R 데이터 분석 입문

Chapter 12

데이터 마이닝 기초 (1)

오 세 종

DANKOOK UNIVERSITY

Contents

개요

- 데이터 마이닝
  - 지금까지는 주로 전통적인 통계 분석 도구를 이용하여 데이터를 분석하는 방법을 배움
  - 이에 더하여 데이터 마이닝 도구를 학습하면 데이터로 부터 다양한 정보를 얻을 수 있음
  - 데이터 마이닝(data mining)은 데이터 안에서 의미 있는 패턴, 추세 등을 발견해 가는 과정
  - 배우게 될 주요 주제: 예측(회귀분석), 분류, 군집화,

1. 단순 선형 회귀

- 정의
  - 종속 변수(y) 와 독립변수(x) 사이의 선형 관계를 파악하고 이를 예측에 활용하는 방법
  - 예) 기온(x) 과 아이스 크림 판매량(y) 사이의 관계식을 찾아낸다. 이를 이용하여 내일의 예상 기온으로 부터 예상 아이스크림 판매량을 예측한다. => 필요한 아이스크림 재료의 양을 예측할 수 있다.
  - 기온(x) 과 아이스 크림 판매량(y) 사이의 관계식을 모델(model) 이라고 한다. (회귀 모델, 예측 모델)
  - 단순 선형 회귀식은 다음과 같은 형태

