

## Chap 2: Linear Regression

---

Seong yeon Park

2025-01-24

Department of Statistic  
Sungshin Women's University

1 Least Squares Method

2 Multiple Regression

3 Distribution of  $\hat{\beta}$

- The data consists of  $(x_1, y_1), \dots, (x_N, y_N)$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- $\beta_0$ : intercept
- $\beta_1$ : slope
- $\varepsilon_i$ : random error
- We obtain  $\beta_0$  and  $\beta_1$  via the least squares method.

# Least Squares Method

- sum of squares of the residuals,  
we minimize  $L$  of the squared distances  $L$  between  $(x_i, y_i)$  and  $(x_i, \beta_0 + \beta_1 x_i)$   
over  $i = 1, 2, \dots, N$ .

$$L = \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_i)^2$$

- Then, by partially differentiating  $L$  by  $\beta_0, \beta_1$  and letting them be zero.

$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^N (x_i (y_i - \beta_0 - \beta_1 x_i)) = 0$$

- $\beta_0$  and  $\beta_1$  are regarded as constants when differentiating  $L$  by  $\beta_1$  and  $\beta_0$ .

# Least Squares Method

- When  $\sum_{i=1}^N (x_i - \bar{x})^2 \neq 0$ ,  $\hat{\beta}_0, \hat{\beta}_1$  instead of  $\beta_0, \beta_1$  which means that they are not the true values but rather estimates obtained from data.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- We center the data as follows,

$$\tilde{x}_1 := x_1 - \bar{x}, \dots, \tilde{x}_N := x_N - \bar{x}, \tilde{y}_1 := y_1 - \bar{y}, \dots, \tilde{y}_N := y_N - \bar{y}$$

- Center the data results in a zero average.
- The formula for calculating the slope from the centralized data is as follows:

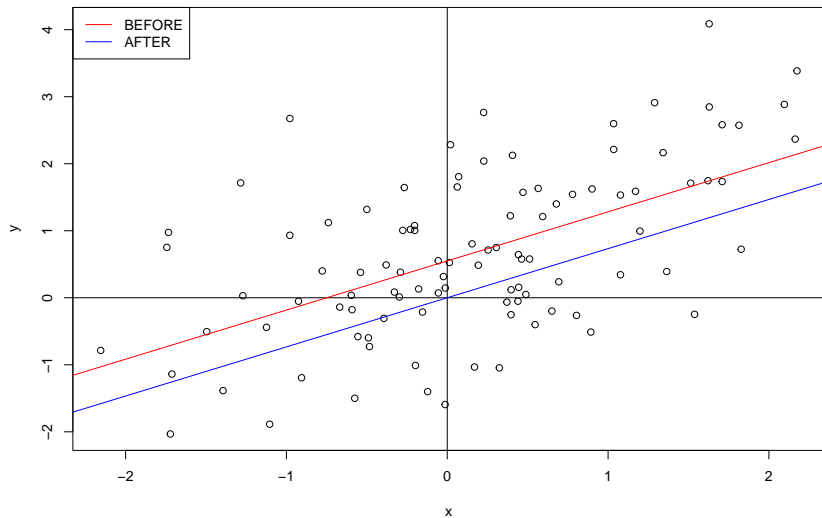
$$\hat{\beta}_1 = \frac{\sum_{i=1}^N \tilde{x}_i \tilde{y}_i}{\sum_{i=1}^N (\tilde{x}_i)^2}$$

## Example

- The two lines  $l$  is obtained from the  $N$  pairs of data and the least squares method, and  $l'$  obtained by shifting  $l$  so that it goes through the origin.

```
min.sq=function(x,y){  
  x.bar=mean(x);y.bar=mean(y)  
  beta.1=sum((x-x.bar)*(y-y.bar))/sum((x-x.bar)^2);beta.0=y.bar-beta.1*x.bar  
  return(list(a=beta.0,b=beta.1))  
}  
a=rnorm(1);b=rnorm(1);  
N=100;x=rnorm(N);y=a*x+b+rnorm(N)  
plot(x,y);abline(h=0);abline(v=0)  
abline(min.sq(x,y)$a,min.sq(x,y)$b,col="red")  
x=x-mean(x);y=y-mean(y)  
abline(min.sq(x,y)$a,min.sq(x,y)$b,col="blue")  
legend("topleft",c("BEFORE","AFTER"),lty=1,col=c("red","blue"))
```

## Example



1 Least Squares Method

2 Multiple Regression

3 Distribution of  $\hat{\beta}$



# Multiple Regression with Matrices

We formulate the least squares method for multiple regression with matrices.

- $L := \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_i)^2,$

$$L = \|y - X\beta\|^2 = (y - X\beta)^T (y - X\beta)$$

- If we define,

$$y := \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, X := \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,p} \end{bmatrix}, \beta := \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

- Partial differentiation with  $L$

$$\nabla L := \begin{bmatrix} \frac{\partial L}{\partial \beta_0} \\ \frac{\partial L}{\partial \beta_1} \end{bmatrix} = -2X^T (y - X\beta)$$

- Set to zero to find the minimum value

$$-2X^T(y - X\beta) = \begin{bmatrix} -2 \sum_{i=1}^N (y_i - \sum_{j=0}^p \beta_j x_{i,j}) \\ -2 \sum_{i=1}^N x_{i,1} (y_i - \sum_{j=0}^p \beta_j x_{i,j}) \\ \vdots \\ -2 \sum_{i=1}^N x_{i,p} (y_i - \sum_{j=0}^p \beta_j x_{i,j}) \end{bmatrix}$$

- When a matrix  $X^T X$  is invertible, we have

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

## When $X^T X$ is irreversible

1.  $N < p + 1$

$$\text{rank}(X^T X) \leq \text{rank}(X) \leq \min\{N, p + 1\} = N < p + 1$$

If  $N > p$ , It is  $X$ \_particular, So there is no inverse matrix.

