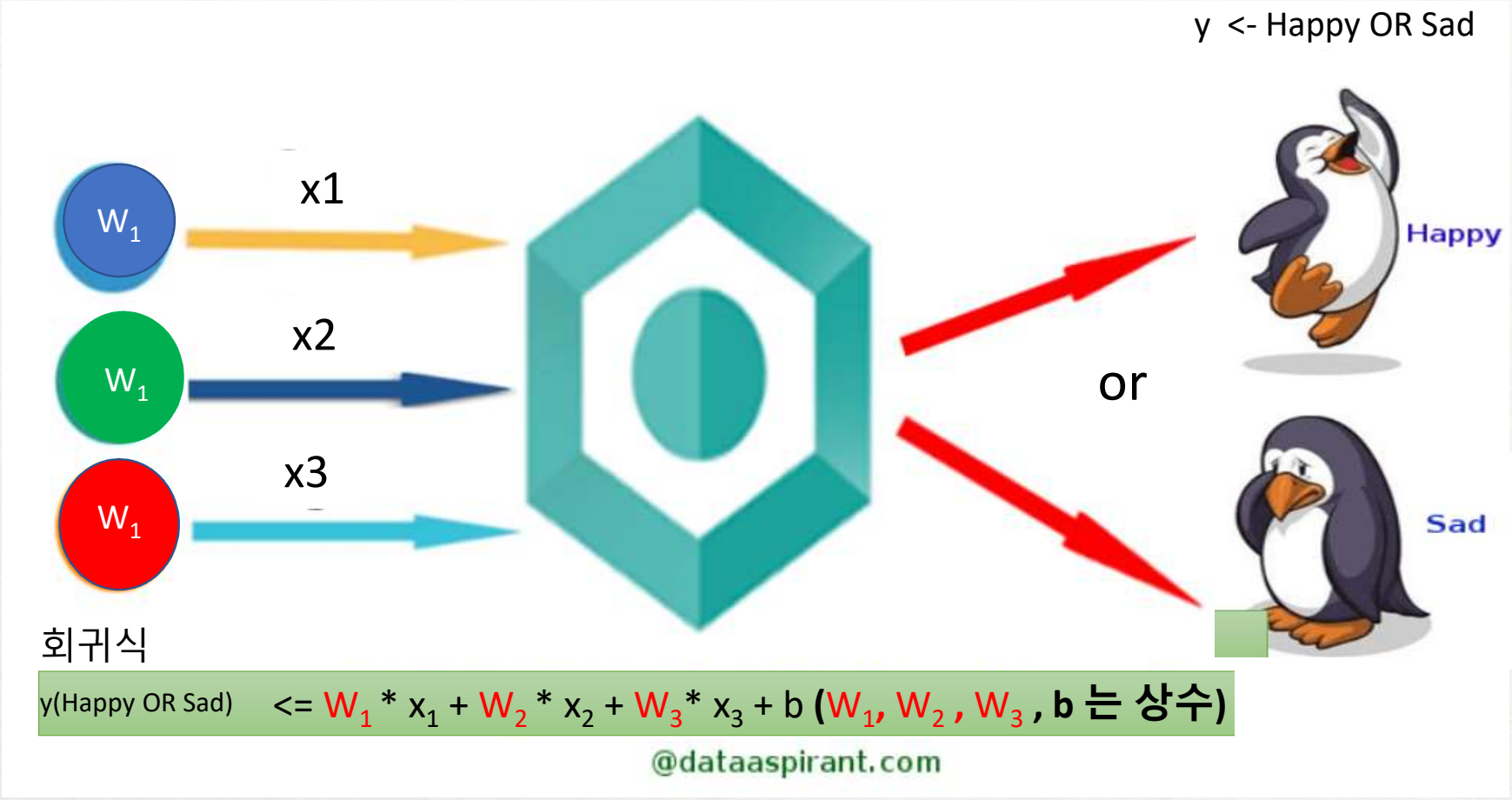
$y = W_1x_1 + W_2x_2 + b$ (W_1, W_2 , b 는 상수)

$y = 177.2x_1+141.4x_2 -253.8$

3823.2 <= 177.2 x(12)+141.4 x(63) -253.8

요약6. 로지스틱 회귀분석

- 회귀식에 의하여 펭귄이 행복한지 아닌가를 분류할 수 있음



요약7. glm()함수 - 로지스틱 회귀식

glm(종속변수 ~독립변수1+독립변수2+독립변수3+독립변수4, 데이터셋)->회귀식 도출

```
> mod.iris <- glm(Species ~Sepal.Length+Sepal.Width+Petal.Length+Petal.Width, data= iris.new)
> mod.iris  # 회귀모델의 상세 내용 확인
```

```
Coefficients:
 (Intercept)  Sepal.Length   Sepal.Width  Petal.Length   Petal.Width
     1.18650      -0.11191      -0.04008       0.22865       0.60925
```

회귀식 $y(\text{품종}) = W_1 * x_1 + W_2 * x_2 + W_3 * x_3 + W_4 * x_4 + b$

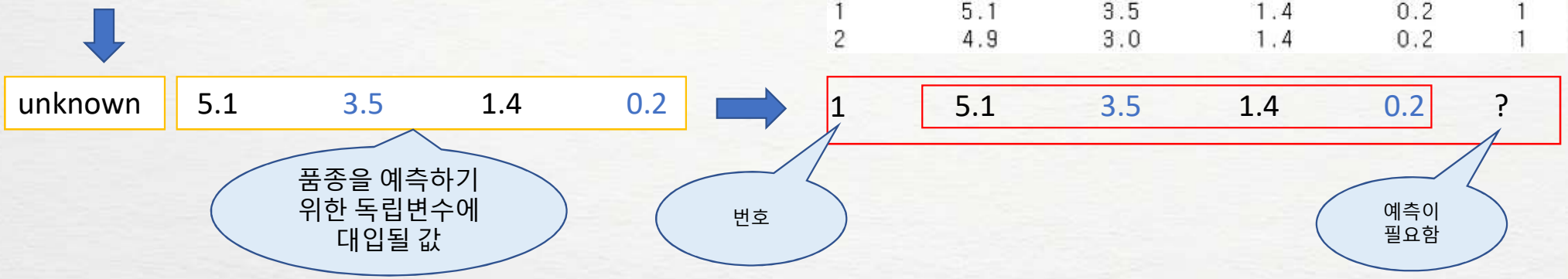
mod.iris $y(\text{품종}) = (-0.11191) * x_1 + (-0.04008) * x_2 + (0.22865) * x_3 + (0.60925) * x_4 + 1.1864$

요약8. 예측을 위한 데이터 생성

* 로지스틱 회귀모델을 이용한 예측을 위한 새로운 데이터생성

```
# 예측 대상 데이터 생성(데이터프레임)
> unknown <- data.frame(rbind(c(5.1, 3.5, 1.4, 0.2)))      #rbind(c(data1,data2…)): data1,data2…를 행으로 구성
> unknown
  1 5.1 3.5 1.4 0.2
```

rbind(c(5.1, 3.5, 1.4, 0.2)) #행으로 구성




```
> head(iris.new)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2       1
2          4.9         3.0          1.4         0.2       1
```

1 5.1 3.5 1.4 0.2 ?