$$y = Wx + b$$

상수인 W와 b 를 찾는 것이 모델을 만드는 과정

1. 단순 선형 회귀

- 정의
  - 기온(x) 과 아이스 크림 판매량(y) 사이의 관계식이 다음과 같다고 하자

$$y = 2500\,x + 45$$

  - 내일의 예상 기온이 32 도 이면 내일의 아이스크림 예상 판매량은 다음과 같이 예측될 수 있다

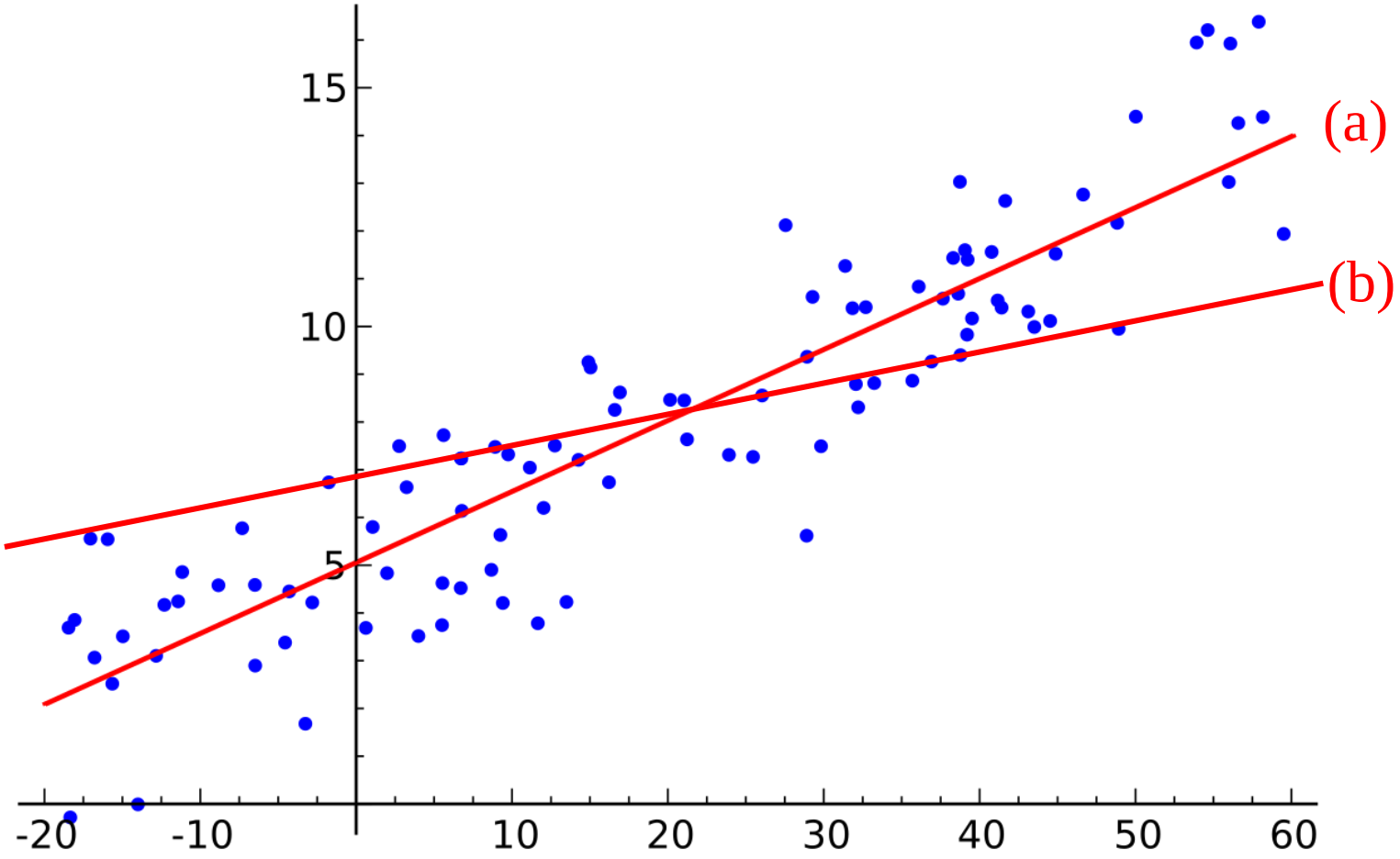
$$y = 2500 \times 32 + 45 = 80045$$

  - 현실세계에서는 두 변수가 선형 관계에 있는 경우가 많아서 선형회귀 분석이 유용하다
  - 두변수가 선형 관계에 있는지 알아보는 방법 : 산점도, 상관계수

