Chapter 2: Linear Regression

Gyeong min Kim

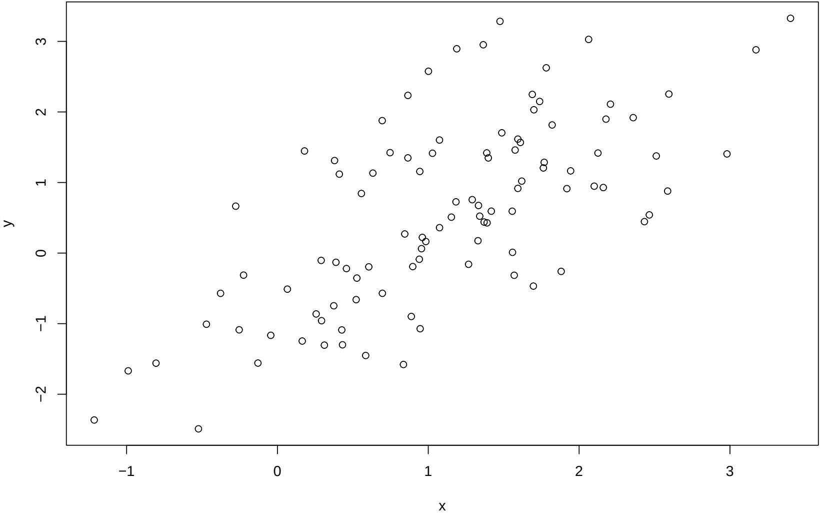
November 19, 2024

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

```
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
```

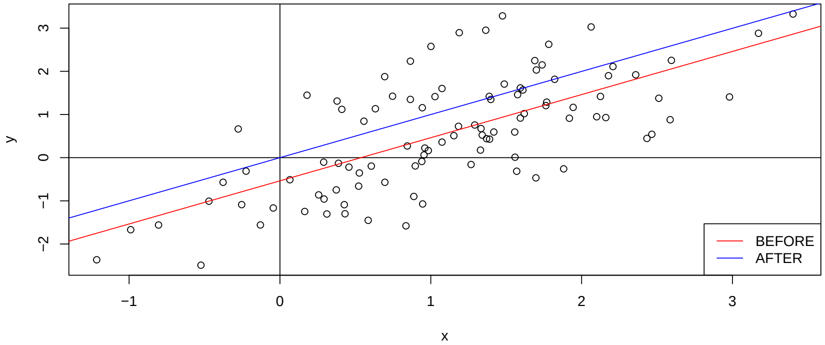
Plot of Simple Linear regression (Original vs. Centering)

```
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)
```

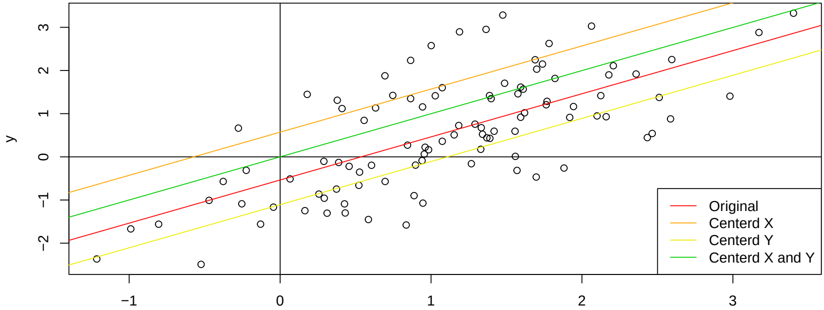
```
## [1] 0.9989396 0.9989396
```



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:

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

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

- $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbf{R}^n$, $X = \begin{bmatrix} 1 & x_{11} & ... & x_{1p} \\ 1 & x_{21} & ... & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & ... & x_{np} \end{bmatrix} \in \mathbf{R}^{n\times(p+1)}$, 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 $\operatorname{rank}(X) = p+1$, *we obtain the diagonalization*

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

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

Proof of Proposition 1

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

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

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

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

- We conclude that $\mathbf{rank(X) = p + 1} \quad \Rightarrow \quad \mathbf{rank(H) = p + 1}$.

Proof of Proposition 1 (cntd.)

- Recall the relationship $HX = X$:

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

- We have

$$HX_i = X_i \quad \text{for} \quad i = 1, \dots, p+1.$$

- $\text{rank}(X) = p + 1 \quad \Rightarrow \quad \dim(\mathcal{E}igen(H)) = p + 1$ where $\mathcal{E}igen(H)$ denotes the eigenspace with eigenvectors corresponding non-zero eigenvalue.
- Therefore, each column of X spans the eigenspace of H, which means X_is are the eigenvectors of the Hat matrix H.

