회귀분석

2017010698
수학과 오서영

목차

0 회귀분석 이란?

회귀분석 : 주어진 데이터로 어떤 함수를 만들어 낸 후,
이 함수를 fitting 하는 작업

- **fitting** : 함수에서 발생하는 차이(잔차의 크기)가 최소화
되도록 함수를 조정해 주는것

- **실행결과** : 어느 정도 신뢰할 수 있는가를 검정해서
통계 예측에 활용가능

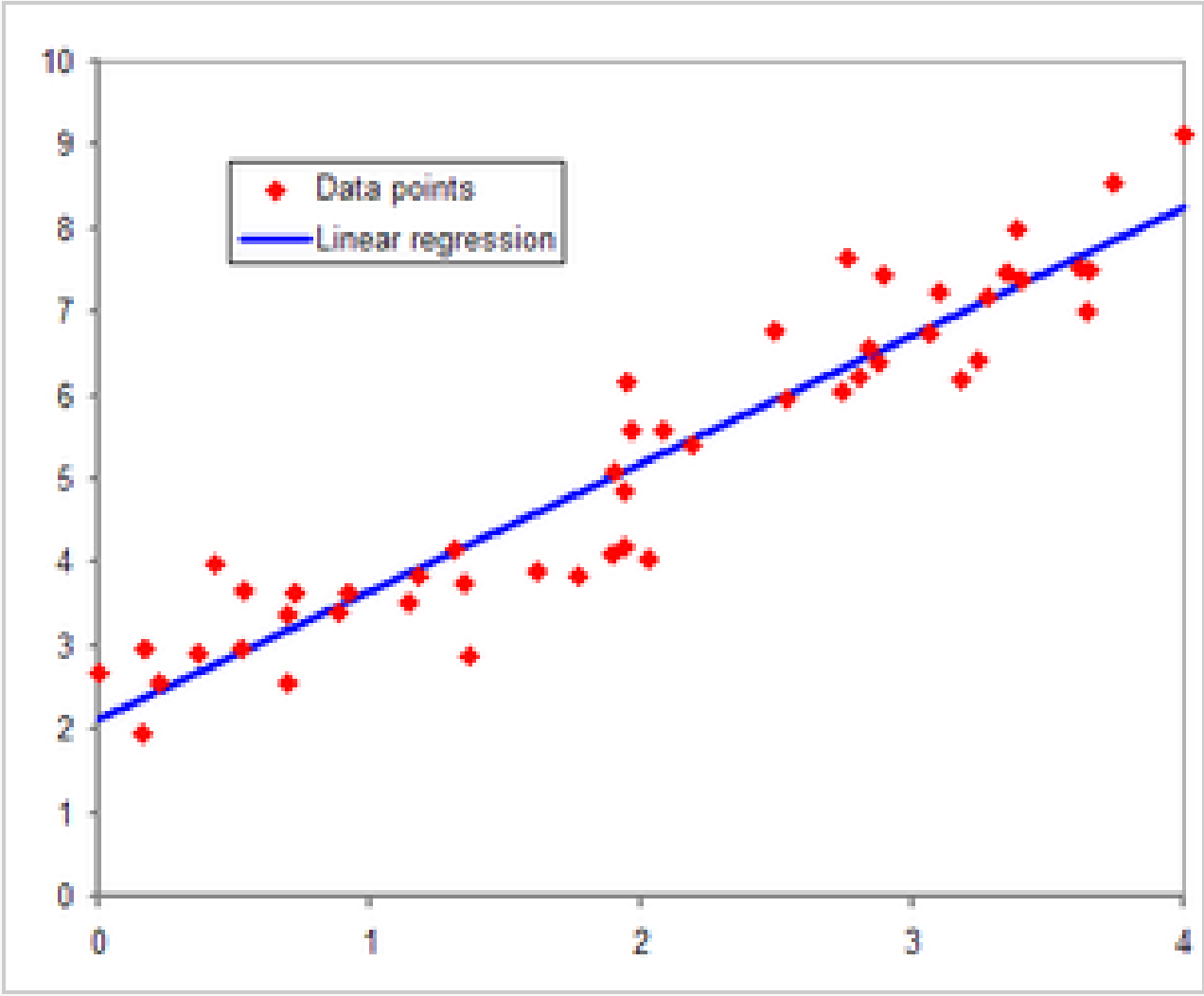
1 단순회귀 분석

단순 회귀 분석 : 종속 변수(y) 와 독립변수(x) 사이의 선형 관계를 파악하고 이를 예측에 활용하는 방법

$$y = Wx + b$$

-하나의 독립변수에 기인하므로 단순회귀라고 한다
- 예) 나이에 따른 평균키 : (키) = W*(나이) + b

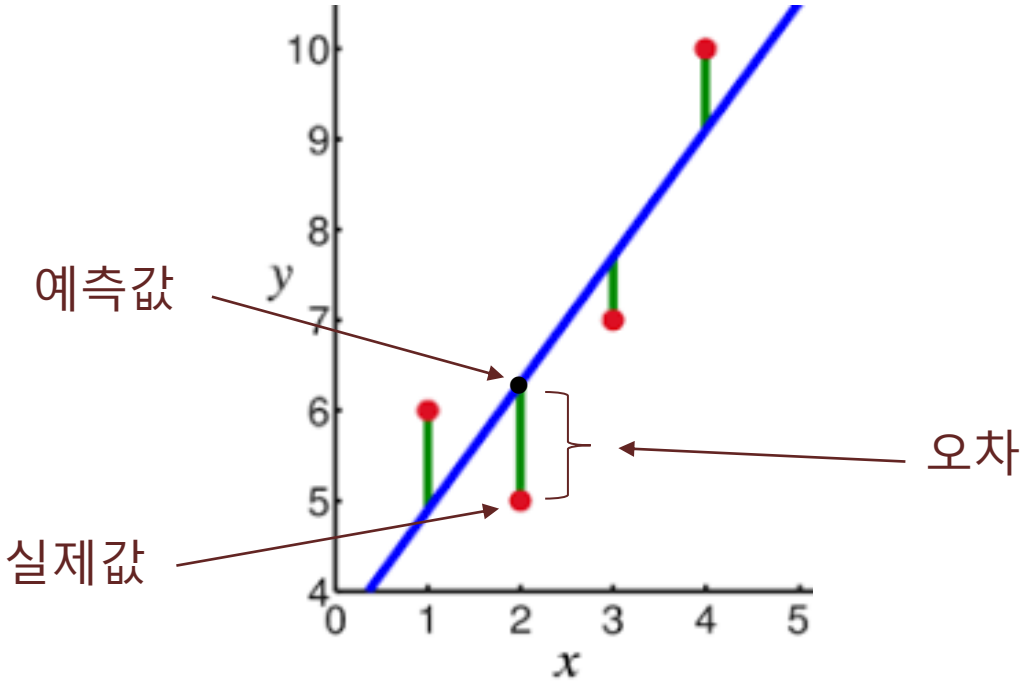
- 목적 : W와 b를 찾는 것



독립변수 1개와 종속변수 1개를 가진 선형 회귀의 예

1 단순회귀 분석

회귀 직선의 W(기울기)와 b(절편) 구하기



1 단순회귀 분석

잔차의 제곱 ↑ "최소 제곱법"

목적함수 $E = \sum_{i=1}^{n} (y_i - wx_i - b)^2$

$$\frac{\partial E}{\partial w} = \sum_{i=1}^{n} (2wx_i^2 + 2x_i(b - y_i)) = 0$$

$$\frac{\partial E}{\partial b} = \sum_{i=1}^{n} (2b + 2(wx_i - y_i)) = 0$$

1

단순회귀 분석

$$a = \frac{n\sum_{k=1}^{n} x_k y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} y_k}{n\sum_{k=1}^{n} x_k^2 - \left(\sum_{k=1}^{n} x_k\right)^2}$$

$$b = \frac{\sum_{k=1}^{n} x_k^2 \sum_{k=1}^{n} y_k - \sum_{k=1}^{n} x_k y_k \sum_{k=1}^{n} x_k}{n\sum_{k=1}^{n} x_k^2 - \left(\sum_{k=1}^{n} x_k\right)^2}$$