요약9. 변수이름 배정

2.3 로지스틱 회귀모델을 이용한 예측을 위한 새로운 데이터생성

```
> names(unknown) <- names(iris)[1:4]          # 새로운 데이터에 변수이름 부여
> unknown                                     # 예측 대상 데이터에 변수이름이 부여된 결과
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1          5.1         3.5          1.4         0.2
```

names(iris)[1:4]

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	1
2	4.9	3.0	1.4	0.2	1

unknown

5.1	3.5	1.4	0.2

변수 이름이 없음

unknown

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	5.1	3.5	1.4	0.2

요약10. 회귀식에 의한 예측

```
# 새로운 데이터 품종 예측
> pred <- predict(mod.iris, unknown)        # 품종 예측
> pred                                      # 예측 결과 출력
        1
0.9174506
```

		Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
unknown	1	5.1	3.5	1.4	0.2

회귀식

mod.iris $y(\text{품종}) = (-0.11191) * x_1 + (-0.04008) * x_2 + (0.22865) * x_3 + (0.60925) * x_4 + 1.1864$

y(품종)=>pred <- (-0.11191) *5.1 + (-0.04008)* 3.5 + (0.22865)* 1.4 + (0.60925) * 0.2 + 1.1864

y(품종)=>pred <- 0.9174506

품종 예측 값

요약11. 예측값 배정

```
> unknown$Species <- round(pred,0)          #unknown에 Species추가,round(pred,0):소수 첫째자리에서 반올림
> unknown
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2       1
```

pred <- 0.9174506

round(pred, 0) : pred를 소수 첫째자리에서 반올림 ⟹ 1

unknown

5.1	3.5	1.4	0.2

round(pred,0) ⟹ 1

unknown$species

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	1

1의 품종은 Setosa

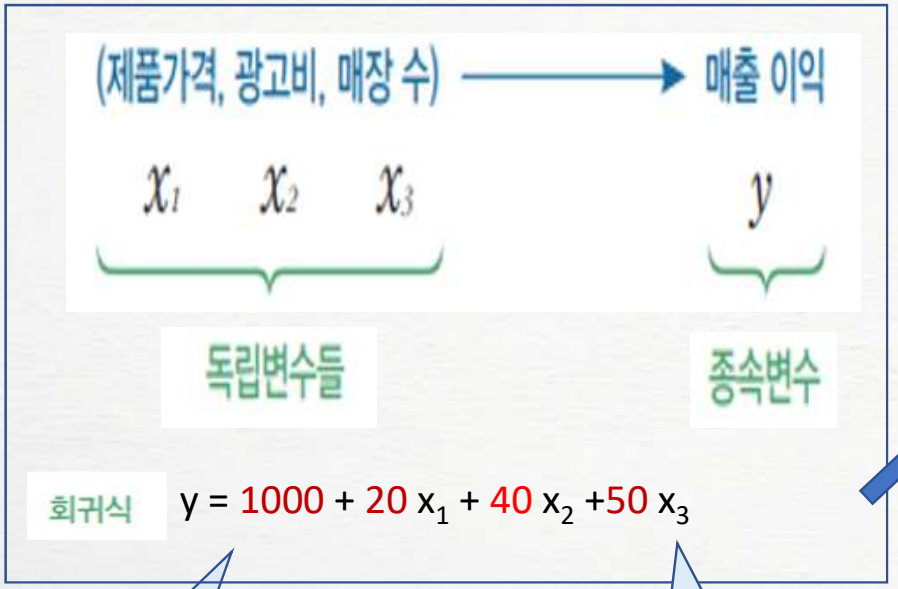
0. 소스분석문제

15

문제1. 다중선형 회귀분석이란?

문제1 ?~????를 채우시오

- **다중선형 회귀분석(regression analysis)**: 독립변수 여러 개가 종속변수에 미치는 영향을 파악하여 예측값 도출 (w_1, w_2, w_3와 b값 도출)



x_1,x_2,x_3에 따라 y값이 변경

제품가격(x_1), 광고비(x_2), 매장수(x_3)를 알면 매출이익(y)을 예측할 수 있다.

회귀식

$Y(\text{매출이익}) = W_1 * x_1 + W_2 * x_2 + W_3 * x_3 + b$ **(W_1, W_2, W_3, b 는 상수)**

x_i 독립변수, y 종속변수

<독립변수, 종속변수, 회귀식>

W_1 <- ? W_2 <- ?? W_3 <- ??? b <- ????

문제2. 다중선형회귀 분석 방법

문제2 (가), (나)에 들어갈 코드를 쓰시오.

```
# 회귀식(lm->회귀식을 만들어 주는 함수)
> mod1 <- lm(weight ~ (가), data=Mw_n)
> mod1
```

```
Call:
lm(formula = weight ~ egg_weight + food, data = Mw_n)

Coefficients:
(Intercept)   egg_weight        food
      3.664       1.745       1.595
```

회귀식

$y(weight) = W_1 * x_1 + W_2 * x_2 + b$

회귀식

y(weight) = (나)

문제3. 다중선형회귀식

문제3 (가), (나),(다)의 결과를 쓰시오.(숫자만 씁니다)

$y(weight) = W_1 * x_1 + W_2 * x_2 + b$ **(W_1, W_2 , b 는 상수)**

$y(weight) = 1.745 * x_1(egg\_weight) + 1.595 * x_2(food) * + 3.664$

>coef(MW_nmodel)[1]

(가)

b 값 출력

$y(weight) = W_1 * x_1 + W_2 * x_2 + b$

$y(weight) = W_1 * x_1 + W_2 * x_2 + 3.664$

>coef(MW_nmodel)[2]

(나)

W_1 값 출력

$y(weight) = W_1 * x_1 + W_2 * x_2 + b$

$y(weight) = 1.745 * x_1 + W_2 * x_2 + b$

>coef(MW_nmodel)[3]

(다)

W_2 값 출력

$y(weight) = W_1 * x_1 + W_2 * x_2 + b$

$y(weight) = W_1 * x_1 + 1.595 * x_2 + b$

문제4. 소득을 예측하는 다중선형 회귀모델

문제4. (가)의 결과를 쓰시오.