Proof of Proposition 1 (cntd.)

- Now, we analyze the relationship between $\mathcal{E}igen(H)$ and $\mathcal{N}(I-H)$ where $\mathcal{N}$ denotes the nullspace of a matrix.
- For arbitray $x \in \mathbf{R}^n$,
$$Hx = x \quad \Rightarrow \quad (I-H)x = \mathbf{0},$$
which means that the eigenvectors of H belong to the nullspace of $I-H$.
- For arbitray $x \in \mathbf{R}^n$,
$$(I-H)x = \mathbf{0} \quad \Rightarrow \quad Hx = x,$$
which means that the vectors of $\mathcal{N}(I-H)$ belong to the $\mathcal{E}igen(H)$.
- Therefore, we have $\mathcal{N}(I-H) = \mathcal{E}igen(H)$, and
$$\dim \mathcal{N}(I-H) = \dim \mathcal{E}igen(H) = p+1.$$
- Then we observe $\text{rank}(I-H) = n-p-1$

Proof of Proposition 1 (cntd.)

- $I - H$ is diagonalizable matrix, since it is the symmetric and square matrix.
- Since $\mathrm{rank}(I - H) = n - p - 1$, we have

$$P^\top (I - H) P = \mathrm{diag}(\underbrace{\lambda_1, ..., \lambda_{n-p-1}}_{n-p-1}, \underbrace{0, ..., 0}_{p+1}),$$

where columns of P are orthonormal and consist of eigenvectors of $I - H$.

Proof of Proposition 1 (cntd.)

Lemma 1

Let a real matrix $D \in \mathbf{R}^{n\times n}$ such that $D^2 = D$. Then, the eigenvalues of D consist of only 0 and 1.

Proof.

Let $D \in \mathbf{R}^{n\times n}$ such that $D^2 = D$, and

$$\exists v \in \mathbf{R}^n, \;\; Dv = \lambda v.$$

Then, we observe

$$Dv = \lambda v \quad \Rightarrow \quad D^2 v = \lambda D v \quad \Rightarrow \quad Dv = \lambda^2 v,$$

since $D^2 = D$. Thus, we have

$$\lambda = \lambda^2 \quad \Rightarrow \quad \lambda = 0 \text{ or } 1.$$

□

- In order to proof Proposition 1, we apply the following:

  1. $\mathrm{rank}(X) = p + 1 \quad \Rightarrow \quad \mathrm{rank}(H) = p + 1$
  2. $\dim \mathcal{N}(I - H) = \dim \mathcal{E}igen(H) = p + 1 \quad \Rightarrow \quad \mathrm{rank}(I - H) = n - p - 1$
  3. $P^\top (I - H) P = \mathrm{diag}(\underbrace{\lambda_1, ..., \lambda_{n-p-1}}_{n-p-1}, \underbrace{0, ..., 0}_{p+1})$
  4. $(I - H)^2 = (I - H) \quad \Rightarrow \quad \lambda = 0 \text{ or } 1.$

- Therefore, if $\mathrm{rank}(X) = p + 1$, we obtain

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

Distribution of RSS values

- Since the columns of P are orthonormal,

$$\exists u \in \mathbf{R}^n, \;\; \varepsilon = Pu, \quad \text{and then} \quad u = P^\top \varepsilon.$$

- We have

$$\begin{aligned} RSS &= \varepsilon^\top (I - H)\varepsilon = u^\top P^\top (I - P) P u \\ &= u^\top \operatorname{diag}(\underbrace{1, ..., 1}_{n-p-1}, \underbrace{0, ..., 0}_{p+1}) u \\ &= \sum_{i=1}^{n-p-1} u_i^2. \end{aligned}$$

