Lecture 12 - Linear Regression

Sim, Min Kyu, Ph.D., mksim@seoultech.ac.kr

SEOULTECH 국립 서울과학기술대학교 IT정책전문대학원

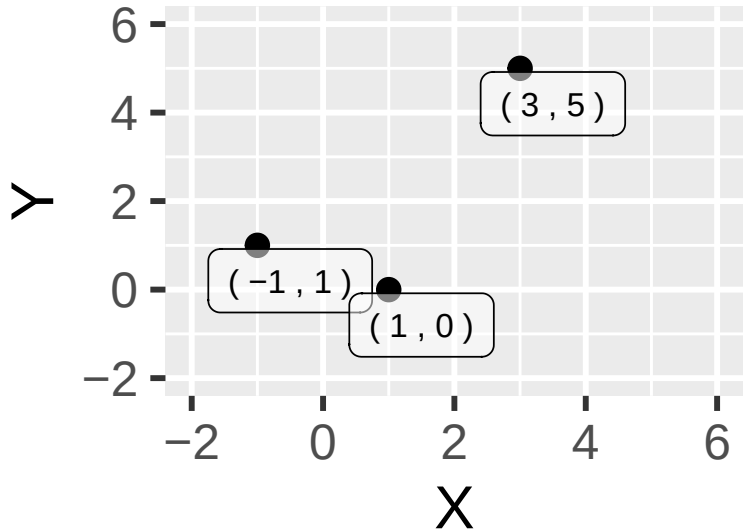
1. I. Minimal example
2. II. 선형회귀의 일반식
3. III. Performance Measure
4. IV. Built-in R function
5. V. Multiple Linear Regression

I. Minimal example

Motivation

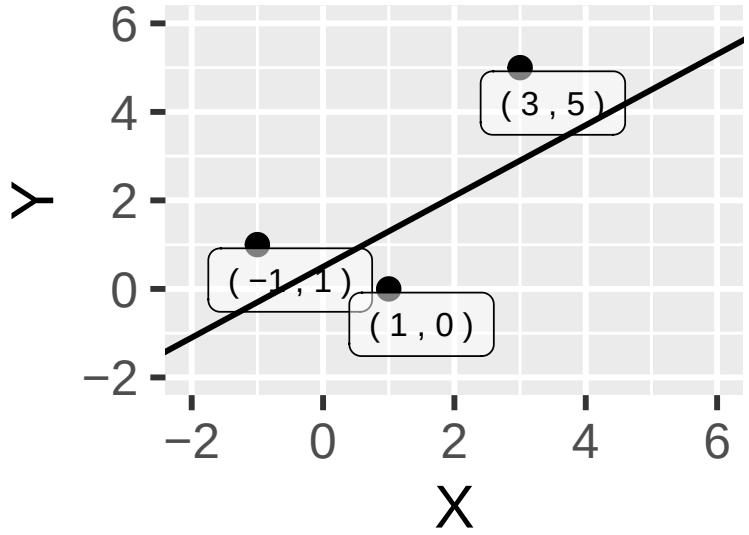
Minimal setting

- 어떤 실험으로 3개의 표본을 얻었고, 각 표본의 (X, Y) 수치가 다음과 같다.
  - $(X_1, Y_1) = (-1, 1)$
  - $(X_2, Y_2) = (1, 0)$
  - $(X_3, Y_3) = (3, 5)$



선형회귀(Linear regression)의 목적

- 관측치들의 "중간"을 통과하는 직선 함수를 찾아내는 것이다.



- 주어진 X_i에 대해서 $\widehat{Y_i}$를 만든다.
  - $\widehat{Y_i}$는 Y_i에 대한 추정치에 해당한다.
  - 회귀직선은 $Y_i - \widehat{Y_i}$의 크기가 작게 되도록 중간을 잘 지나가야 된다.
- 수식으로 요약하면 아래와 같다.
  - $\widehat{Y_i} = \alpha + \beta X_i$
  - $\epsilon_i = Y_i - \widehat{Y_i}$
  - $Y_i = \alpha + \beta X_i + \epsilon_i$

잔차(residual)의 의미

- 실제값(Y_i)과 회귀모형을 이용한 추정치($\widehat{Y_i}$)의 차이이다.
- $\epsilon_i = Y_i - \widehat{Y_i}$
- 회귀모형(regression model)이 미처 설명하지 못하는 부분에 해당하므로 **에러(error)**의 의미를 가진다.