여러 독립변수를 이용하여 예측값 계산

```
> b <- coef(mod1)[1]                          #y = 177x1+141x2 -200,
> W1 <- coef(mod1)[2]                         #y = 177x1+141x2-200,
> W2 <- coef(mod1)[3]                         #y = 177x1+141x2-200,
> education  <- 10                            # 교육정도
> prestige <-  10                             # 평판
> income<- W1*education + W2*prestige + b
> income                                      # 소득
```

b<-200
W1<-177
W2<-141

(가)

교육연수 10, 평판도 10일 경우, 소득은 (가)가 될 것이다.

⇒ x_1 ->education(10), x_2->prestige(10) 를 입력
⇒ y->income의 값을 예측

$y = W_1x_1 + W_2x_2 + b$ (W_1, W_2 , b 는 상수)

$y = 177x_1+141x_2-200$

문제5. glm()함수 – 로지스틱 회귀식

문제5. (가)의 회귀식을 쓰시오.

glm(종속변수 ~독립변수1+독립변수2+독립변수3+독립변수4, 데이터셋)->회귀식 도출

```
> mod약자 <- glm(cluster ~Hamburger+Pizza+Cat+Dog+Summer, data= FC약자) # 로지스틱 회귀모델 도출
> mod약자                                                              # 회귀모델의 상세 내용 확인
```

```
Call:  glm(formula = cluster ~ ., data = FC약자)

Coefficients:
(Intercept)    Hamburger        Pizza          Cat          Dog       Summer
    0.18082     -0.04299     -0.01183      0.08036      0.13993      0.17516
```

y(분류) = $W_1 * x_1 + W_2 * x_2 + W_3 * x_3 + W_4 * x_4 + W_5 * x_5 + b$

Mod약자

y(분류) = (가)

문제6. 로지스틱 회귀모델을 이용한 사용자 분류

문제6. (가)를 쓰시오.

```
# 예측 대상 데이터 생성(데이터프레임)
> ksj <- data.frame(rbind(  (가)     )))        # 고수정 설문 결과
> ksj
```

	X1	X2	X3
1	10	10	10

```
> head(FC약자)
```

	Hamburger	Pizza	Cat
1	10	9	8
2	10	10	7
3	10	10	10
4	10	8	6
5	6	8	3
6	10	10	10

ksj	10	10	10

문제7. 로지스틱 회귀모델을 이용한 사용자 분류

문제7. (가),(나)를 쓰시오.

로지스틱 회귀모델을 이용한 새로운 데이터 군집 예측

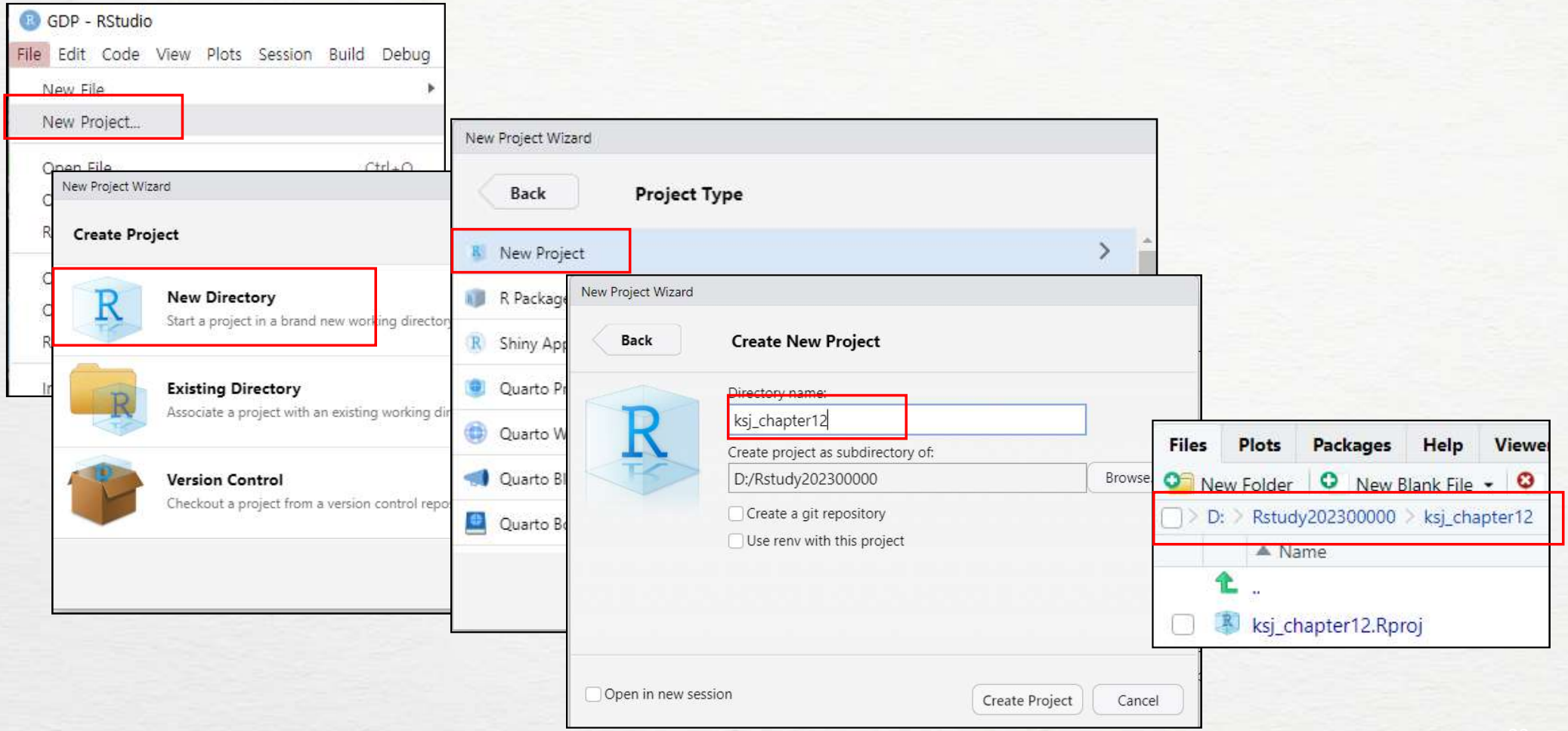
```
# 새로운 데이터 군집 예측
> pred <- predict(mod약자 , ksj)          # 고수정 군집 예측
> pred                                    # 예측 결과 출력
   (가)
> round(pred,0)
   (나)
```