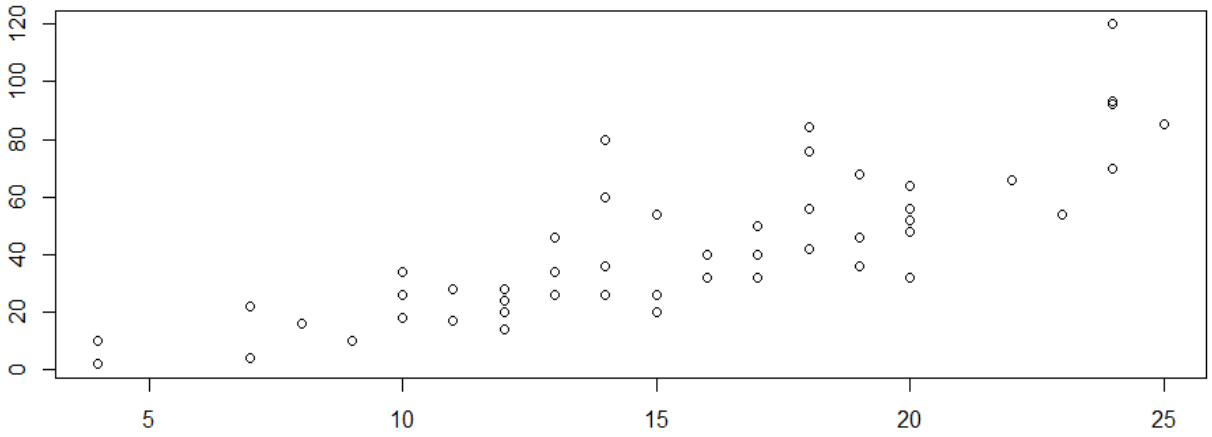
1 단순회귀 분석

R 을 이용하여 회귀 모델 구하기
- 주행속도(speed) 와 제동 거리(dist) 사이의 회귀식

```
head(cars)
plot(dist~speed, data=cars)
```

```
> head(cars)
  speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10
```

mile ft



1 단순회귀 분석

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

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

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

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

```
coef(model)[1] # b
coef(model)[2] # W
```

```
> coef(model)
(Intercept)       speed
 -17.579095    3.932409
> coef(model)[1]
(Intercept)
  -17.57909
> coef(model)[2]
   speed
3.932409
```

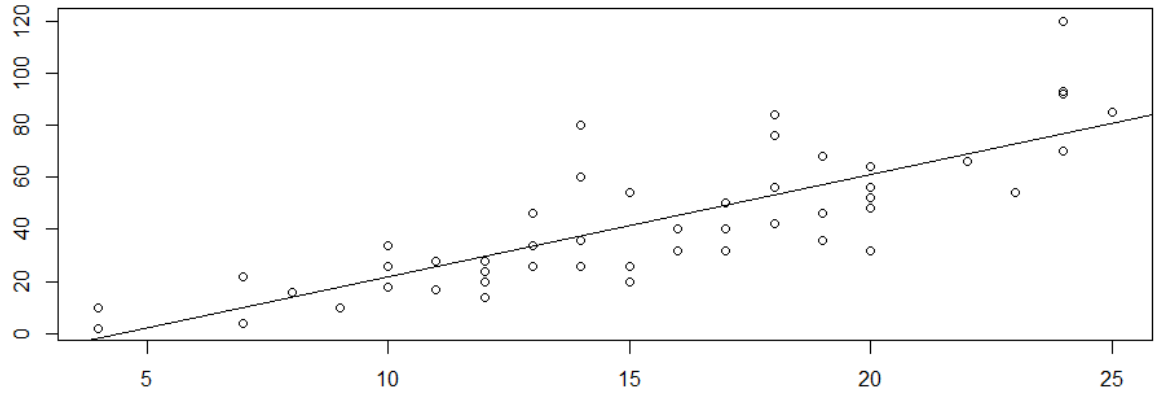
1 단순회귀 분석

완성된 모델

" dist = 3.932 × speed – 17.579 "

회귀식을 산점도에 표현

```
plot(dist~speed, data=cars)
abline(coef(model))
```



2 다중회귀 분석

: 독립변수가 2개 이상인 경우

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \epsilon.$$

2 다중회귀 분석

사례: 연봉 예측 모델

특정 직군의 연봉을 3가지 변수(교육년수, 여성비율, 평판)를 가지고 예측

데이터셋 : car 패키지의 Prestige

```
library(car)
head(Prestige)
```

```
> head(Prestige)
                    education income 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
>
```

