3) Given $y = \beta_0 + \beta_1 x + \epsilon$

where ϵ is noise and $N(\epsilon : 0, \sigma_e^2)$ i.e. error (noise) follows a normal distribution.

$$\hat{y}_i = \beta_0 + \beta_1 x_i \Rightarrow y_i - \hat{y}_i = \epsilon$$

Function of ϵ, $f(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{1}{2\pi\sigma^2}(\epsilon_i)^2\right)}$

The likelihood is given by,

$$L = \frac{1}{(2\pi\sigma^2)^{-N/2}} e^{-\left(\frac{1}{2\sigma^2}\sum_1^{N}\epsilon^2\right)}$$

Log likelihood is, $l = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\epsilon_i^2$

$$l = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\epsilon_i^2$$

where N is number of datapoints

So, for the likelihood to be maximum, we choose,

$$\beta_{1_{MLE}} = \arg\max_{\beta_1} - \sum_{i=1}^{N}\epsilon_i^2$$

In other words, $\beta_{1_{MLE}} = \arg\min_{\beta_1} \frac{1}{N}\sum_{i=1}^{N}\epsilon_i^2$

or $\beta_{1_{MLE}} = \arg\min_{\beta_1} \underline{\underline{MSE}}$

which essentially means that if we minimize MSE, we get maximum likelihood. Conversely, it is clear that maximization of likelihood leads to minimized MSE.