		Hamburger	Pizza	Cat
ksj	1	10	10	10

Mod약자 $y(분류) = (-0.01) * x_1 + (0.07) * x_2 + (0.06) * x_3 + 0.7$

y(분류)=>pred <- (가)

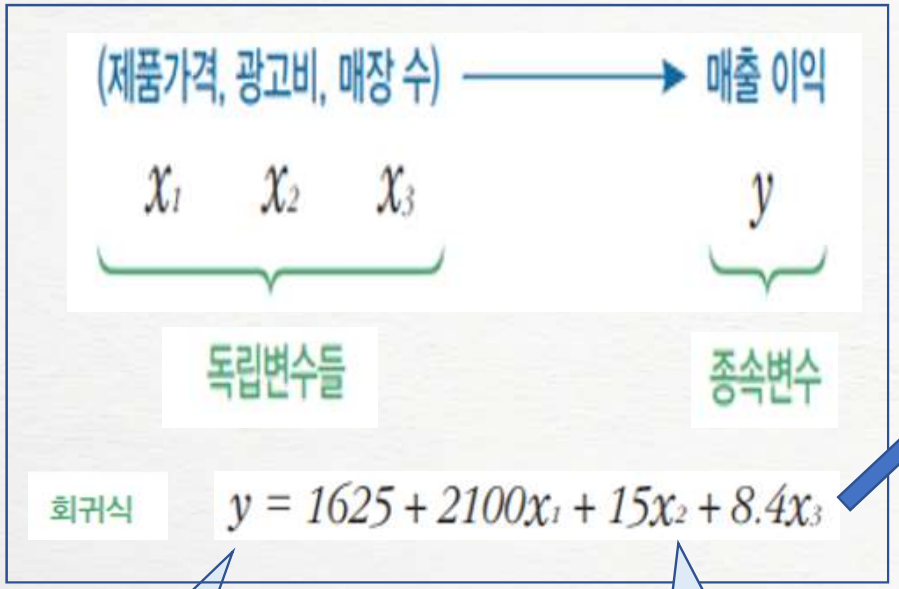
* 프로젝트 시작


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
Shiny App
Quarto Pr
Quarto W
Quarto Bl
Quarto Bo
New Project Wizard
Back
Create New Project
Directory name:
ksj_chapter12
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 Viewe
New Folder
New Blank File
D: Rstudy202300000 ksj_chapter12
Name
..
ksj_chapter12.Rproj

I. 다중선형 회귀분석

1. 다중선형 회귀분석이란?

- **다중선형 회귀분석(regression analysis)**: 독립변수 여러 개가 종속변수에 미치는 영향을 파악하여 예측값 도출 (w_1, w_2, w_3와 b값 도출)



회귀식

Y(매출이익) = $W_1 * x_1 + W_2 * x_2 + W_3 * x_3 + b$ **(W_1, W_2, W_3, b 는 상수)**

x_i 독립변수, y 종속변수

<독립변수, 종속변수, 회귀식>

W_1 <- 2100 W_2 <- 15 W_3 <- 8.4 b <- 1625

x_1,x_2,x_3에 따라 y값이 변경

제품가격(x_1), 광고비(x_2), 매장수(x_3)를 알면 매출이익(y)을 예측할 수 있다.

2. 다중선형 회귀분석을 이용한 병아리 몸무게 예측

2.1 데이터 준비

코드

```
> w <- read.csv("ch5-1.csv", header = TRUE)
>w_n <- w[,2:5]
>head(w_n)
```

	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

ch5-1.csv

chick_nm	weight	egg_weigl	movemer	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. 다중선형 회귀분석을 이용한 병아리 몸무게 예측

2.2 선형관계확인

- 몸무게와 종란무게, 먹는양은 상관도가 높음->이동량(movement)는 제외

코드

```
>Mw_n<- w_n[, c(1,2,4)]
>head(Mw_n)
```

w_n[, c(1,2,4)]

	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

movement제 외함

Mw_n

	weight	egg_weight	food
1	140	65	14
2	128	62	12
3	140	65	13
4	135	65	13
5	145	69	13
6	138	65	13

2. 다중선형 회귀분석을 이용한 병아리 몸무게 예측

2.2 선형관계확인

- 몸무게는 종란무게, 먹는 양과 상관도가 높음 - >movement는 제외

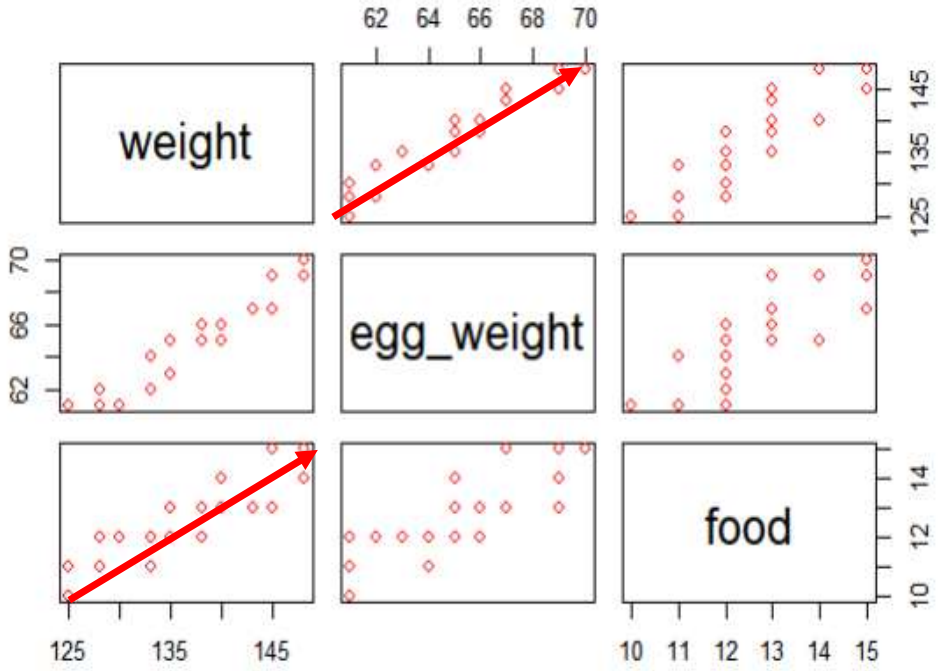
몸무게와 종란 무게,
먹는 양은
상관관계가 있음

코드

```
# 산점도를 통해 선형 관계 확인
>plot(Mw_n, col="red")
```

Mw_n

	weight	egg_weight	food
1	140	65	14
2	128	62	12
3	140	65	13
4	135	65	13
5	145	69	13
6	138	65	13