Simple linear regression

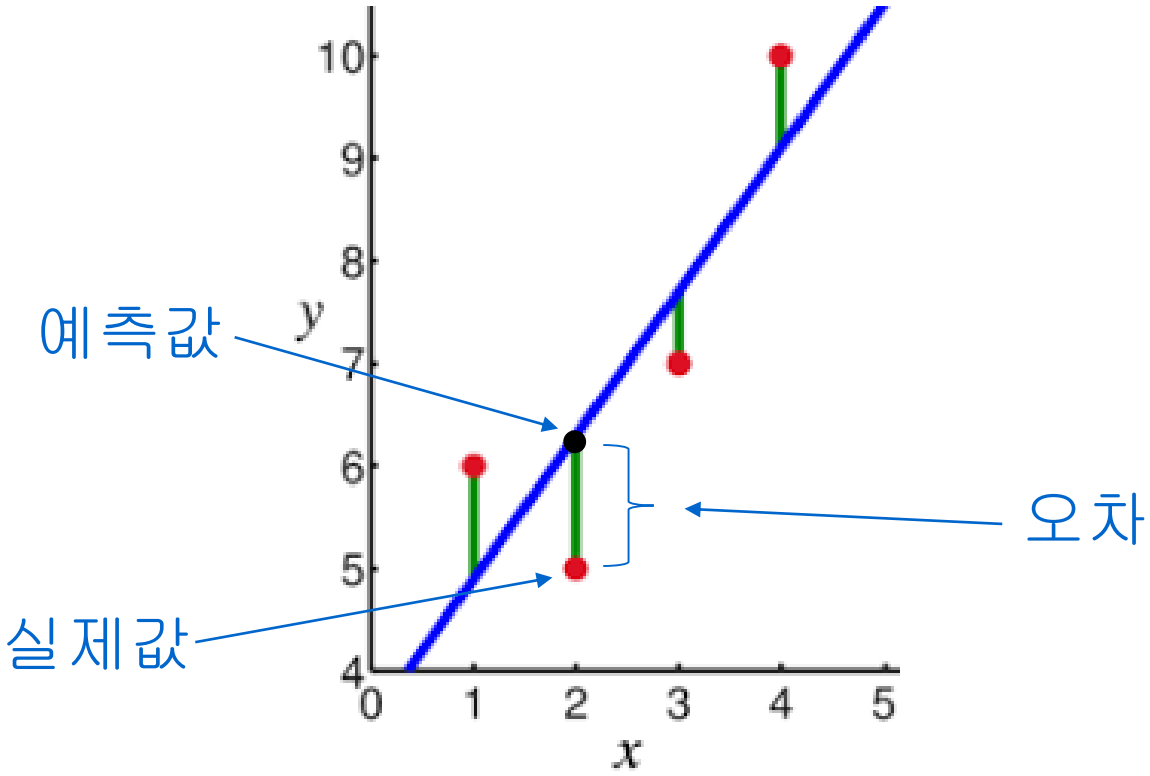
1. 단순 선형 회귀

- 회귀식에서 W와 b를 찾는 방법



1. 단순 선형 회귀

- 회귀식에서 W와 b를 찾는 방법



"최소 제곱법"

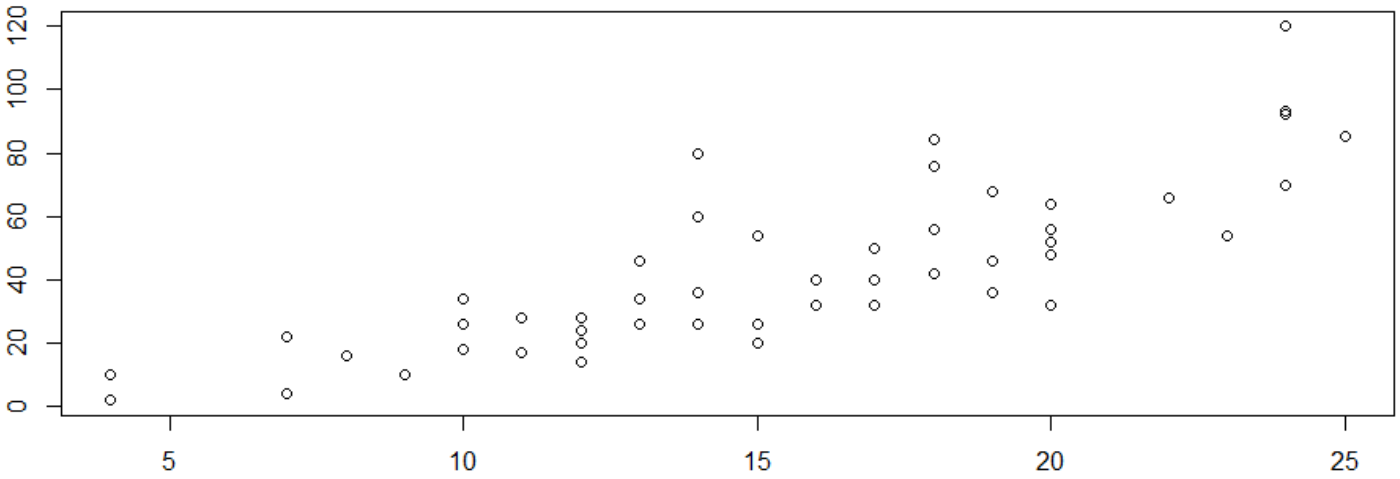
1. 단순 선형 회귀

- 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. 단순 선형 회귀

- R 을 이용하여 회귀 모델 구하기

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

종속 독립

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

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

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

b *W*

```
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. 단순 선형 회귀

- R 을 이용하여 회귀 모델 구하기
  - 완성된 모델

$$dist = 3.932 \times speed - 17.579$$

Speed 가 1 mile 증가할때마다
제동거리가 약 4 ft 씩 증가한다

  - 모델을 이용한 예측 : 속도가 각각 30, 35, 40 일 때 예상 제동 거리는 ?

