Chapter 2: Linear Regression

Gyeong min Kim

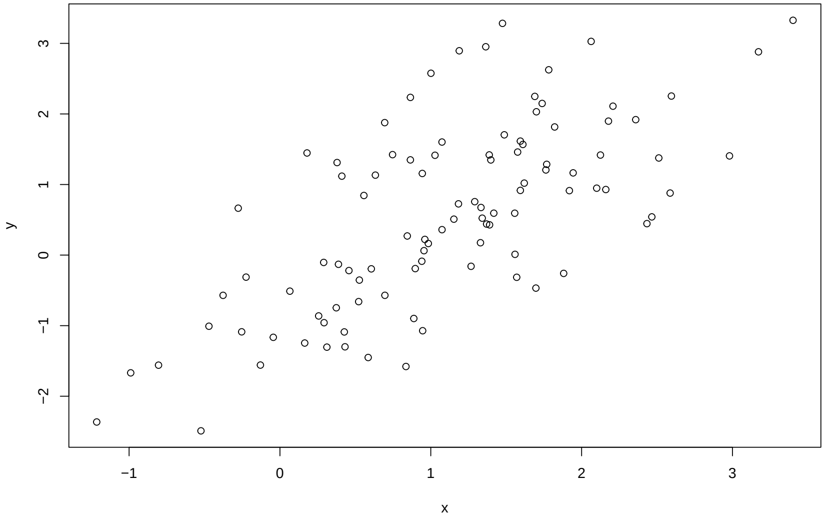
November 19, 2024

Department of Statistics
Sungshin Women's University

Outline

Data Generation

```
beta = c(-0.5, 1)
n = 100 ; x = rnorm(n, mean = 1) ; y = beta[1] + beta[2] * x + rnorm(n)
plot(x, y)
```



Least Squares algorithm for Simple Linear Regression

```r
ls = function(x, y){
  beta_hat1 = crossprod(x - mean(x), y - mean(y)) / crossprod(x - mean(x))
  beta_hat0 = mean(y) - beta_hat1 * mean(x)

  return(list("intercept" = as.numeric(beta_hat0),
              "slope" = as.numeric(beta_hat1)))
}

beta ; ls(x, y)
```

```
## [1] -0.5  1.0
```

```
## $intercept
## [1] -0.5366322
##
## $slope
## [1] 0.9989396
```

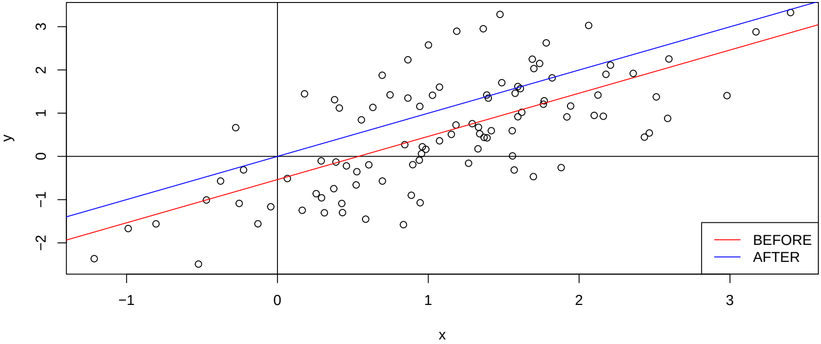
Plot of Simple Linear regression (Original vs. Centering)

```
c(ls(x, y)$intercept, ls(x - mean(x), y - mean(y))$intercept)
```

```
## [1] -5.366322e-01  4.660343e-17
```

```
c(ls(x, y)$slope, ls(x - mean(x), y - mean(y))$slope)
```

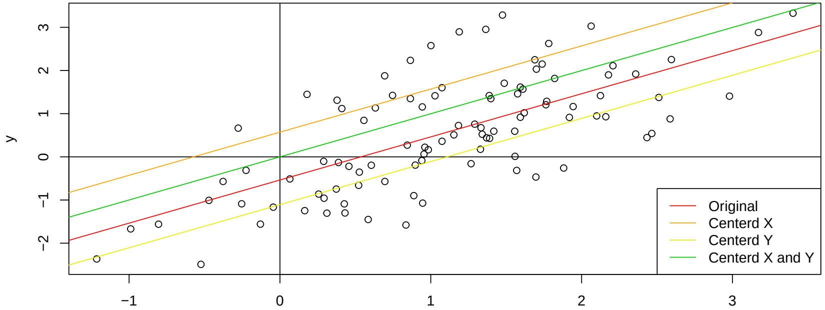
```
## [1] 0.9989396 0.9989396
```



Plot of Simple Linear regression (Original vs. Centering)*

```r
rbind("Original" = ls(x, y),
      "Centerd X" = ls(x - mean(x), y),
      "Centerd Y" = ls(x, y - mean(y)),
      "Centerd X and Y" = ls(x - mean(x), y - mean(y)))
```

```
##                    intercept    slope
## Original           -0.5366322   0.9989396
## Centerd X          0.5710793    0.9989396
## Centerd Y          -1.107712    0.9989396
## Centerd X and Y 4.660343e-17 0.9989396
```



Outline

Multiple Linear Regression scheme

- Consider the multiple linear regression:

$$y = X\beta + \varepsilon.$$

- $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \in \mathbf{R}^n \quad \text{where} \quad \varepsilon_1, \varepsilon_2, ..., \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2).$

