Math behind linear regression

we have been told in class that the cost function is always

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \underbrace{(\theta_i^T x_i + b)}_{})^2$$

where $y_{pred} = \theta_i^T x_i + b$

But have ever wondered they might have come to this conclusion that we have to minimize the sum of squared error

The concept used behind this is known as maximising the log likelihood.

So we know that there is a noise associated to predicted value since we assume each y_i value is generated by a linear model plus some random noise:

$$y_i = \theta x_i + b + \epsilon_i$$

gaussian noise follows normal distribution acc. to central limit theorem

Where $\epsilon_i \sim N(0, \sigma^2)$ i.e; normal distribution with mean 0 and variance σ^2

ie; $y_i | x_i \cong N(\underbrace{w x_i + b}_{\hat{y_i}}, \sigma^2)$

mean of distribution is the prediction

The probability density function for a normal distribution is

$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

For our case $\Rightarrow$ $$p(y_i | x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(y_i - (\theta x_i + b))^2}{2\sigma^2}\right)$$

maximum likelihood $\Rightarrow$ $$\prod_{i=1}^{n} p(y_i | x_i)$$

$$\prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(y_i - (\theta x_i + b))^2}{2\sigma^2}\right)\right]$$

$$\prod_{i=1}^{n} a = a^n$$ basic property

$$\prod_{i=1}^{n} \exp(x_i) = \exp\left(\sum x_i\right)$$

$$\mathcal{L}(\theta, b) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \cdot \exp\left(-\frac{\sum (y_i - (\theta x_i + b))^2}{2\sigma^2}\right)$$

Taking log on both sides

$$\log(a.b) = \log a + \log b$$

$$\log L(\theta, b) = n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - (\theta x_i + b)\right)^2$$

Let us assume variance $\sigma^2 = 1$ for simplicity

constant

$$\log L(\theta, b) = n \log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(y_i - (\theta x_i + b)\right)^2$$

Hence we have to optimize the cost function which maximizing log likelihood is equivalent to minimizing the sum of squared errors.

$$\log L(\theta, b) = \min_{(\theta, b)} -\frac{1}{2}\sum_{i=1}^{n}\left(y_i - (\theta x_i + b)\right)^2$$

We are always thought this way to remember the loss function but if someone is interested like me why and how it came? This is the derivation behind it.