Chapter 2 : Linear Regression

Newton's three sisters

February 16, 2025

Department of Statistics
Sungshin Women's University

Outline

Simple Linear Regression

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

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

  - β_0: intercept
  - β_1: slope
  - ε_i: random error
- We obtain β_0 and β_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 $i = 1, 2, ..., N$.

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

- Then, by partially differentiating L by β_0, β_1 and letting them be zero.

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

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

- β_0 and β_1 are regarded as constants when differentiating L by β_1 and β_0.

Least Squares Method

- When $\sum_{i=1}^{N}(x_i - \bar{x})^2 \neq 0$,
  $\hat{\beta}_0$, $\hat{\beta}_1$ instead of β_0, β_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,

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

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

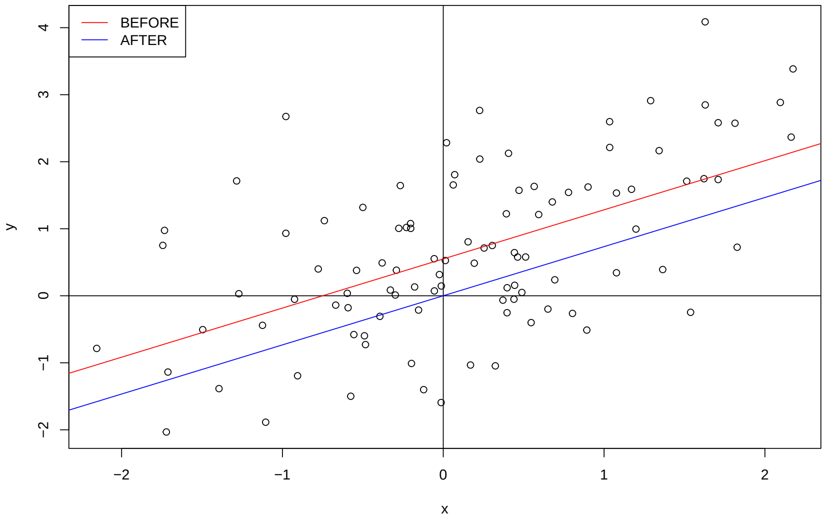
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} \tilde{x}_i \tilde{y}_i}{\sum_{i=1}^{N} (\tilde{x}_i)^2}$$

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



Outline

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

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

Multiple Regression

- 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^TX is invertible, we have

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

When $X^T X$ is irreversible

1. $N < p + 1$

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

   If $N > p$, It is X_particular, So there is no inverse matrix.

2. Two columns in X coincide.

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

Outline

Distribution of $\hat{\beta}$ 제목 생각해보기

- y have been obtained from the covariates X multiplied by the (true) coefficients β plus some noise ϵ.

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

- The true β 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),$\cdots$,$(x_N, y_N) \in R^p$ X R
- $x_i \in R^p$ is the row vector consisting of p values excluding the leftmost one in the ith row of X.

Density function

- We assume that each element $\epsilon_1, \cdots, \epsilon_N$ in the random variable ϵ is independent of the others and Gaussian distribution with mean zero and variance σ^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 $\epsilon_1, \cdots, \epsilon_N$ 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 ϵ 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 β consist of $p+1$ values.
- $V(\hat{\beta}_i) = E(\hat{\beta}_i - \beta_i)^2, i = 0, 1, \cdots, p$, 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 ith row and jth 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 $\hat{\beta}$

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

$$
= E\begin{bmatrix}\hat{\beta}_0-\beta_0\\ \hat{\beta}_1-\beta_1\\ \vdots\\ \hat{\beta}_p-\beta_p\end{bmatrix}\left[\hat{\beta}_0-\beta_0, \hat{\beta}_1-\beta_1, \cdots, \hat{\beta}_p-\beta_p\right]
$$

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

$$
= (X^TX)^{-1}X^TE\epsilon\epsilon^TX(X^TX)^{-1} = \sigma^2(X^TX)^{-1}
$$

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

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

Outline

Hat matrix

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

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

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

- Some properties

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

Residual Sum of Square

- RSS defined

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

- Using hat matrix

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

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

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 $N - (p + 1)$

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

Residual Sum of Squares (RSS) and Eigenvalue

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

$$
\begin{aligned}
\mathrm{RSS} &= \varepsilon^T (I - H)\varepsilon = (P^T v)^T (I - H) P^T v = v^T P (I - H) P^T v \\
&= [v_1, \cdots, v_{N-p-1}, v_{N-p}, \cdots, v_n] \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & \cdots & \vdots \\ \vdots & 0 & 1 & \cdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \cdots & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_{N-p-1} \\ v_{N-p} \\ \vdots \\ v_N \end{bmatrix} \\
&= \sum_{i=1}^{N-p-1} v_i^2
\end{aligned}
$$

Statistical Properties of the Residual Sum of Squares (RSS)