- We observe

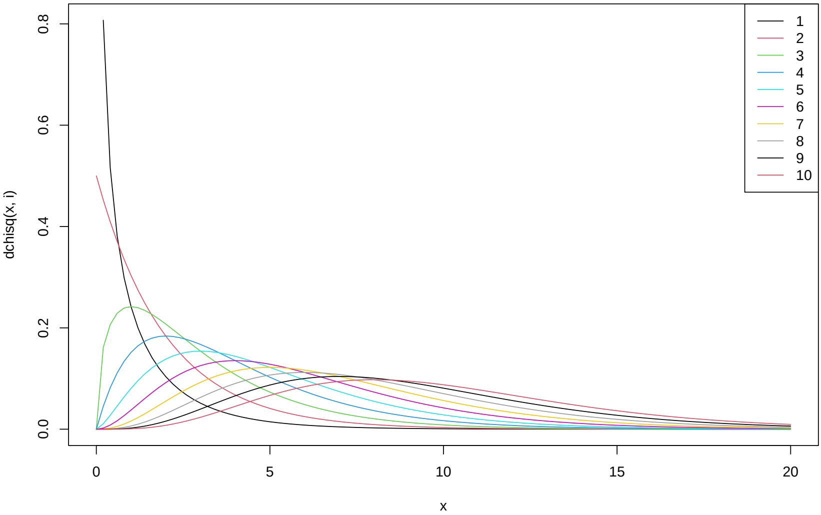
$$\begin{aligned} \mathbf{E}(u) &= \mathbf{E}(P\varepsilon) = \mathbf{0} \\ \operatorname{Cov}(u) &= \mathbf{E}(uu^\top) = P^\top E(\varepsilon\varepsilon^\top) P = \sigma^2 P^\top P = \sigma^2 I_n. \end{aligned}$$

- Since $u \sim N_n(0, \sigma^2 I_n)$ and then $\frac{u}{\sigma} \sim N_n(0, I_n)$,

$$\frac{RSS}{\sigma^2} = \sum_{i=1}^{n-p-1} \left(\frac{u_i}{\sigma}\right)^2 \sim \mathcal{X}^2_{n-p-1}.$$

Plot of Chi-squared distribution

```r
i = 1 ; curve(dchisq(x, i), 0, 20, col = i)
for(i in 2:10) curve(dchisq(x, i), 0, 20, col = i, add = TRUE, ann = FALSE)
legend("topright", legend = 1:10, lty = 1, col = 1:10)
```



Outline

Distribution of $\hat{\beta}$

- In this section, we consider whether each of the $\beta_j,\ j = 0, 1, ..., p$, is zero or not based on the data.
- Due to fluctuations in the N random variables $\varepsilon_1, ..., \varepsilon_n$, the data occurred by chance.
- Since the estimator $\hat{\beta} = (X^\top X)^{-1} X^\top y$ have randomness, we use the distribution of the estimator:

$$\hat{\beta} \sim N_{p+1}\left(\beta,\ \sigma^2 (X^\top X)^{-1}\right).$$

Known or Unknwon σ^2

- If we know σ^2, use the z-statistics

$$z = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{(X^\top X)^{-1}_{jj}}} \sim N(0,1).$$

- If we do not know σ^2, use the t-statistics

$$t = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}}} \sim t_{n-p-1}.$$

Unknwon σ^2

- When σ^2 is unknown, we need to estimate σ^2. Since $\frac{RSS}{\sigma^2} \sim \mathcal{X}^2_{n-p-1}$,

