

We have:

$$\begin{aligned}
 \text{RSS}(\beta) &= (\underline{Y} - \underline{X}\beta)^T (\underline{Y} - \underline{X}\beta) = \sum_{i=1}^N (Y_i - (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p}))^2 \\
 &= \underline{Y}^T \underline{Y} - \underline{Y}^T \underline{X} \beta - \underbrace{\beta^T \underline{X}^T \underline{Y}}_{= (\beta^T \underline{X}^T \underline{Y})^T} + \beta^T \underline{X}^T \underline{X} \beta \\
 &= \underline{Y}^T \underline{Y} - 2 \underline{Y}^T \underline{X} \beta + \beta^T \underline{X}^T \underline{X} \beta
 \end{aligned}$$

We want to minimise $\text{RSS}(\beta)$ w.r.t. to β , so we differentiate RSS w.r.t. β and set the derivatives to 0.

Recall that if $\underline{a}$ and $\underline{b}$ are vectors and $\underline{A}$ is a matrix with appropriate dimensions, then

$$\frac{\partial \underline{a}^T \underline{b}}{\partial \underline{b}} = \underline{a} \quad \text{and} \quad \frac{\partial \underline{b}^T \underline{A} \underline{b}}{\partial \underline{b}} = (\underline{A} + \underline{A}^T) \underline{b}$$

We get

$$\frac{\partial \underline{Y}^T \underline{X} \beta}{\partial \beta} = (\underline{Y}^T \underline{X})^T = \underline{X}^T \underline{Y}$$

$$\frac{\partial \beta^T \underline{X}^T \underline{X} \beta}{\partial \beta} = (\underline{X}^T \underline{X} + \underbrace{(\underline{X}^T \underline{X})^T}_{\underline{X}^T \underline{X}}) \beta = 2 \underline{X}^T \underline{X} \beta$$

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

We need to solve

$$-2 \underline{X}^T \underline{Y} + 2 \underline{X}^T \underline{X} \hat{\beta} = 0$$

$$\rightarrow \underline{X}^T \underline{X} \hat{\beta} = \underline{X}^T \underline{Y}$$

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

When $\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$, the likelihood of y is given by

$$\begin{aligned} L(\beta, \sigma^2; y, X) &= p(y|X) \stackrel{ind}{=} \prod_{i=1}^N p(y_i | x_i) \\ &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2\sigma^2} (y_i - x_i^T \beta)^2} \\ &= \frac{1}{(2\pi \sigma^2)^{N/2}} \cdot e^{-\frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^N (y_i - x_i^T \beta)^2}_{RSS(\beta)}} \end{aligned}$$

The log-likelihood is the:

$$l(\beta, \sigma^2; y, X) = \log L(\beta, \sigma^2; y, X) = -\frac{N}{2} \log(2\pi \sigma^2) - \frac{1}{2\sigma^2} RSS(\beta)$$

We see that the β that maximises the log-likelihood is the same as the one that minimises $RSS(\beta)$, i.e. the least squares estimate of β .

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

We have:

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

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