

## Chapter 2 : Linear Regression

---

Newton's three sisters

February 16, 2025

Department of Statistics

Sungshin Women's University

- 1 Least Squares Method
- 2 Multiple Regression
- 3 Distribution of  $\hat{\beta}$
- 4 Distribution of the RSS Values
- 5 Hypothesis Testing for  $\hat{\beta}_j \neq 0$
- 6 Coefficient of Determination and the Detection of Collinearity
- 7 Confidence and Prediction Intervals

- 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.

- 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  $\sum_{i=1}^N (x_i - \bar{x})^2 \neq 0$ ,

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$\hat{\beta}_0, \hat{\beta}_1$  instead of  $\beta_0, \beta_1$  which means that they are not the true values but rather estimates obtained from data.

- 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:

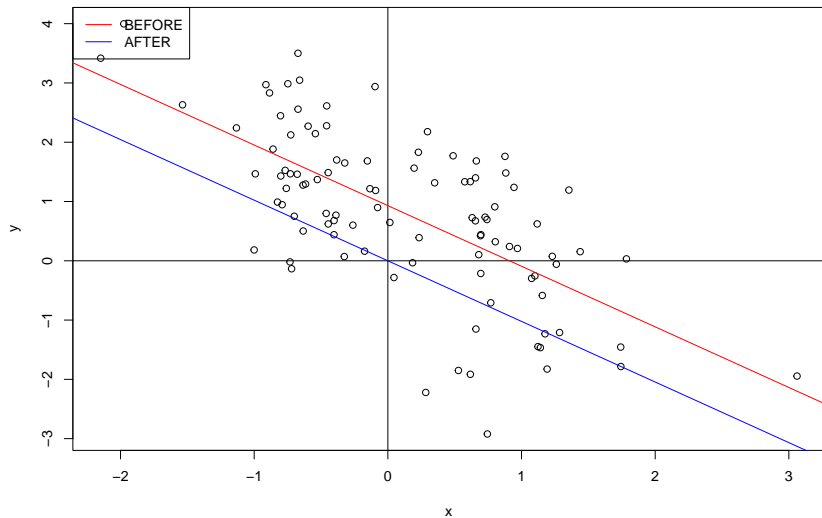
$$\hat{\beta}_1 = \frac{\sum_{i=1}^N \tilde{x}_i \tilde{y}_i}{\sum_{i=1}^N (\tilde{x}_i)^2}$$

## 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  $\hat{\beta}$
- 4 Distribution of the RSS Values
- 5 Hypothesis Testing for  $\hat{\beta}_j \neq 0$
- 6 Coefficient of Determination and the Detection of Collinearity
- 7 Confidence and Prediction Intervals



# 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  $\hat{\beta}$**
- 4 Distribution of the RSS Values
- 5 Hypothesis Testing for  $\hat{\beta}_j \neq 0$
- 6 Coefficient of Determination and the Detection of Collinearity
- 7 Confidence and Prediction Intervals

- $y$  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  $E \epsilon \epsilon^T = \sigma^2 I$ .

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

- 1 Least Squares Method
- 2 Multiple Regression
- 3 Distribution of  $\hat{\beta}$
- 4 Distribution of the RSS Values**
- 5 Hypothesis Testing for  $\hat{\beta}_j \neq 0$
- 6 Coefficient of Determination and the Detection of Collinearity
- 7 Confidence and Prediction Intervals

- Hat matrix defined by  $\hat{y} = Hy$

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

$$H \triangleq X(X^T X)^{-1} X^T$$

- Some properties

$$H^2 = X(X^T X)^{-1} X^T \cdot X(X^T X)^{-1} X^T = X(X^T X)^{-1} X^T = H$$

$$(I - H)^2 = I - 2H + H^2 = I - H$$

$$HX = X(X^T X)^{-1} X^T \cdot X = X$$

- RSS defined

$$\text{RSS} \triangleq \|y - \hat{y}\|^2$$

- Using hat matrix

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

$$\text{RSS} \triangleq \|y - \hat{y}\|^2 = \{(I - H)\varepsilon\}^T (I - H)\varepsilon = \varepsilon^T (I - H)^2 \varepsilon = \varepsilon^T (I - H)\varepsilon$$

- They are only zeros and ones
- Dimensions of the eigenspaces of  $H$  and  $I - H$  are both  $p + 1$