$$\mathbf{E}\left(\frac{RSS}{n-p-1}\right) = \sigma^2,$$

where $RSS/(n-p-1)$ is unbiased estimator of σ^2

$$\left(\therefore \hat{\sigma}^2 = \frac{RSS}{n-p-1}\right)$$

Unknwon σ^2

- Let $U \sim N(0,1)$ and $V \sim \mathcal{X}_{df}$, then

$$\frac{U}{\sqrt{V/df}} \sim t_{df}.$$

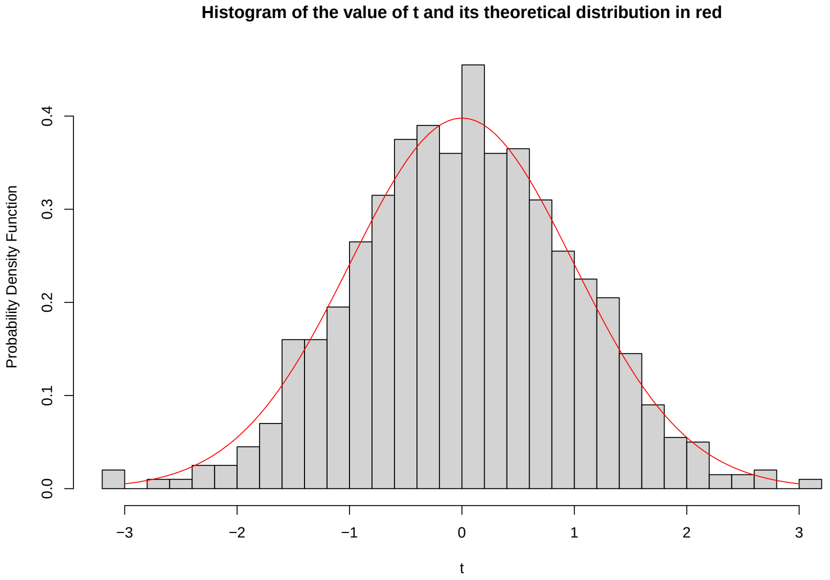
- t-statistics

$$\begin{aligned} t &= \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}}} \\ &= \frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{(X^\top X)^{-1}_{jj}}}\sqrt{\frac{\sigma^2}{\hat{\sigma}^2}} = \frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{(X^\top X)^{-1}_{jj}}}/\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}} \\ &= \frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{(X^\top X)^{-1}_{jj}}}/\sqrt{\frac{RSS}{\sigma^2}/(n-p-1)} \\ &= \frac{U}{\sqrt{V/df}} \sim t_{df} = t_{n-p-1} \end{aligned}$$

[Example 1] Under $H_0 : \beta_j = 0$

```r
n = 100 ; p = 1 ; rep = 1000
T = NULL
t = rep(0, rep)
for(i in 1:rep){
  x = rnorm(n) ; y = rnorm(n)
  fit = lm(y ~ x)
  RSS = crossprod(y - fit$fitted.values)
  sigma_hat = sqrt( RSS / (n - p - 1) )
  statistics = fit$coefficients[2] / ( sigma_hat / sqrt(crossprod(x - mean(x))) )
  t[i] = statistics
}
```

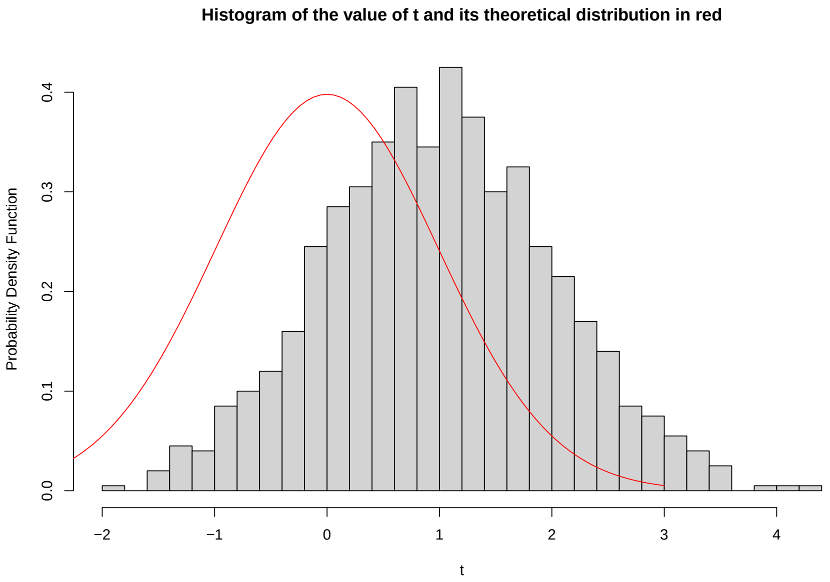
Plot of Example 1



[Example 2] Not under $H_0 : \beta_j = 0$

```r
n = 100 ; p = 1 ; rep = 1000
T = NULL
t = rep(0, rep)
for(i in 1:rep){
  x = rnorm(n) ; y = 0.1*x + rnorm(n)
  fit = lm(y ~ x)
  RSS = crossprod(y - fit$fitted.values)
  sigma_hat = sqrt( RSS / (n - p - 1) )
  statistics = fit$coefficients[2] / ( sigma_hat / sqrt(crossprod(x - mean(x))) )
  t[i] = statistics
}
```

Plot of Example 2



Outline

Three sum of squares

- Let $W \in \mathbf{R}^{n\times n}$ be a matrix such that all the elements are $1/n$. Then we have $Wy = (\bar{y}, ..., \bar{y}) \in \mathbf{R}^n$
- Total sum of squares (TSS):

$$TSS \triangleq \|y - \bar{y} \cdot \mathbf{1}\|_2^2 = \|y - Wy\|_2^2 = \|(I - W)y\|_2^2$$

- Residual sum of squares (RSS):

$$RSS \triangleq \|y - \hat{y}\|_2^2 = \|y - Hy\|_2^2 = \|(I - H)y\|_2^2$$

- Explained sum of squares (ESS):

$$ESS \triangleq \|\hat{y} - \bar{y} \cdot \mathbf{1}\|_2^2 = \|Hy - Wy\|_2^2 = \|(H - W)y\|_2^2$$