2. 다중선형 회귀분석을 이용한 병아리 몸무게 예측

2.3 회귀모델 구하기

lm() 함수를 이용하여 쉽게 회귀식을 구할 수 있음

코드

```
>MW_nmodel <- lm(weight~egg_weight+food, Mw_n)     # 회귀모델 구하기
>MW_nmodel

Call:
lm(formula = weight ~ egg_weight + food, data = Mw_n)

Coefficients:
(Intercept)   egg_weight         food
      3.664        1.745        1.595
```

y(weight) = 1.745 *x_1(egg_weight)*+ 1.595 *x_2(food)* +3.664

y(weight) = W_1 * x_1(egg_weight)*+ W_2 * x_2(food) + b

2. 다중선형 회귀분석을 이용한 병아리 몸무게 예측

2.4 회귀식 구하기

$y(weight) = W_1 * x_1 + W_2 * x_2 + b$ **(W_1, W_2 , b 는 상수)**

$y(weight) = 1.745 * x_1(egg\_weight) * + 1.595 * x_2(food) * + 3.664$

코드

```
>coef(MW_nmodel)[1]
(Intercept)
    3.66385
```

b 값 출력

$y(weight) = W_1 * x_1 + W_2 * x_2 + b$

$y(weight) = W_1 * x_1 + W_2 * x_2 + 3.664$

```
>coef(MW_nmodel)[2]
egg_weight
  1.745323
```

W_1 값 출력

$y(weight) = W_1 * x_1 + W_2 * x_2 + b$

$y(weight) = 1.745 * x_1 + W_2 * x_2 + b$

```
>coef(MW_nmodel)[3]
    food
1.595467
```

W_2 값 출력

$y(weight) = W_1 * x_1 + W_2 * x_2 + b$

$y(weight) = W_1 * x_1 + 1.595 * x_2 + b$

2. 다중선형 회귀분석을 이용한 병아리 몸무게 예측

2.5 닭의 몸무게 예측

- 회귀식에 의하여 종란 무게를 입력하면 닭의 무게를 예측할 수 있음

코드

```
b    <- coef(MW_nmodel)[1]              # b=-3.664
W1 <- coef(MW_nmodel)[2]                #W1=1.745
W2 <- coef(MW_nmodel)[3]                #W2=1.595

egg_weight <- 71                        # 종란무게(egg_weight )=71
food <- 15                              # 먹는양(food )=15
weight <- W1*egg_weight+W2*food+ b
weight                                  # 닭의 몸무게
```

```
egg_weight
  151.5138
```

종란무게가 71,먹는양이 15이면 닭의 무게는 151.5138이 될 것이다.

$y(weight) = W_1 * x_1 + W_2 * x_2 + b$ **(W_1, W_2 , b 는 상수)**

$y(weight) = 1.745\ x_1 + 1.595\ x_2 + 3.664$

II. 로지스틱 회귀분석

1. 로지스틱 회귀분석의 개념

로지스틱 회귀(logistic regression)

종속변수의 값의 형태가 연속형 숫자가 아닌 범주형 값인 경우를 다루기 위해서 만들어진 통계적 방법

iris 데이터셋에서 4개의 측정값을 가지고 품종을 예측

R에서 로지스틱 회귀 모델은 glm() 함수 이용

2. 꽃의 품종을 예측하기 위한 로지스틱 회귀모델

2.1 데이터 준비

코드

```
>iris.new <- iris
>iris.new$Species <- as.integer(iris.new$Species)        # 범주형 자료를 정수로 변환
>head(iris.new)
```

iris

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

iris.new

```
> head(iris.new)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2       1
2          4.9         3.0          1.4         0.2       1
3          4.7         3.2          1.3         0.2       1
4          4.6         3.1          1.5         0.2       1
5          5.0         3.6          1.4         0.2       1
6          5.4         3.9          1.7         0.4       1
```

품종이 정수로 변환

as.integer(iris.new$Species)

2. 꽃의 품종을 예측하기 위한 로지스틱 회귀모델

코드 2.2 회귀모델 도출

```
# 로지스틱 회귀모델 도출
> mod.iris <- glm(Species ~Sepal.Length+Sepal.Width+Petal.Length+Petal.Width, data= iris.new)
> mod.iris                                                    # 회귀모델의 상세 내용 확인

Call:  glm(formula = Species ~ ., data = iris.new)

Coefficients:
 (Intercept)  Sepal.Length   Sepal.Width  Petal.Length   Petal.Width
     1.18650      -0.11191      -0.04008       0.22865       0.60925

Degrees of Freedom: 149 Total (i.e. Null);  145 Residual
Null Deviance:      100
Residual Deviance: 6.961        AIC: -22.87
```

- Species ~.
  회귀모델에서 종속변수가 Species이고, 나머지 변수들은 모두 독립변수이다.
- data=iris.new
  회귀모델 도출에 사용할 데이터셋이 iris.new이다.

$y(\text{품종}) = W_1 * x_1 + W_2 * x_2 + W_3 * x_3 + W_4 * x_4 + b$

mod.iris

$y(\text{품종}) = (-0.11191) * x_1 + (-0.04008) * x_2 + (0.22865) * x_3 + (0.60925) * x_4 + 1.1864$

2. 꽃의 품종을 예측하기 위한 로지스틱 회귀모델

2.3 로지스틱 회귀모델을 이용한 예측을 위한 새로운 데이터생성

코드

```
# 예측 대상 데이터 생성(데이터프레임)
> unknown <- data.frame(rbind(c(5.1, 3.5, 1.4, 0.2)))
> unknown
  1 5.1 3.5 1.4 0.2
```

```
> head(iris.new)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2       1
2          4.9         3.0          1.4         0.2       1
3          4.7         3.2          1.3         0.2       1
4          4.6         3.1          1.5         0.2       1
5          5.0         3.6          1.4         0.2       1
6          5.4         3.9          1.7         0.4       1
```

unknown

5.1	3.5	1.4	0.2

1	5.1	3.5	1.4	0.2	?