2. Two columns in  $X$  coincide.

$$X^T X z = 0 \Rightarrow z^T X^T X z = 0 \Rightarrow \|X_z\|^2 = 0 \Rightarrow X_z = 0$$

1 Least Squares Method

2 Multiple Regression

3 Distribution of  $\hat{\beta}$

- $y$  have been obtained from the covariates  $X$  multiplied by the (true) coefficients  $\beta$  plus some noise  $\epsilon$ .

$$y = X\beta + \epsilon$$

- The true  $\beta$  is unknown and different from the estimate  $\hat{\beta}$ .
- We have estimated  $\hat{\beta}$  via the least squares method from the  $N$  pairs of data  $(x_1, y_1), \dots, (x_N, y_N) \in R^p \times R$
- $x_i \in R^p$  is the row vector consisting of  $p$  values excluding the leftmost one in the  $i$ th row of  $X$ .

- We assume that each element  $\epsilon_1, \dots, \epsilon_N$  in the random variable  $\epsilon$  is independent of the others and Gaussian distribution with mean zero and variance  $\sigma^2$ .  $N(0, \sigma^2)$

$$f_i(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\epsilon_i^2}{2\sigma^2}}$$

- We may express the distributions of  $\epsilon_1, \dots, \epsilon_N$  by

$$f(\epsilon) = \prod_{i=1}^N f_i(\epsilon_i) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{\epsilon^T \epsilon}{2\sigma^2}}$$

This is  $N(0, \sigma^2 I)$ ,  $I$  is a unit matrix of size  $N$ .

## Independent if and only if their covariance is zero

- For the proof,

$$\hat{\beta} = (X^T X)^{-1} X^T (X\beta + \epsilon) = \beta + (X^T X)^{-1} X^T \epsilon$$

- Since the average of  $\epsilon \in R^N$  is zero, the average of  $\epsilon$  multiplied from left by the constant matrix  $(X^T X)^{-1} X^T$  is zero.

$$E[\hat{\beta}] = \beta$$

- In general, we say that an estimate is unbiased if its average coincides with the true value.

## Covariance matrix of $\hat{\beta}$

- $\hat{\beta}$  and its average  $\beta$  consist of  $p + 1$  values.
- $V(\hat{\beta}_i) = E(\hat{\beta}_i - \beta_i)^2, i = 0, 1, \dots, p$ , the covariance  $\sigma_{i,j} := E(\hat{\beta}_i - \beta_i)(\hat{\beta}_j - \beta_j)^T$  can be defined for each pair  $i \neq j$ .
- matrix consisting of  $\sigma_{i,j}$  in the  $i$ th row and  $j$ th column as to the covariance matrix of  $\hat{\beta}$ .

$$E \begin{bmatrix} (\hat{\beta}_0 - \beta_0)^2 & (\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) & \cdots & (\hat{\beta}_0 - \beta_0)(\hat{\beta}_p - \beta_p) \\ (\hat{\beta}_1 - \beta_1)(\hat{\beta}_0 - \beta_0) & (\hat{\beta}_1 - \beta_1)^2 & \cdots & (\hat{\beta}_1 - \beta_1)(\hat{\beta}_p - \beta_p) \\ \vdots & \vdots & \ddots & \vdots \\ (\hat{\beta}_p - \beta_p)(\hat{\beta}_0 - \beta_0) & (\hat{\beta}_p - \beta_p)(\hat{\beta}_1 - \beta_1) & \cdots & (\hat{\beta}_p - \beta_p)^2 \end{bmatrix}$$



## Covariance matrix of $\hat{\beta}$

$$\begin{aligned} E & \begin{bmatrix} (\hat{\beta}_0 - \beta_0)^2 & (\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) & \cdots & (\hat{\beta}_0 - \beta_0)(\hat{\beta}_p - \beta_p) \\ (\hat{\beta}_1 - \beta_1)(\hat{\beta}_0 - \beta_0) & (\hat{\beta}_1 - \beta_1)^2 & \cdots & (\hat{\beta}_1 - \beta_1)(\hat{\beta}_p - \beta_p) \\ \vdots & \vdots & \ddots & \vdots \\ (\hat{\beta}_p - \beta_p)(\hat{\beta}_0 - \beta_0) & (\hat{\beta}_p - \beta_p)(\hat{\beta}_1 - \beta_1) & \cdots & (\hat{\beta}_p - \beta_p)^2 \end{bmatrix} \\ &= E \begin{bmatrix} \hat{\beta}_0 - \beta_0 \\ \hat{\beta}_1 - \beta_1 \\ \vdots \\ \hat{\beta}_p - \beta_p \end{bmatrix} [\hat{\beta}_0 - \beta_0, \hat{\beta}_1 - \beta_1, \dots, \hat{\beta}_p - \beta_p] \\ &= E(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T = E(X^T X)^{-1} X^T \epsilon (X^T X)^{-1} X^T \epsilon^T \\ &= (X^T X)^{-1} X^T E \epsilon \epsilon^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1} \end{aligned}$$

We have determined that the covariance matrix of  $\epsilon$  is  $E \epsilon \epsilon^T = \sigma^2 I$ .

$$\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$$