회귀모형을 만드는 것

$$\widehat{Y_i} = \alpha + \beta X_i$$

$$\epsilon_i = Y_i - \widehat{Y_i}$$

- 회귀모형을 만드는 것은 각각의 샘플에서 발생하는 잔차의 총합을 줄이는 α와 β를 찾는 것이다.

회귀 분석의 목적함수

잔차의 총합에 대한 정의

- 방법 1. 적절한 α와 β를 찾아서 $|\epsilon_1| + |\epsilon_2| + |\epsilon_3|$를 최소화 한다.
- 방법 2. 적절한 α와 β를 찾아서 $\epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2$를 최소화 한다.

- 방법 1
  - 비전문가에게 설명하기 쉽다.
  - Outlier에 대한 영향이 방법2보다 적다.
- 방법 2
  - 최소자승법(LS, least-square method)이라고 불린다.
  - 미분후에 연속함수가 등장하기에 계산이 쉽다.
  - Outlier에 대한 영향이 크며, 이에 대한 대안으로 여러가지의 회귀분석에 대한 변형 기법들이 발전되어 있다.

Risk, Loss, Cost, and Penalty.

- 방법 1의 절대값 함수나 방법 2의 제곱 함수는
  - 하나의 샘플에 대한 에러를 측정하는 함수이며, 음수값을 가질 수 없다.
  - 고전통계에서는 Risk 함수라고 부른다.
  - 기계학습에서는 Loss 함수라고 부른다.

- Risk함수나 Loss함수의 합은
  - 모형에서 최소화시켜야할 목적함수에 해당한다.
  - 고전통계에서는 Penalty 함수라고 부른다.
  - 기계학습에서는 Cost 함수라고 부른다.

- 벌점과 벌금이 최적의 운전 습관을 결정하듯이, Penalty함수가 α와 β를 찾아 회귀모형을 결정한다.

- “If you cannot measure it, you cannot improve it.” - Peter Drucker

Minimal example의 α와 β를 찾기

Step 0. Setting

0. $(X_1, Y_1) = (-1, 1), (X_2, Y_2) = (1, 0), (X_3, Y_3) = (3, 5)$
1. $\widehat{Y_i} = \alpha + \beta X_i$
2. $\epsilon_i = Y_i - \widehat{Y_i}$
3. $L = \epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2$
4. L을 최소화 하는 α와 β를 찾자!

Step 1. $\widehat{Y_i} = \alpha + \beta X_i$

$$\begin{aligned} \widehat{Y_1} &= \alpha + \beta X_1 = \alpha + \beta(-1) \\ \widehat{Y_2} &= \alpha + \beta X_2 = \alpha + \beta(1) \\ \widehat{Y_3} &= \alpha + \beta X_3 = \alpha + \beta(3) \end{aligned}$$

Step 2. $\epsilon_i = Y_i - \widehat{Y_i}$

$$\begin{aligned}
\epsilon_1 &= Y_1 - \widehat{Y_1} = 1 - (\alpha - \beta) \\
\epsilon_2 &= Y_2 - \widehat{Y_2} = 0 - (\alpha + \beta) \\
\epsilon_3 &= Y_3 - \widehat{Y_3} = 5 - (\alpha + 3\beta)
\end{aligned}$$

Step 3. $L = \epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2$

$$\begin{aligned}
L &:= \epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2 \\
&= (1 - \alpha + \beta)^2 + (\alpha + \beta)^2 + (5 - \alpha - 3\beta)^2 \\
&= 3\alpha^2 + 11\beta^2 + 6\alpha\beta - 12\alpha - 28\beta + 26 \qquad (1)
\end{aligned}$$