직업

교육년수

수입(연봉)

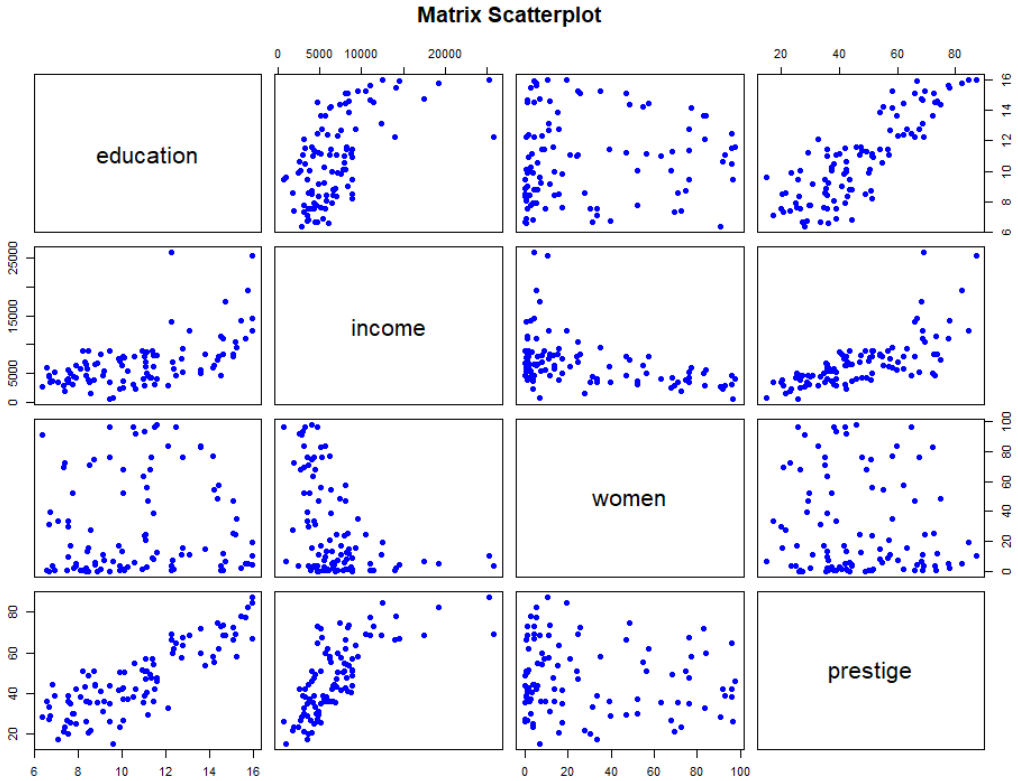
여성비율

직업에 대한 평판

2

다중회귀 분석

```
newdata <- Prestige[,c(1:4)]
plot(newdata, pch=16, col="blue",
     main="Matrix Scatterplot")
```



2 다중회귀 분석

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

```
> summary(mod1)

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

Residuals:
    Min      1Q  Median      3Q     Max
-7715.3  -929.7  -231.2   689.7 14391.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -253.850   1086.157  -0.234    0.816
education    177.199    187.632   0.944    0.347
prestige     141.435     29.910   4.729 7.58e-06
women        -50.896      8.556  -5.948 4.19e-08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2575 on 98 degrees of freedom
Multiple R-squared:  0.6432,    Adjusted R-squared:  0.6323
F-statistic: 58.89 on 3 and 98 DF,  p-value: < 2.2e-16
```