**Proof** using  $\text{rank}(X) = p + 1$

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

$$\text{rank}(H) \geq \text{rank}(HX) = \text{rank}(X) = p + 1$$

- Hat matrix defined by

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

$$H \triangleq X(X^T X)^{-1} X^T$$

- some properties

$$H^2 = X(X^T X)^{-1} X^T \cdot X(X^T X)^{-1} X^T = X(X^T X)^{-1} X^T = H$$

$$(I - H)^2 = I - 2H + H^2 = I - H$$

$$HX = X(X^T X)^{-1} X^T \cdot X = X$$

- RSS defined

$$RSS \triangleq ||y - \hat{y}||^2$$

- Using hat matrix

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

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

## Eigenvalues of $H$ and Null space of $(I - H)$

- Proof by contrapositive

$$Hx = x \Rightarrow (I - H)x = 0$$

$$(I - H)x = 0 \Rightarrow Hx = x$$

- Dimensions of the eigenspaces of  $H$  is  $p + 1$

**Proof** using  $\text{rank}(X) = p + 1$

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

$$\text{rank}(H) \geq \text{rank}(HX) = \text{rank}(X) = p + 1$$

- Dimensions of the null space of  $I - H$  is  $N - (p + 1)$

$$P(I - H)P^T = \text{diag}(\underbrace{1, \dots, 1}_{N-p-1}, \underbrace{0, \dots, 0}_{p+1})$$



- We define  $v = P\varepsilon \in \mathbb{R}^N$ , then from  $\varepsilon = P^T v$

$$\text{RSS} = \varepsilon^T (I - H) \varepsilon = (P^T v)^T (I - H) P^T v = v^T P (I - H) P^T v$$

$$= [v_1, \dots, v_{N-p-1}, v_{N-p}, \dots, v_N] \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & \ddots & 0 & \dots & \dots & \vdots \\ \vdots & 0 & 1 & \dots & \dots & 0 \\ 0 & 0 & 0 & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & \dots & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_{N-p-1} \\ v_{N-p} \\ \vdots \\ v_N \end{bmatrix} = \sum_{i=1}^{N-p-1} v_i^2$$

- Let  $w \in \mathbb{R}^{N-p-1}$

- Average

$$E[v] = E[P\varepsilon] = 0$$

$$E[w] = 0$$

- Covariance

$$E[vv^t] = E[P\varepsilon(P\varepsilon)^T] = PE[\varepsilon\varepsilon^t]P = P\sigma^2IP^T = \sigma^2I$$

$$E[ww^T] = \sigma^2I$$

- We have RSS

$$\frac{RSS}{\sigma^2} \sim \chi_{N-p-1}^2$$

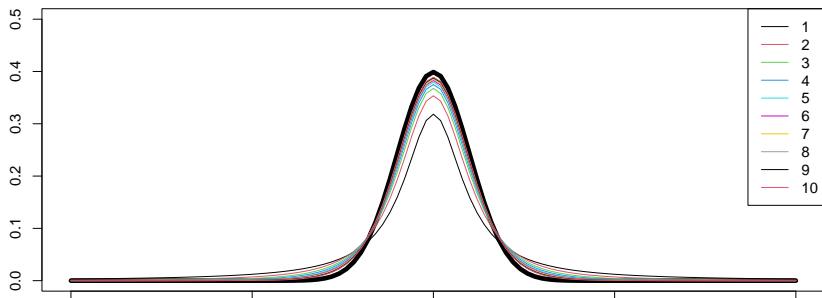
- 1 Least Squares Method
- 2 Multiple Regression
- 3 Distribution of  $\hat{\beta}$
- 4 Distribution of the RSS Values
- 5 Hypothesis Testing for  $\hat{\beta}_j \neq 0$
- 6 Coefficient of Determination and the Detection of Collinearity
- 7 Confidence and Prediction Intervals

## Test statistic $T$

- A  $t$  distribution with  $N - P - 1$  degrees of freedom
- We decide that hypothesis  $\beta_j = 0$  should be rejected.
- $U \sim N(0, 1)$ ,  $V \sim \chi_m^2$ ,

$$T \triangleq U / \sqrt{V/m}$$

```
curve(dnorm(x), -10, 10, ann = FALSE, ylim = c(0, 0.5), lwd = 5)
for(i in 1:10)curve(dt(x, df= i), -10, 10, col = i, add = TRUE, ann = FALSE)
legend("topright", legend = 1:10, lty = 1, col = 1:10)
```



- $\alpha = 0.01, 0.05$
- Reject the null hypothesis

- 1 Least Squares Method
- 2 Multiple Regression
- 3 Distribution of  $\hat{\beta}$
- 4 Distribution of the RSS Values
- 5 Hypothesis Testing for  $\hat{\beta}_j \neq 0$
- 6 Coefficient of Determination and the Detection of Collinearity
- 7 Confidence and Prediction Intervals