```
> speed = 30
> dist = 3.932 * speed - 17.579
> print(dist)
[1] 100.381
> speed = 35
> dist = 3.932 * speed - 17.579
> print(dist)
[1] 120.041
> speed = 40
> dist = 3.932 * speed - 17.579
> print(dist)
[1] 139.701
```

1. 단순 선형 회귀

- R 을 이용하여 회귀 모델 구하기
  - 모델의 예측값과 실제값 사이에 어느정도 차이가 나는지 알아보자

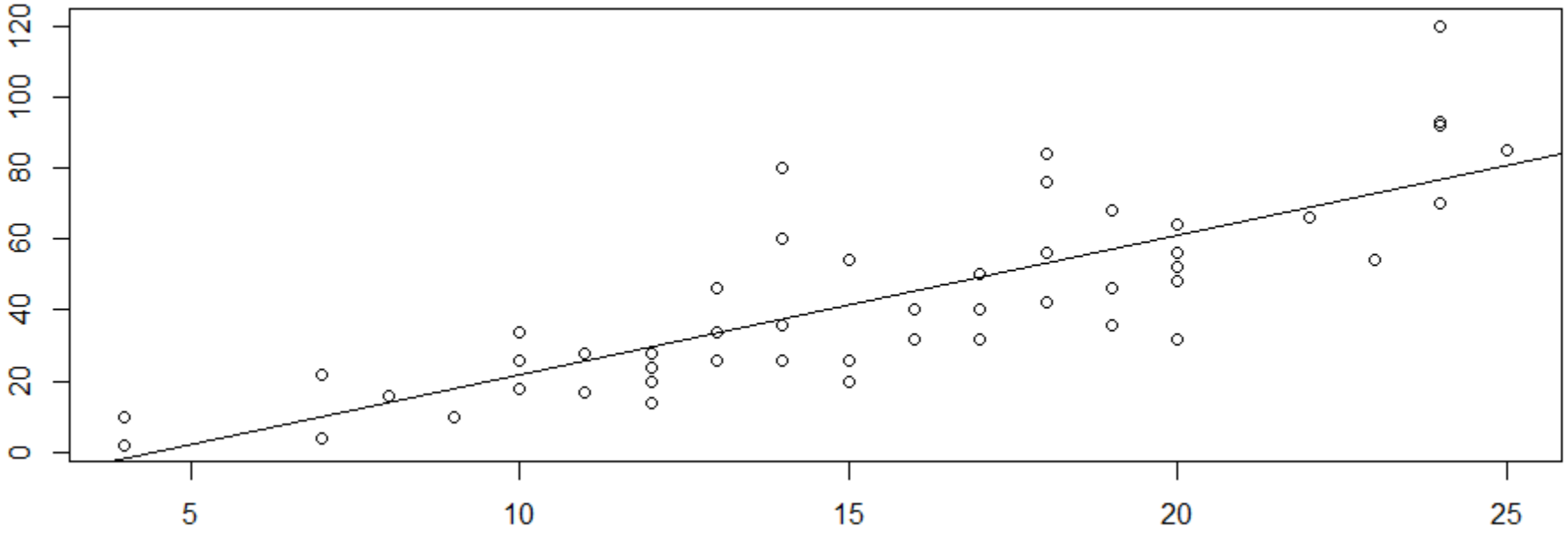
```
speed <- cars[,1]
pred <- 3.932 * speed - 17.579
pred
compare <- cbind(pred,cars[,2],abs(pred-cars[,2]))
compare
```

