

$$\begin{aligned}
 \text{RSS}(\beta) &= (Y - X\beta)^T (Y - X\beta) \\
 &= Y^T Y - Y^T X \beta - \underbrace{\beta^T X^T Y}_{= Y^T X \beta} + \beta^T X^T X \beta \\
 &= Y^T Y - 2 Y^T X \beta + \beta^T X^T X \beta
 \end{aligned}$$

Recall that $\frac{\partial a^T b}{\partial b} = a$ and $\frac{\partial b^T A b}{\partial b} = (A + A^T) b$

We have:

$$\frac{\partial \overbrace{Y^T X \beta}^a}{\partial \beta} = X^T Y$$

$$\frac{\partial \beta^T \overbrace{X^T X}^A \beta}{\partial \beta} = \left(X^T X + \underbrace{(X^T X)^T}_{= X^T X} \right) \beta = 2 X^T X \beta$$

$$\rightarrow \frac{\partial \text{RSS}(\beta)}{\partial \beta} = -2 X^T Y + 2 X^T X \beta$$

To minimise $\text{RSS}(\beta)$ wrt β we solve the following for β :

$$-2 X^T Y + 2 X^T X \beta = 0$$

$$\rightarrow X^T X \beta = X^T Y$$

$$\rightarrow \beta = (X^T X)^{-1} X^T Y, \text{ as long as } X^T X \text{ is invertible}$$

When $\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$, then the likelihood of $Y|X=x$ is given by

$$\begin{aligned} L(\beta, \sigma^2) &= p(y_1, \dots, y_N | x_1, \dots, x_N) \\ &= \prod_{i=1}^N p(y_i | x_i) \\ &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2\sigma^2} (y_i - x_i^T \beta)^2} \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^N (y_i - x_i^T \beta)^2}_{= \text{RSS}(\beta)}} \end{aligned}$$

The log-likelihood is then:

$$\ell(\beta, \sigma^2) = \log L(\beta, \sigma^2) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \text{RSS}(\beta)$$

The β that maximises $\ell(\beta, \sigma^2)$ is the one that minimises $\text{RSS}(\beta)$.

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

We have:

$$\begin{aligned} E(\hat{\beta} | X) &= E\left((X^T X)^{-1} X^T Y | X\right) \\ &= (X^T X)^{-1} X^T \underbrace{E(Y | X)}_{X\beta} = \underbrace{(X^T X)^{-1} X^T X}_{=I} \beta = \beta \end{aligned}$$

$$\begin{aligned} \text{Var}(\hat{\beta} | X) &= \text{Var}\left((X^T X)^{-1} X^T Y | X\right) \\ &= (X^T X)^{-1} X^T \underbrace{\text{Var}(Y | X)}_{\sigma^2 I} ((X^T X)^{-1} X^T)^T \\ &= (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} \underbrace{X^T X (X^T X)^{-1}}_{=I} = \sigma^2 (X^T X)^{-1} \end{aligned}$$

Recall that if $Z \sim N(\mu, \sigma^2)$, so that

$$\frac{Z - \mu}{\sigma} \sim N(0, 1) \quad \text{and}$$

$U \sim \chi^2_\nu$, then

$$\frac{(Z - \mu)/\sigma}{\sqrt{U/\nu}} \sim t_\nu$$