Step 4. Minimization

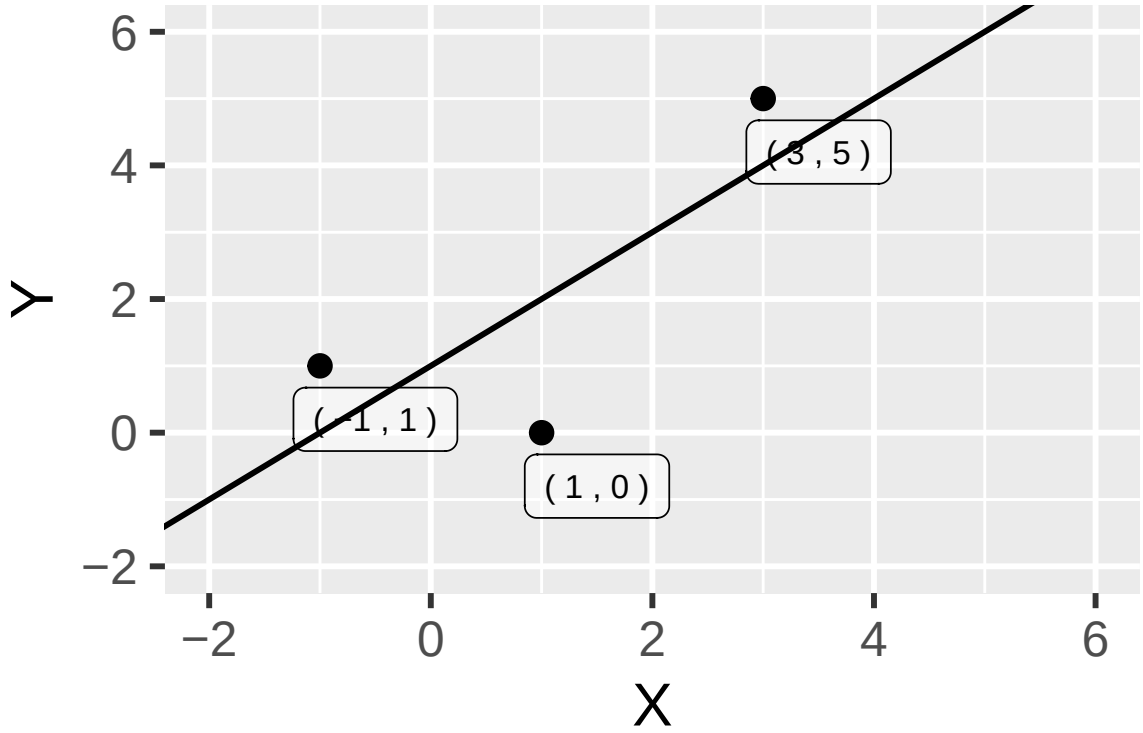
- (Review) $y = 3x^2 - 5x + 4$가 최소가 되는 x의 지점은 $y' = 6x - 5$가 0이 되는 지점인 $x = 5/6$ 이다.
- 마찬가지로 L이 최소가 되는 지점에서는
  - L을 α에 관하여 미분한 것이 0이 된다.
  - L을 β에 관하여 미분한 것이 0이 된다.

$$\frac{\partial L}{\partial \alpha} = 6\alpha + 6\beta - 12 \quad (2)$$

$$\frac{\partial L}{\partial \beta} = 22\beta + 6\alpha - 28 \quad (3)$$

- 위의 두 식이 모두 0이 되는 α와 β의 값은 $\hat{\alpha} = 1$, $\hat{\beta} = 1$이다.
- 따라서 선형회귀식은 $\widehat{Y_i} = 1 + X_i$ 이다.

정리



- 몇 가지 용어
  - α: intercept
  - β: slope
  - $\sum \epsilon_i^2$: 잔차의 제곱합 (Sum of Squared Error, SSE)

II. 선형회귀의 일반식

Setting

- Given $(X_i, Y_i),\ i = 1, 2, 3, ..., n$,
- We shall find a linear relationship between X and Y, $Y_i = \alpha + \beta X_i + \epsilon_i$,
- where $L = \sum \epsilon_i^2$ is to be minimized by choosing the suitable $\hat{\alpha}$ and $\hat{\beta}$.