```
> compare
         pred
 [1,] -1.851   2  3.851
 [2,] -1.851  10 11.851
 [3,]  9.945   4  5.945
 [4,]  9.945  22 12.055
 [5,] 13.877  16  2.123
 [6,] 17.809  10  7.809
 [7,] 21.741  18  3.741
 [8,] 21.741  26  4.259
 [9,] 21.741  34 12.259
[10,] 25.673  17  8.673
[11,] 25.673  28  2.327
[12,] 29.605  14 15.605
[13,] 29.605  20  9.605
[14,] 29.605  24  5.605
[15,] 29.605  28  1.605
```

1. 단순 선형 회귀

- 회귀식을 산점도에 표현

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



[연습문제 1]

1. state.x77 데이터셋에서 문맹률(Illiteracy)을 가지고 살인범죄율(Murder)을 예측하는 회귀 모델을 만드시오

2. 회귀 모델을 이용하여 문맹률이 0.5, 1.0, 1.5 일때 살인범죄율을 예측하여 보시오

2. 중선형 회귀

- 정의
  - Multiple linear regression
  - 독립변수가 2개 이상인 경우
  - 예) 키(x1) 와 몸무게(x2)를 가지고 혈당수치(y)를 를 예측
    - 독립변수 : 키, 몸무게
    - 종속 변수 : 혈당수치
  - 중선형 회귀식의 형태

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

2. 중선형 회귀

- 사례: 연봉 예측 모델
  - 특정 직군의 연봉을 3가지 변수(교육년수, 여성비율, 평판)를 가지고 예측해보자
  - 참조 : https://rpubs.com/FelipeRego/MultipleLinearRegressionInRFirstSteps
  - 데이터셋 : car 패키지의 Prestige

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

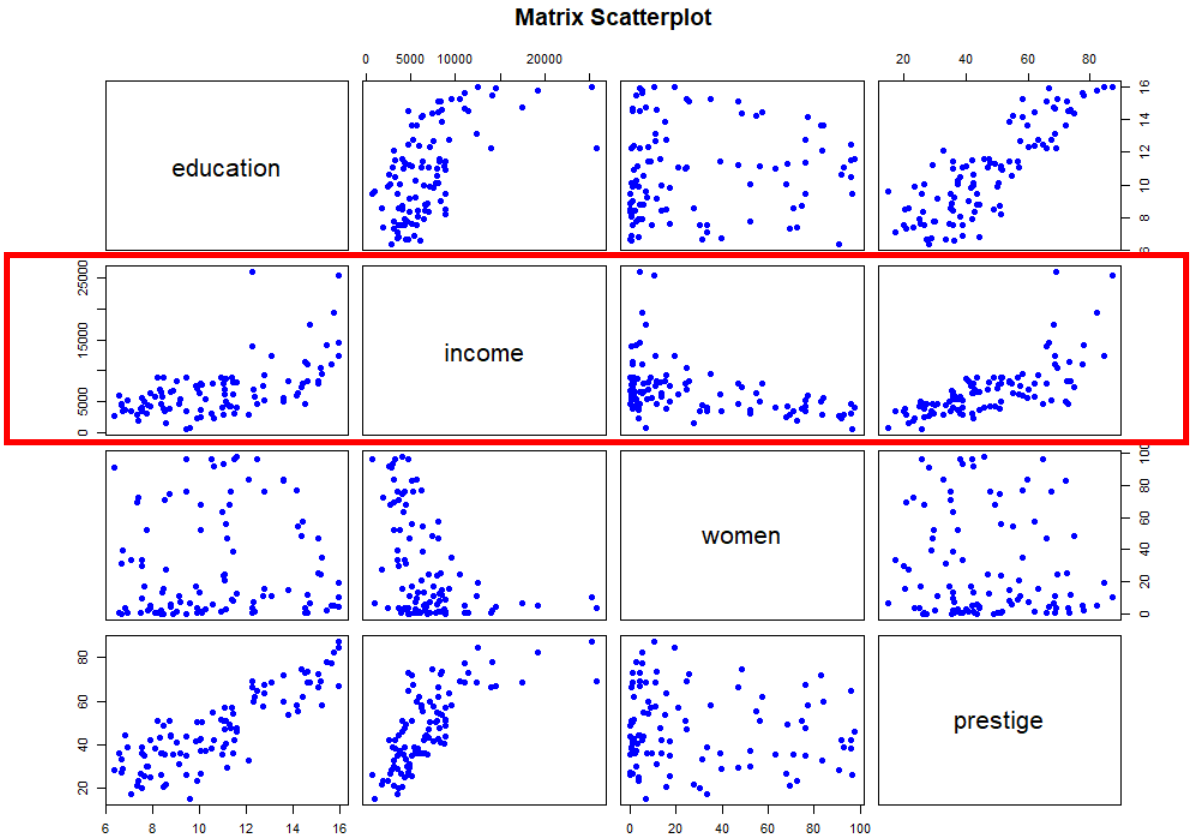
```
> head(Prestige)
                    education income women prestige census type
gov.administrators      13.11  12351 11.16     68.8   1113 prof
general.managers        12.26  25879  4.02     69.1   1130 prof
accountants             12.77   9271 15.70     63.4   1171 prof
purchasing.officers     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 여성비율 수입(연봉) 직업에 대한 평판