2. 꽃의 품종을 예측하기 위한 로지스틱 회귀모델

2.3 로지스틱 회귀모델을 이용한 예측을 위한 새로운 데이터생성

코드

```
> names(unknown) <- names(iris)[1:4]          # 새로운 데이터에 변수이름 부여
> unknown                                     # 예측 대상 데이터에 변수이름이 부여된 결과
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1          5.1         3.5          1.4         0.2
```

```
> head(iris.new)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2       1
2          4.9         3.0          1.4         0.2       1
3          4.7         3.2          1.3         0.2       1
4          4.6         3.1          1.5         0.2       1
5          5.0         3.6          1.4         0.2       1
6          5.4         3.9          1.7         0.4       1
```

unknown

5.1	3.5	1.4	0.2

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	5.1	3.5	1.4	0.2

2. 꽃의 품종을 예측하기 위한 로지스틱 회귀모델

2.4 로지스틱 회귀모델을 이용한 새로운 데이터 품종 예측

코드

```
# 새로운 데이터 품종 예측
> pred <- predict(mod.iris, unknown)          # 품종 예측
> pred                                        # 예측 결과 출력
        1
0.9174506
```

		Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
unknown	1	5.1	3.5	1.4	0.2

mod.iris $y(\text{품종}) = (-0.11191) * x_1 + (-0.04008) * x_2 + (0.22865) * x_3 + (0.60925) * x_4 + 1.1864$

y(품종)=>pred <- (-0.11191) *5.1 + (-0.04008)* 3.5 + (0.22865)* 1.4 + (0.60925) * 0.2 + 1.1864

y(품종)=>pred <- 0.9174506

2. 꽃의 품종을 예측하기 위한 로지스틱 회귀모델

2.4 로지스틱 회귀모델을 이용한 새로운 데이터 품종 예측

코드

```
> round(pred,0)
```

```
> round(pred,0)
1
1
```

예측 결과 출력(소수 첫째 자리에서 반올림)

#0.9174506을 반올림하면 1

		Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
unknown	1	5.1	3.5	1.4	0.2

mod.iris $y(\text{품종}) = (-0.11191) * x_1 + (-0.04008) * x_2 + (0.22865) * x_3 + (0.60925) * x_4 + 1.1864$

y(품종)=>pred <- (-0.11191) *5.1 + (-0.04008)* 3.5 + (0.22865)* 1.4 + (0.60925) * 0.2 + 1.1864

y(품종)=>pred <- 0.9174506

round(pred,0) <- 1

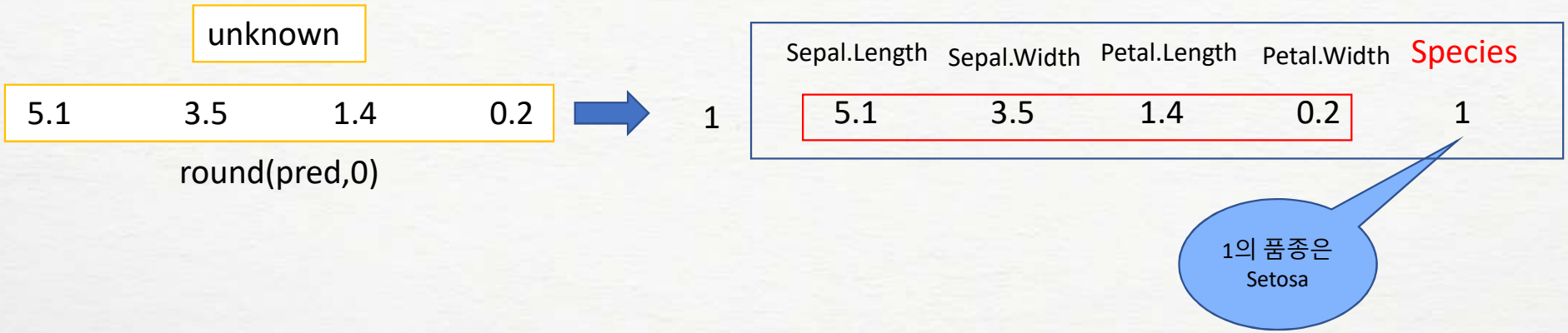
0.9174506 반올림한 결과

2. 꽃의 품종을 예측하기 위한 로지스틱 회귀모델

2.4 로지스틱 회귀모델을 이용한 새로운 데이터 품종 예측

코드

```
> unknown$Species <- round(pred,0)          #unknown에 Species추가
> unknown
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2       1
```



Ⅲ. 응용

41

1. 소득을 예측하는 다중선형 회귀모델

1.1 데이터 준비

코드

```
>install.packages("car")
>library(car)
>head(Prestige)
>newdata <- Prestige[,c(1:4)]                    # 회귀식 작성을 위한 데이터 준비
>head(newdata)
```

소득은 교육연수(education), 여성비율(women),
평판도(prestige) 등의 값을 통하여 예측가능하다

Prestige[,c(1:4)]

```
> head(Prestige)
                    education ncome women prestige census type
gov.administrators      13.11  12351 11.16     68.8   1113 prof
general.managers        12.26  25879  4.02     69.1   1130 prof
accountants             12.77   9271 15.70     63.4   1171 prof
purchasing.officers     11.42   8865  9.11     56.8   1175 prof
chemists                14.62   8403 11.68     73.5   2111 prof
physicists              15.64  11030  5.13     77.6   2113 prof
```