Step 1. Normal equation (정규방정식)

$$L = \sum \epsilon_i^2 = \sum (Y_i - \widehat{Y_i})^2 = \sum (Y_i - \alpha - \beta X_i)^2$$

$$\frac{\partial L}{\partial \alpha} = -\sum 2(Y_i - \alpha - \beta X_i) = 0 \tag{4}$$

$$\Rightarrow \quad \sum Y_i = \sum (\alpha + \beta X_i)$$

$$\Rightarrow \quad \frac{\sum Y_i}{n} = \alpha + \beta \frac{\sum X_i}{n}$$

$$\Rightarrow \quad \overline{Y} = \alpha + \beta \overline{X} \tag{5}$$

$$\frac{\partial L}{\partial \beta} = -\sum 2(Y_i - \alpha - \beta X_i)X_i = 0 \tag{6}$$

$$\Rightarrow \quad \sum (Y_i - \alpha - \beta X_i)X_i = 0$$

$$\Rightarrow \quad \sum (Y_i - \widehat{Y_i})X_i = 0 \tag{7}$$

Step 2. α와 β의 계산

$$\begin{aligned} \widehat{Y}_i &= \alpha + \beta X_i \\ \overline{Y} &= \alpha + \beta \overline{X} \text{ (from (5))} \end{aligned} \tag{8}$$

- 위의 두 식에서 각각 좌변과 우변을 빼주면, 아래의 등식이 성립한다. (α를 제거하는 일종의 수학적 트릭)

$$\begin{aligned} \Rightarrow \quad & \widehat{Y}_i - \overline{Y} = \beta(X_i - \overline{X}) \\ \Rightarrow \quad & \widehat{Y}_i - Y_i + Y_i - \overline{Y} = \beta(X_i - \overline{X}) \\ \Rightarrow \quad & (\widehat{Y}_i - Y_i) + (Y_i - \overline{Y}) = \beta(X_i - \overline{X}) \\ \Rightarrow \quad & (\widehat{Y}_i - Y_i)(X_i - \overline{X}) + (Y_i - \overline{Y})(X_i - \overline{X}) = \beta(X_i - \overline{X})^2 \\ \Rightarrow \quad & \sum(\widehat{Y}_i - Y_i)(X_i - \overline{X}) + \sum(Y_i - \overline{Y})(X_i - \overline{X}) = \beta \sum(X_i - \overline{X})^2 \end{aligned}$$

(Step 2. continued)

- 마지막 식에서 좌변의 첫 번째 항은 0이다 (Step. 3에 의해서). 따라서 아래와 같이 $\hat{\beta}$와 $\hat{\alpha}$에 대한 식을 얻는다.

$$\hat{\beta} = \frac{\sum(Y_i - \overline{Y})(X_i - \overline{X})}{\sum(X_i - \overline{X})^2} = \frac{SXY}{SXX} \tag{9}$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} \text{ (from (5))} \tag{10}$$

Step 3. Why $\sum(\widehat{Y_i} - Y_i)(X_i - \overline{X}) = 0$?

$$
\begin{aligned}
& \sum(\widehat{Y_i} - Y_i)(X_i - \overline{X}) \\
= \; & \sum(\widehat{Y_i} - Y_i)(X_i) - \overline{X}\sum(\widehat{Y_i} - Y_i) \\
= \; & 0 - 0
\end{aligned}
$$

(First term: from (7), Second term: sum of residuals)

General formula to the minimal example.

1. Data

```
X <- c(-1,1,3)
Y <- c(1,0,5)
n <- length(X)
```

2. Some statistics

```
mean(X)
```

```
## [1] 1
```

```
mean(Y)
```

```
## [1] 2
```

```
sum(X^2)
```

```
## [1] 11
```

```
sum(Y^2)
```

```
## [1] 26
```

```
sum(X*Y)
```

```
## [1] 14
```