2. 중선형 회귀

- 사례: 연봉 예측 모델

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



2. 중선형 회귀

- 사례: 연봉 예측 모델

```
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
```

2. 중선형 회귀

- 사례: 연봉 예측 모델
  - 중선형 회귀식

$$\begin{aligned} \text{income} = -253.850 &+ 177.199 \times \text{education} \\ &+ 141.435 \times \text{prestige} \\ &- 50.896 \times \text{women} \end{aligned}$$

	Estimate
(Intercept)	-253.850
education	177.199
prestige	141.435
women	-50.896

  - 어떤 직군의 평균교육연수가 9.5년, 여성비율이 20%, 평판도가 80 이라면 예상 평균 연봉은?

$$\begin{aligned} \text{income} = -253.850 &+ 177.199 \times 9.5 \\ &+ 141.435 \times 80 \\ &- 50.896 \times 20 \\ &= 13762.26 \end{aligned}$$

2. 중선형 회귀

- 사례: 연봉 예측 모델
  - 모델 평가

```
> 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

2. 중선형 회귀

- 변수 선택
  - 독립변수들이 많을 때 그중에서 종속변수를 잘 설명할 수 있는 변수들만 모아서 모델을 만들면 좋을 것이다.
  - 이 작업을 자동으로 하는 방법이 있다.

```
library(MASS)
newdata2 <- Prestige[,c(1:5)]
head(newdata2)
mod2 <- lm(income ~ education + prestige +
           women + census, data= newdata2)
step <- stepAIC(mod2, direction="both")
```

동일한 명령어

```
mod2 <- lm(income ~., data= newdata2)
```

income 을 제외한 나머지 모든 변수들을 의미

2. 중선형 회귀

```
> step <- stepAIC(mod2, direction="both")
Start:  AIC=1607.93
income ~ education + prestige + women + census

            Df Sum of Sq       RSS    AIC
- census     1    639658 649654265 1606.0
- education  1   5558323 654572930 1606.8
<none>                   649014607 1607.9
- prestige   1 143207106 792221712 1626.3
- women      1 212639294 861653901 1634.8

Step:  AIC=1606.03
income ~ education + prestige + women

            Df Sum of Sq       RSS    AIC
- education  1   5912400 655566665 1605.0
<none>                   649654265 1606.0
+ census     1    639658 649014607 1607.9
- prestige   1 148234959 797889223 1625.0
- women      1 234562232 884216497 1635.5