newdata

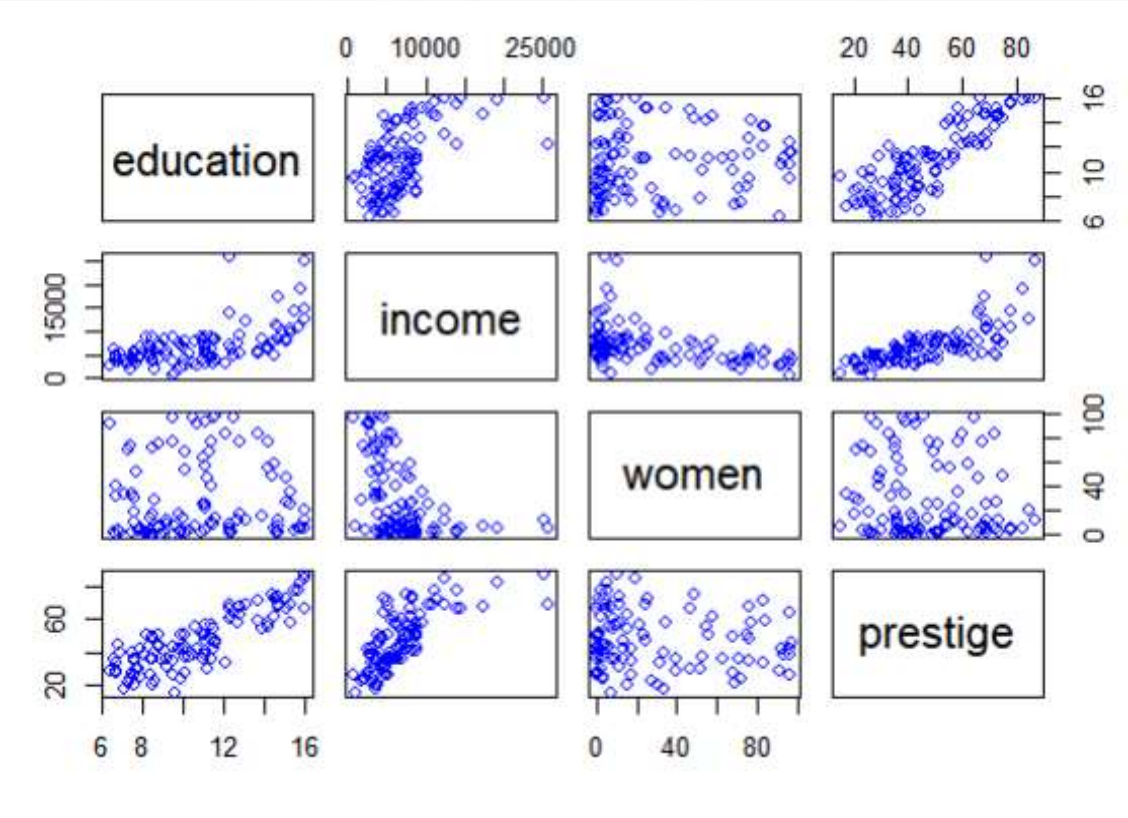
```
> newdata
                    education ncome women prestige
gov.administrators      13.11  12351 11.16     68.8
general.managers        12.26  25879  4.02     69.1
accountants             12.77   9271 15.70     63.4
purchasing.officers     11.42   8865  9.11     56.8
chemists                14.62   8403 11.68     73.5
physicists              15.64  11030  5.13     77.6
```

1. 소득을 예측하는 다중선형 회귀모델

1.2 산점도

코드

```
> plot(newdata, col="blue“)                    # 산점도를 통해 변수 간 관계 확인
```



1. 소득을 예측하는 다중선형 회귀모델

1.3 다중선형회귀 분석

코드

```
# 회귀식
> mod1 <- lm(income ~ education + prestige + women, data=newdata)
> mod1
```

```
Call:
lm(formula = income ~ education + prestige + women, data = newdata)

Coefficients:
(Intercept)    education     prestige        women
     -253.8        177.2        141.4        -50.9
```

- **income ~ education + prestige + women**

  회귀모델에서 무엇이 독립변수이고 무엇이 종속변수인지 지정하는 것으로, ~ 앞에 있는 것이 종속변수, ~ 뒤쪽에 있는 것이 독립변수이다. 독립변수가 여러 개이면 +로 연결한다.

- **data=newdata**

  회귀모델 도출에 사용할 데이터셋을 지정한다. 변수명 income, education, prestige, women은 newdata에 속한 열의 이름이다.

$$y(income) = W_1 * x_1(education) + W_2 * x_2(prestige) + W_3 * x_3(women) + b$$

$$y(income) = (177.2) * x_1(education) + (141.4) * x_2(prestige) + (-50.9) * x_3(women) - 253.8$$

1. 소득을 예측하는 다중선형 회귀모델

1.4 다중선형회귀식 구하기

코드

```
> coef(mod1)[1]
(Intercept)
  -253.8497
```

b 값 출력

$y = W_1x_1 + W_2x_2 + W_3x_3 - 253.8$

```
> coef(mod1)[2]
education
  177.199
```

W_1 값 출력

$y = 177.2x_1 + W_2x_2 + W_3x_3 + b$

```
> coef(mod1)[3]
prestige
141.4354
```

W_2 값 출력

$y = W_1x_1 + 141.4x_2 + W_3x_3 + b$

```
> coef(mod1)[4]
   women
-50.8957
```

W_3 값 출력

$y = W_1x_1 + W_2x_2 + -50.9x_3 + b$

```
Coefficients:
(Intercept)    education     prestige        women
     -253.8        177.2        141.4        -50.9
```

$y = W_1x_1 + W_2x_2 + W_3x_3 + b$ (W_1, W_2 , W_3 ,b 는 상수)

$y = 177.2x_1 + 141.4x_2 - 50.9x_3 - 253.8$

1. 소득을 예측하는 다중선형 회귀모델

응용1. income의 예측값이 얼마인지 화면을 캡처하시오.

1.5 소득 예측

코드

```
> b <- coef(mod1)[1]                    #y = 177.2x1+141.4x2-50.9x3 -253.8
> W1 <- coef(mod1)[2]                   #y = 177.2x1+141.4x2-50.9x3 -253.8
> W2 <- coef(mod1)[3]                   #y = 177.2x1+141.4x2-50.9x3 -253.8
> W3 <- coef(mod1)[4]                   #y = 177.2x1+141.4x2-50.9x3 -253.8

> education  <- 12                      # 교육정도
> Prestige <-  63                       # 평판
> women <-  12                          #여성 비율
> income<- W1*education + W2*prestige + W3*women + b  # y = 177.2X12 + 141.4X63 - 50.9X12  - 253.8
> income                                # 소득
```

교육연수 12, 평판도 63,
여성비율 12일
경우소득은 ?이 될 것이다.