- $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbf{R}^n, \ \ X = \begin{bmatrix} 1 & x_{11} & ... & x_{1p} \\ 1 & x_{21} & ... & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & ... & x_{np} \end{bmatrix} \in \mathbf{R}^{n\times(p+1)}, \ \ \text{and} \ \ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbf{R}^{p+1}.$

Solution of Multiple Linear Regression

- When a matrix $X^\top X \in \mathbf{R}^{(p+1)\times(p+1)}$ is invertible, we have

$$\hat{\beta} = \left(X^\top X\right)^{-1} X^\top y.$$

Data Generation and Least Squares Estimator

```r
beta = c(1, 2, 3)
n = 100 ; p = 3
X = cbind("intercept" = 1, "x1" = rnorm(n), "x2" = rnorm(n))
y = X %*% beta + rnorm(n)
solve(t(X) %*% X) %*% t(X) %*% y # Original
```

```
##                [,1]
## intercept 0.9586774
## x1        2.0309721
## x2        3.0245980
```

```r
C = cbind(1, X[,2] - mean(X[,2]), X[,3] - mean(X[,3]))
solve(t(C) %*% C) %*% t(C) %*% (y - mean(y)) # Centered X and Y
```

```
##              [,1]
## [1,] 1.526557e-16
## [2,] 2.030972e+00
## [3,] 3.024598e+00
```

Rank Condition of Design matrix X

- We may notice that the matrix $X^\top X$ is not invertible under each of the following conditions:
  1. $N < p + 1$
  2. Two columns in X coincide.

Outline

Moments of the estimator $\hat{\beta}$

$$\hat{\beta} = \left(X^\top X\right)^{-1} X^\top y$$

- The estimate $\hat{\beta}$ of β depends on the value of ε because N pairs of data $(x_1, y_1), ..., (x_n, y_n)$ randomly occur.

$$\mathbf{E}\left(\hat{\beta}\right) = \left(X^\top X\right)^{-1} X^\top \mathbf{E}\left(y\right) = \left(X^\top X\right)^{-1} X^\top X\beta = \beta$$
$$\mathrm{Var}\left(\hat{\beta}\right) = \left(X^\top X\right)^{-1} X^\top \mathrm{Var}(\varepsilon) X \left(X^\top X\right)^{-1} = \sigma^2 \left(X^\top X\right)^{-1}$$

$$\left(\therefore \hat{\beta} \sim N(\beta, \sigma^2 \left(X^\top X\right)^{-1})\right)$$

Outline

Hat matrix H

- We explore the properties of the matrix

$$H \triangleq X(X^\top X)^{-1}X^\top \in \mathbf{R}^{n\times n}.$$

- The following are easy to derive but useful in the later part of this book:

$$\begin{aligned} H^2 &= X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top = X(X^\top X)^{-1}X^\top = H \\ (I-H)^2 &= I - 2H + H^2 = I - 2H + H = I - H \\ HX &= X(X^\top X)^{-1}X^\top X = X. \end{aligned}$$

- Moreover, if we set $\hat{y} = X\hat{\beta}$, we have

$$\hat{y} = X\hat{\beta} = X(X^\top X)^{-1}X^\top y = Hy.$$

- And we observe

$$\begin{aligned} y - \hat{y} &= (I-H)y = (I-H)(X\beta + \varepsilon) \\ &= X\beta + \varepsilon - HX\beta - H\varepsilon = X\beta + \varepsilon - X\beta - H\varepsilon \\ &= (I-H)\varepsilon. \end{aligned}$$

The RSS with respect to Hat matrix H

- Observe the equation

$$RSS = \|y - \hat{y}\|^2 = \varepsilon^\top (I - H)^\top (I - H)\varepsilon = \varepsilon^\top (I - H)^2 \varepsilon = \varepsilon^\top (I - H)\varepsilon.$$

- To analysis RSS, we explore the properties of Hat matrix H.

Proposition 1

If $\mathrm{rank}(X) = p + 1$, *we obtain the diagonalization*

$$P(I - H)P^\top = \mathrm{diag}(\underbrace{1, ..., 1}_{N-p-1}, \underbrace{0, ..., 0}_{p+1}),$$

where P *is orthonormal matrix whose columns consist of eigenvectors of matrix* $I - H$.

Proof of Proposition 1

- If $\mathrm{rank}(X) = p + 1$, we have

$$\begin{aligned}\mathrm{rank}(H) &= \mathrm{rank}\left(X(X^\top X)^{-1} \cdot X^\top\right) \\ &\le \min\left\{\mathrm{rank}(X(X^\top X)^{-1}), \mathrm{rank}(X)\right\} \\ &\le \mathrm{rank}(X) = p + 1\end{aligned}$$

- If $\mathrm{rank}(X) = p + 1$, we have

$$\begin{aligned}\mathrm{rank}(H) &\ge \min\left\{\mathrm{rank}(H), \mathrm{rank}(X)\right\} \\ &\ge \mathrm{rank}(HX) = \mathrm{rank}(X) = p + 1\end{aligned}$$

- We conclude if $\mathrm{rank}(X) = p + 1$, $\mathrm{rank}(H) = p + 1$.

Proof of Proposition 1 (cntd.)

- Recall the relationship $HX = X$:

$$HX = H \begin{bmatrix} | & | & & | \\ X_1 & X_2 & ... & X_{p+1} \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ X_1 & X_2 & ... & X_{p+1} \\ | & | & & | \end{bmatrix}$$

- For i-th column, $i = 1, ..., p+1$, we have

$$HX_i = X_i.$$