Step:  AIC=1604.96
income ~ prestige + women

            Df Sum of Sq        RSS    AIC
<none>                   655566665 1605.0
+ education  1   5912400  649654265 1606.0
+ census     1    993735  654572930 1606.8
- women      1 234647032  890213697 1634.2
- prestige   1 811037947 1466604612 1685.1
```

최종

2. 중선형 회귀

- 선택된 변수로 모델을 다시 생성

```
newdata2 <- Prestige[,c(1:5)]
head(newdata2)
mod3 <- lm(income ~ prestige + women,
           data= newdata2)
summary(mod3)
```

```
> summary(mod3)

Call:
lm(formula = income ~ prestige + women, data = newdata2)

Residuals:
    Min      1Q  Median      3Q     Max
-7620.9 -1008.7  -240.4   873.1 14180.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  431.574    807.630   0.534    0.594
prestige     165.875     14.988  11.067  < 2e-16 ***
women        -48.385      8.128  -5.953 4.02e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2573 on 99 degrees of freedom
Multiple R-squared:   0.64,     Adjusted R-squared:  0.6327
F-statistic: 87.98 on 2 and 99 DF,  p-value: < 2.2e-16
```

[연습문제 2]

1. mlbench 패키지의 BostonHousing 데이터셋은 보스턴 지역의 지역정보 및 평균주택 가격 (medv) 정보를 담고 있다. 다른 변수들을 이용하여 medv 를 예측하는 모델을 만드시오 (medv 를 예측하는데 도움이 되는 변수들만 사용할 것)

2. 만들어진 모델로 부터 임의의 데이터에 대한 medv 값을 예측하여 보시오

데이터 불러오기

```
> library(mlbench)
> data(BostonHousing)
> head(BostonHousing)
     crim zn indus chas   nox    rm  age    dis rad tax ptratio      b lstat
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
  medv
1 24.0
2 21.6
3 34.7
4 33.4
5 36.2
6 28.7
```

3. 로지스틱 회귀

- 일반적인 회귀 문제에서는 종속변수가 수치데이터(양적 자료)임

- 예측 해야 할 종속 변수가 수치데이터가 아닌 범주형 데이터 (Yes or No, Patient or Healthy) 일 때를 로지스틱 회귀라고 한다.

- 예) iris 데이터셋에서 4개의 측정 데이터로 부터 품종 (Species) 를 예측해 보자

* 범주나 그룹을 예측하는 문제를 '분류(classification)' 문제 라고 한다

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

- iris 품종 예측
  - 다음 데이터의 품종을 예측해보자

```
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
```

3. 로지스틱 회귀

- iris 품종 예측
  - 다음 데이터의 품종을 예측해보자 (model, predict() 함수 이용)

```
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

- iris 품종 예측
  - 여러 개의 데이터에 대해 예측할 때

```
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

- iris 품종 예측
  - 얼마나 정확히 예측했는지 알아보자

```
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

- Note
  - 로지스틱 회귀의 경우에도 종속 변수가 숫자여야 하기 때문에 문자형으로 되어 있는 범주데이터는 숫자(1,2,3,..)로 변환한후 작업을 해야 한다.
  - 범주형 데이터가 factor 이면 as.integer() 함수를 통해 쉽게 숫자로 바꿀수 있다

```
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 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 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 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
```

실습x

[연습문제 3]

1. ucla_admit.csv 의 데이터셋으로 부터 gre, gpa, rank 를 가지고 합격여부(admit) 를 예측하는 로지스틱 모델을 만드시오 (0: 불합격, 1:합격)

2. 만들어진 모델에 대해 ucla_admit.csv 의 데이터를 넣어 합격여부를 예측하고 실제값과 예측값을 보이시오.

3. 만들어진 모델의 예측 정확도를 보이시오

ucla_admit.csv

admit	gre	gpa	rank
0	380	3.61	3
1	660	3.67	3
1	800	4	1
1	640	3.19	4
0	520	2.93	4
1	760	3	2
1	560	2.98	1
0	400	3.08	2
1	540	3.39	3
0	700	3.92	2
0	800	4	4
0	440	3.22	1