=>회귀식에 의하여 education, prestige, women를 입력하면 income을 예측할 수 있음

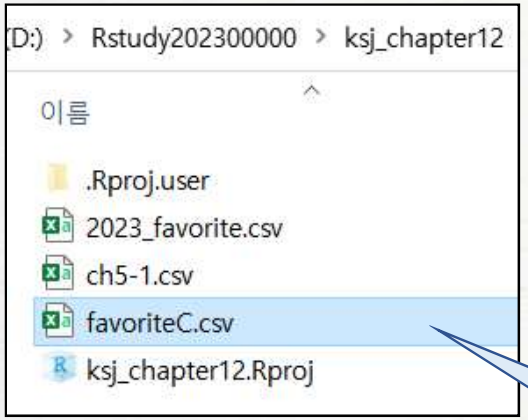
$y = W_1x_1 + W_2x_2 + W_3x_3 + b$ (W_1, W_2 , W_3 ,b 는 상수)

$y = 177.2x_1+141.4x_2-50.9x_3 -253.8$

2. 로지스틱 회귀모델을 이용한 사용자 분류

2.1 데이터 준비

코드



favoriteC.csv

No	ClassNum	Hamburge	Pizza	Cat	Dog	Summer	cluster
1	201608045	10	9	8	7	6	3
2	201612010	10	10	7	6	3	1
3	201612038	10	10	10	10	0	2
4	201712010	10	8	6	2	5	1
5	201712039	6	8	3	9	7	3

숫자

Lms에서
favoriteC.csv파일을
다운로드하고
약자_chapter12로 이동

2. 로지스틱 회귀모델을 이용한 사용자 분류

2.1 데이터 준비

코드

```
> F약자 <- read.csv("favoriteC.csv", header = TRUE)    # CSV파일을 [F약자]의 이름으로 저장
> head(F약자)
```

	No	ClassNum	Hamburger	Pizza	Cat	Dog	Summer	cluster
1	1	201608045	10	9	8	7	6	3
2	2	201612010	10	10	7	6	3	1
3	3	201612038	10	10	10	10	0	2
4	4	201712010	10	8	6	2	5	1
5	5	201712039	6	8	3	9	7	3
6	6	201809065	10	10	10	10	5	3

3열~8열

```
> FC약자 <- F약자[, 3:8]
```

#3열에서 8열까지 선택하여 [FC약자]이름으로 저장

	Hamburger	Pizza	Cat	Dog	Summer	cluster
1	10	9	8	7	6	3
2	10	10	7	6	3	1
3	10	10	10	10	0	2
4	10	8	6	2	5	1
5	6	8	3	9	7	3
6	10	10	10	10	5	3

2. 로지스틱 회귀모델을 이용한 사용자 분류

2.2 회귀모델 도출

코드

```
> mod약자 <- glm(cluster ~Hamburger+Pizza+Cat+Dog+Summer, data= FC약자) # 로지스틱 회귀모델 도출
> mod약자                                                               # 회귀모델의 상세 내용 확인
```

```
Call:  glm(formula = cluster ~ ., data = FC약자)

Coefficients:
(Intercept)    Hamburger        Pizza          Cat          Dog       Summer
    0.18082     -0.04299     -0.01183      0.08036      0.13993      0.17516

Degrees of Freedom: 58 Total (i.e. Null);  53 Residual
Null Deviance:      37.39
Residual Deviance: 9.756        AIC: 75.26
```

$$y(\text{분류}) = W_1 * x_1 + W_2 * x_2 + W_3 * x_3 + W_4 * x_4 + W_5 * x_5 + b$$

Mod약자

$$y(\text{분류}) = (-0.04299) * x_1 + (-0.01183) * x_2 + (0.08036) * x_3 + (0.13993) * x_4 + (0.17516) * x_5 + 0.18082$$

2. 로지스틱 회귀모델을 이용한 사용자 분류

2.3 로지스틱 회귀모델을 이용한 예측을 위한 새로운 사용자 생성

코드

```
# 예측 대상 데이터 생성(데이터프레임)
> ksj <- data.frame(rbind(c(7, 6, 5, 10, 8)))          # 고수정 설문 결과
> ksj
  X1 X2 X3 X4 X5
1  7  6  5 10  8
```

```
> head(FC약자)
  Hamburger Pizza Cat Dog Summer cluster
1        10     9   8   7      6       3
2        10    10   7   6      3       1
3        10    10  10  10      0       2
4        10     8   6   2      5       1
5         6     8   3   9      7       3
6        10    10  10  10      5       3
```

ksj	7	6	5	10	8	?

2. 로지스틱 회귀모델을 이용한 사용자 분류

2.3 로지스틱 회귀모델을 이용한 예측을 위한 새로운 사용자 생성

코드

```
> names(ksj) <- names(FC약자)[1:5]
> ksj
```

```
  Hamburger Pizza Cat Dog Summer
1         7     6   5  10      8
```

```
# 새로운 데이터에 변수이름 부여
# 예측 대상 데이터
```

```
> head(FC약자)
  Hamburger Pizza Cat Dog Summer cluster
1        10     9   8   7      6       3
2        10    10   7   6      3       1
3        10    10  10  10      0       2
4        10     8   6   2      5       1
5         6     8   3   9      7       3
6        10    10  10  10      5       3
```

ksj

	Hamburger	Pizza	Cat	Dog	Summer
1	7	6	5	10	8

?

2. 로지스틱 회귀모델을 이용한 사용자 분류

2.4 로지스틱 회귀모델을 이용한 새로운 데이터 군집 예측

코드

```
# 새로운 데이터
> pred <- predict(mod약자 , ksj)          # 고수정 군집 예측
> pred                                    # 예측 결과 출력
        1
3.011273
```

		Hamburger	Pizza	Cat	Dog	Summer
ksj	1	7	6	5	10	8

Mod약자

$$y(\text{분류}) = (-0.04299) * x_1 + (-0.01183) * x_2 + (0.08036) * x_3 + (0.13993) * x_4 + (0.17516) * x_5 + 0.18082$$

y(분류)=>pred <- $(-0.04299) * x_1 + (-0.01183) * x_2 + (0.08036) * x_3 + (0.13993) * x_4 + (0.17516) * x_5 + 0.18082$

y(분류)=>pred <- 3.011273

2. 로지스틱 회귀모델을 이용한 사용자 분류

응용2. ksj의 cluster는 어떤 값으로 예측되는지 결과를 캡처하시오.

2.4 로지스틱 회귀모델을 이용한 새로운 데이터 군집 예측

코드

```
> round(pred,0)                          # 예측 결과 출력(소수 첫째 자리에서 반올림)
                                         #3.011273을 반올림하면 (가)
  (가)

> ksj$cluster <- round(pred,0)           #ksj의 cluster에 예측 결과 대입
> ksj
  Hamburger Pizza Cat Dog Summer cluster
1         7     6   5  10      8
```

```
> head(FC약자)
  Hamburger Pizza Cat Dog Summer cluster
1        10     9   8   7      6       3
2        10    10   7   6      3       1
3        10    10  10  10      0       2
4        10     8   6   2      5       1
5         6     8   3   9      7       3
6        10    10  10  10      5       3
```

	Hamburger	Pizza	Cat	Dog	Summer	cluster
ksj	1	7	6	5	10	8	(가)

(가)번째 군집으로 분류

오늘도 잘했어요