income = - 253.850
+ 177.199 × education
+ 141.435 × prestige
– 50.896 × women

2 다중회귀 분석

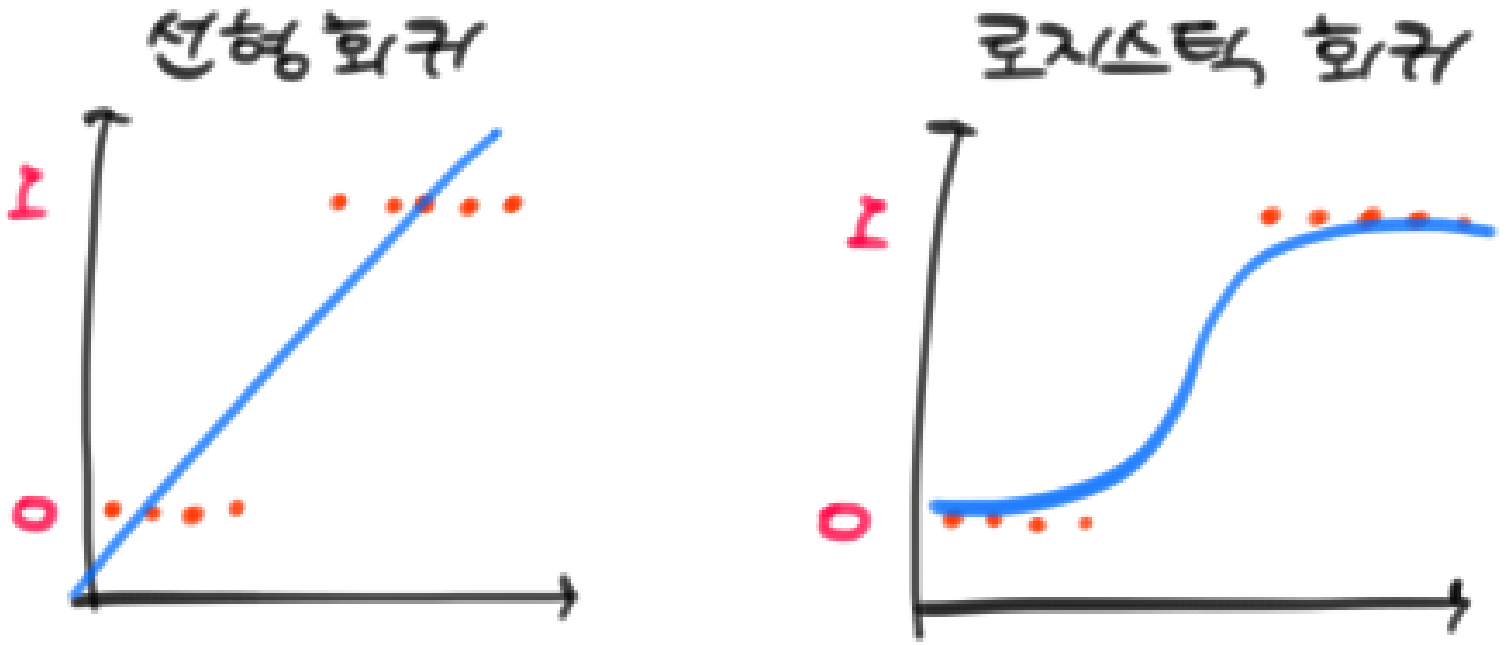
```
> summary(mod1)

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

Residuals:
    Min      1Q  Median      3Q     Max
-7715.3  -929.7  -231.2   689.7 14391.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -253.850   1086.157  -0.234    0.816
education    177.199    187.632   0.944    0.347
prestige     141.435     29.910   4.729 7.58e-06 ***
women        -50.896      8.556  -5.948 4.19e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2575 on 98 degrees of freedom
Multiple R-squared:  0.6432,    Adjusted R-squared:  0.6323
F-statistic: 58.89 on 3 and 98 DF,  p-value: < 2.2e-16
```

income 을 설명하는데
얼마나 중요한 변수인가

모델이 income 을
얼마나 설명할 수 있는가

구한 모델이
의미 있는 모델인가

3 로지스틱 회귀



종속변수가 연속형인 단순, 다중회귀와 다르게
로지스틱 회귀는 종속변수가 0 or 1인 경우 사용한다.