- We define a matrix  $W \in \mathbb{R}^{N \times N}$  such that all the elements are  $1/N$   
 $Wy \in \mathbb{R}^N$  are  $\bar{y} = Wy = \sum_{i=1}^N y_i$  for  $y_1, \dots, y_N \in \mathbb{R}$
- Residual sum of squares RSS

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

- Explained sum of squares ESS

$$\text{ESS} \triangleq \|\hat{y} - \bar{y}\|^2 = \|\hat{y} - Wy\|^2 = \|(H - W)y\|^2$$

- Total sum of squares TSS

$$\text{TSS} \triangleq \|y - \bar{y}\|^2 = \|(I - W)y\|^2$$

- We have relation  $TSS = RSS + ESS$

**Proof**



- Coefficient of determination

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

- Correlation between the covariates and response

$$\hat{\rho} \triangleq \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2 \sum_{i=1}^N (y_i - \bar{y})^2}}$$

$$\begin{aligned} \frac{\text{ESS}}{\text{TSS}} &= \frac{\hat{\beta}_1^2 \|x - \bar{x}\|^2}{\|y - \bar{y}\|^2} = \left\{ \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2} \right\}^2 \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \\ &= \frac{\left\{ \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) \right\}^2}{\sum_{i=1}^N (x_i - \bar{x})^2 \sum_{i=1}^N (y_i - \bar{y})^2} = \hat{\rho}^2 \end{aligned}$$

- Variance inflation factors

$$\text{VIF} \triangleq \frac{1}{1 - R_{X_j|X_{-j}}^2}$$

- The minimum value of VTI is one, and we say that the collinearity of covariate is strong when its VIF value is large

- 1 Least Squares Method
- 2 Multiple Regression
- 3 Distribution of  $\hat{\beta}$
- 4 Distribution of the RSS Values
- 5 Hypothesis Testing for  $\hat{\beta}_j \neq 0$
- 6 Coefficient of Determination and the Detection of Collinearity
- 7 Confidence and Prediction Intervals

- We have showed how to obtain the estimate  $\hat{\beta}$  of  $\beta \in \mathbb{R}^{p+1}$ , confidence interval of  $\hat{\beta}$  as follows

$$\beta_i = \hat{\beta}_i \pm t_{N-p-1}(\alpha/2)\text{SE}(\hat{\beta}_i), \quad \text{for } i = 0, 1, \dots, p$$

- Confidence interval of  $x_*\hat{\beta}$  for another point  $x_* \in \mathbb{R}^{p+1}$

- The average

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

- The variance

$$V[x_*\hat{\beta}] = x_*V(\hat{\beta})x_*^T = \sigma^2x_*(X^TX)^{-1}x_*^T$$

- We define

$$\hat{\sigma} \triangleq \sqrt{\text{RSS}/(N-p-1)}, \quad \text{SE}(x_*\hat{\beta}) \triangleq \hat{\sigma}\sqrt{x_*(X^TX)^{-1}x_*^T}$$

- $C \sim t_{N-p-1}$
- variance in the difference between  $x_*\hat{\beta}$  and  $y_* \triangleq x_*\beta + \varepsilon$

$$V[x_*\hat{\beta} - (x_*\beta + \varepsilon)] = V[x_*(\hat{\beta} - \beta)] + V[\varepsilon] = \sigma^2 x_*(X^T X)^{-1} x_*^T + \sigma^2$$

- Similarly, we can derive the following

$$P \triangleq \frac{x_*\hat{\beta} - y_*}{\text{SE}(x_*\hat{\beta} - y_*)} = \frac{x_*\hat{\beta} - y_*}{\sigma(1 + \sqrt{x_*(X^T X)^{-1} x_*^T})} / \sqrt{\frac{\text{RSS}}{\sigma^2} / (N - p - 1)} \sim t_{N-p-1}$$

- The confidence and prediction intervals

$$\begin{aligned} x_*\beta &= x_*\hat{\beta} \pm t_{N-p-1}(\alpha/2)\hat{\sigma}\sqrt{x_*(X^T X)^{-1} x_*^T} \\ y_* &= x_*\hat{\beta} \pm t_{N-p-1}(\alpha/2)\hat{\sigma}\sqrt{1 + x_*(X^T X)^{-1} x_*^T} \end{aligned}$$

# Q & A

**Thank you :)**