- Let $w \in \mathbb{R}^{N-p-1}$ be
  - Average
    $E[v] = E[P\varepsilon] = 0$
    $E[w] = 0$
  - Covariance
    $E[vv^t] = E[P\varepsilon(P\varepsilon)^T] = PE[\varepsilon\varepsilon^t]P = P\sigma^2 IP^T = \sigma^2 I$
    $E[ww^T] = \sigma^2 I$
- We have RSS

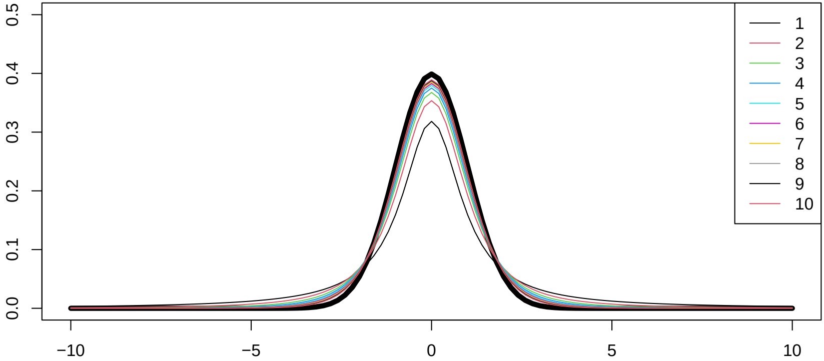
$$\frac{RSS}{\sigma^2} \sim \chi^2_{N-p-1}$$

Outline

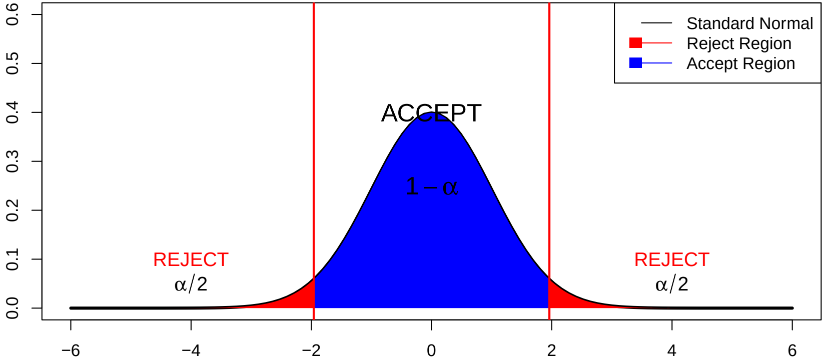
Test statistic T

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

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



- $\alpha = 0.01,\ 0.05$
- Null hypothesis $\beta_j = 0$



Example 23

- For $p = 1$, since

$$X^T X = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_N \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = N \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{N}\sum_{i=1}^N x_i^2 \end{bmatrix}$$

- The inverse is

$$(X^T X)^{-1} = \frac{1}{N} \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{N}\sum_{i=1}^N x_i^2 \end{bmatrix}^{-1} = \frac{1}{\sum_{i=1}^N (x_i - \bar{x})^2} \begin{bmatrix} \frac{1}{N}\sum_{i=1}^N x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

- Which means that

$$B_0 = \frac{\frac{1}{N}\sum_{i=1}^N x_i^2}{\sum_{i=1}^N (x_i - \bar{x})^2} \quad \text{and} \quad B_1 = \frac{1}{\sum_{i=1}^N (x_i - \bar{x})^2}$$

Example 23 (contd.)

- For $B = (X^T X)^{-1}, B\sigma^2$ is covariance matrix of $\hat{\beta}$
- $B_j\sigma^2$ is the variance of $\hat{\beta}_j$
- Because $\bar{x}$ is positive, the correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$ is negative

$$t = \frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}(\hat{\beta}_j)} \sim t_{N-p-1}$$

Statistical Independence in Regression

- It remains to be shown that U and V are independent

$$U \triangleq \frac{\hat{\beta}_j - \beta_j}{\sqrt{B_j}\sigma} \sim N(0,1) \quad \text{and} \quad V \triangleq \chi^2_{N-p-1}$$

- Sufficient to show tha $y - \hat{y}$ and $\hat{\beta} - \beta$ are independent

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

- From $E\varepsilon\varepsilon^T = \sigma^2 I$ and $HX = X$,

$$E(\hat{\beta} - \beta)(y - \hat{y})^T = 0$$

RSS, ESS, and TSS

- 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 $y_1, \cdots, y_N \in \mathbb{R}$
- Residual sum of squares RSS

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

- Explained sum of squres ESS

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

- Total sum of squres TSS

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

Sample - based correlation

- 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

- Variance inflation factors

$$\mathrm{VIF} \triangleq \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

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

Outline

Confidence Interval

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

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

- 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

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

- We define

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

Confidence and Prediction Intervals in Regression

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

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

- Similarly, we can derive the following

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

- The confidence and prediction intervals

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

Q & A

Thank you :)