3 로지스틱 회귀

Iris 품종 예측

```
head(iris)
# 종속변수가 숫자형 이어야 함. 범주형 변수를 숫자로 변환
mod3 <- glm(as.integer(Species) ~., data= iris)
summary(mod3)
```

```
> summary(mod3)

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

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.59215  -0.15368   0.01268   0.11089   0.55077

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.18650    0.20484   5.792 4.15e-08 ***
Sepal.Length -0.11191    0.05765  -1.941   0.0542 .
Sepal.Width  -0.04008    0.05969  -0.671   0.5030
Petal.Length  0.22865    0.05685   4.022 9.26e-05 ***
Petal.Width   0.60925    0.09446   6.450 1.56e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.04800419)

    Null deviance: 100.0000  on 149  degrees of freedom
Residual deviance:   6.9606  on 145  degrees of freedom
AIC: -22.874

Number of Fisher Scoring iterations: 2
```

3 로지스틱 회귀

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

```
pred <- 1.18650 + 5.1*(-0.11191)+
     3.5*(-0.04008)+
     1.4*0.22865+
     0.2*0.60925
pred
```

```
> pred
[1] 0.917439
```

1에 가장 가까우므로 1 (setosa)로 판단

```
> unique(iris$Species)
[1] setosa     versicolor virginica
Levels: setosa versicolor virginica
> as.integer(unique(iris$Species))
[1] 1 2 3
```

로지스틱 회귀

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

```
unknown <- data.frame(rbind(c(5.1, 3.5, 1.4, 0.2)))
names(unknown) <- names(iris)[1:4]
unknown
mod3
pred <- predict(mod3, unknown)
pred
```

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

```
> pred
        1
0.9174506
```

3 로지스틱 회귀

```
test <- iris[,1:4]
pred <- predict(mod3, test)
pred
pred <- round(pred,0) # find nearest integer
pred
```

```
> pred <- round(pred,0) # find nearest integer
> pred
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
 20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
 39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57
  1   1   1   1   1   1   1   1   1   1   1   1   2   2   2   2   2   2   2
 58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76
  2   2   2   2   2   2   2   2   2   2   2   2   2   3   2   2   2   2   2
 77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95
  2   2   2   2   2   2   2   3   2   2   2   2   2   2   2   2   2   2   2
 96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114
  2   2   2   2   2   3   3   3   3   3   3   3   3   3   3   3   3   3   3
115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133
  3   3   3   3   3   2   3   3   3   3   3   3   3   3   3   3   3   3   3
134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150
  2   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3
```

3 로지스틱 회귀

얼마나 정확히 예측했는지

```
pred == as.integer(iris[,5])
acc <- mean(pred == as.integer(iris[,5]))
acc
```

```
> pred == as.integer(iris[,5])
    1     2     3     4     5     6     7     8     9    10    11    12    13
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
   14    15    16    17    18    19    20    21    22    23    24    25    26
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
   27    28    29    30    31    32    33    34    35    36    37    38    39
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
   40    41    42    43    44    45    46    47    48    49    50    51    52
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
   53    54    55    56    57    58    59    60    61    62    63    64    65
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
   66    67    68    69    70    71    72    73    74    75    76    77    78
 TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
   79    80    81    82    83    84    85    86    87    88    89    90    91
 TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
   92    93    94    95    96    97    98    99   100   101   102   103   104
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
  105   106   107   108   109   110   111   112   113   114   115   116   117
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
  118   119   120   121   122   123   124   125   126   127   128   129   130
 TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
  131   132   133   134   135   136   137   138   139   140   141   142   143
 TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
  144   145   146   147   148   149   150
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
> acc <- mean(pred == as.integer(iris[,5]))
> acc
[1] 0.9733333
```

3 로지스틱 회귀

```
class(iris$Species)
iris$Species
as.integer(iris$Species)
```

```
> class(iris$Species)
[1] "factor"
> iris$Species
  [1] setosa     setosa     setosa     setosa     setosa     setosa
  [7] setosa     setosa     setosa     setosa     setosa     setosa
 [13] setosa     setosa     setosa     setosa     setosa     setosa
 [19] setosa     setosa     setosa     setosa     setosa     setosa
 [25] setosa     setosa     setosa     setosa     setosa     setosa
 [31] setosa     setosa     setosa     setosa     setosa     setosa
 [37] setosa     setosa     setosa     setosa     setosa     setosa
 [43] setosa     setosa     setosa     setosa     setosa     setosa
 [49] setosa     setosa     versicolor versicolor versicolor versicolor
 [55] versicolor versicolor versicolor versicolor versicolor versicolor
```

```
> as.integer(iris$Species)
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [75] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3
[112] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[149] 3 3
```

끝