3. Estimators

- $SXY = \sum X_i Y_i - n\overline{XY}$
- $SXX = \sum X_i^2 - n\overline{X}^2$
- $\hat{\beta} = \frac{SXY}{SXX}$
- $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$

```
SXY <- sum(X*Y)-n*mean(X)*mean(Y)
SXX <- sum(X^2)-n*mean(X)^2
beta_hat <- SXY/SXX
beta_hat
```

```
## [1] 1
```

```
alpha_hat <- mean(Y)-beta_hat*mean(X)
alpha_hat
```

```
## [1] 1
```

III. Performance Measure

Sum of Squares (SST, SSR, SSE)

- Notations
  - Y_i: 실제값
  - $\widehat{Y_i}$: 추정값
  - $\overline{Y}$: Y_i 들의 평균

$$
\begin{aligned}
\sum (Y_i - \overline{Y})^2 &= \sum (Y_i - \widehat{Y_i} + \widehat{Y_i} - \overline{Y})^2 \\
&= \sum [(Y_i - \widehat{Y_i}) + (\widehat{Y_i} - \overline{Y})]^2 \\
&= \sum [(Y_i - \widehat{Y_i})^2 + (\widehat{Y_i} - \overline{Y})^2 + 2(Y_i - \widehat{Y_i})(\widehat{Y_i} - \overline{Y})] \\
&= \sum (Y_i - \widehat{Y_i})^2 + \sum (\widehat{Y_i} - \overline{Y})^2 \\
SST &= SSE + SSR
\end{aligned}
$$

$$\begin{aligned} \sum (Y_i - \overline{Y})^2 &= \sum (Y_i - \widehat{Y}_i)^2 + \sum (\widehat{Y}_i - \overline{Y})^2 \\ SST &= SSE + SSR \end{aligned}$$

- SST
  - Sum of Squared Total
  - Y_i의 편차의 제곱합
  - Y_i의 total variation
  - Regression으로 설명해야할 타겟
- SSE
  - Sum of Squared Error
  - 에러의 제곱합, 잔차의 제곱합
  - Cost 함수, Penalty 함수에 해당
  - 작을수록 좋음
  - SST에 비해서 상대적으로 작으면 좋음
  - 모델의 개선을 통해서 향상 시킬수 있는 영역
- SSR
  - SST중에서 Regression하는 부분
  - 클수록 좋음
  - $R^2 = \frac{SSR}{SST}$ 은 전체의 variation (SST)를 Regression이 설명하는 비율을 의미하며, 회귀에서 가장 중요한 퍼포먼스 척도가 된다.

IV. Built-in R function

```r
X <- c(-1,1,3)
Y <- c(1,0,5)
lm(Y ~ X) # linear model
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
## (Intercept)            X
##           1            1
```

```
summary(lm(Y ~ X))
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##  1  2  3
##  1 -2  1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.000      1.658   0.603    0.655
## X              1.000      0.866   1.155    0.454
##
## Residual standard error: 2.449 on 1 degrees of freedom
## Multiple R-squared:  0.5714, Adjusted R-squared:  0.1429
## F-statistic: 1.333 on 1 and 1 DF,  p-value: 0.4544
```

V. Multiple Linear Regression

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

- Y: mpg
- X_1: disp
- X_2: wt

```
lm(mtcars$mpg ~ mtcars$disp + mtcars$wt)
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ mtcars$disp + mtcars$wt)
##
## Coefficients:
## (Intercept)  mtcars$disp    mtcars$wt
##    34.96055     -0.01772     -3.35083
```

$$\widehat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2$$

$$\widehat{Y} = 34.96 - 0.018 X_1 - 3.35 X_2$$

```
summary(lm(mtcars$mpg ~ mtcars$disp + mtcars$wt))
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ mtcars$disp + mtcars$wt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4087 -2.3243 -0.7683  1.7721  6.3484
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.96055    2.16454  16.151 4.91e-16 ***
## mtcars$disp -0.01773    0.00919  -1.929  0.06362 .
## mtcars$wt   -3.35082    1.16413  -2.878  0.00743 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.917 on 29 degrees of freedom
## Multiple R-squared:  0.7809, Adjusted R-squared:  0.7658
## F-statistic: 51.69 on 2 and 29 DF,  p-value: 2.744e-10
```

"Man can learn nothing unless he proceeds from the known to the unknown - Claude Bernard"