Mathematical explanation for Linear Regression working

Suppose we are given a dataset:

Experience (X)	Salary (y) (in lakhs)
2	3
6	10
5	4
7	13

Given is a Work vs Experience dataset of a company and the task is to predict the salary of a employee based on his / her work experience.
This article aims to explain how in reality Linear regression mathematically works when we use a pre-defined function to perform prediction task.
Let us explore **how the stuff works when Linear Regression algorithm gets trained.**

Iteration 1 – In the start, θ0 and θ1 values are randomly chosen. Let us suppose, θ0 = 0 and θ1 = 0.

- **Predicted values after iteration 1 with Linear regression hypothesis.**

$$h_\theta = \begin{bmatrix} \theta_0 & \theta_1 \end{bmatrix} \begin{bmatrix} x_0 & x_0 & x_0 & x_0 \\ x_1 & x_2 & x_3 & x_4 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 6 & 5 & 7 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}$$

- **Cost Function – Error**

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}[h_\theta(x_i) - y_i]^2$$
$$= \frac{1}{2\times 4}\left[(0-3)^2 + (0-10)^2 + (0-4)^2 + (0-3)^2\right]$$
$$= \frac{1}{8}[\,9 + 100 + 16 + 9\,]$$
$$= 16.75$$

- **Gradient Descent – Updating θ0 value**
  Here, j = 0

$$\theta_j := \theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}\left[\left(h_\theta(x_i) - y_i\right)x_i\right]$$
$$= 0 - \frac{0.001}{4}\left[(0-3) + (0-10) + (0-4) + (0-3)\right]$$
$$= \frac{0.001}{4}\left[-3 + (-10) + (-4) + (-3)\right]$$
$$= \frac{0.001}{4}[20]$$
$$= 0.005$$

- **Gradient Descent – Updating θ1 value**
  Here, j = 1

$$\theta_j := \theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}\left[\left(h_\theta(x_i) - y_i\right)x_i\right]$$
$$= 0 - \frac{0.001}{4}\left[(0-3)2 + (0-10)6 + (0-4)5 + (0-3)7\right]$$
$$= \frac{0.001}{4}\left[-6 + (-60) + (-20) + (-21)\right]$$
$$= \frac{0.001}{4}[107]$$
$$= 0.02657$$

Iteration 2 – θ0 = 0.005 and θ1 = 0.02657

- **Predicted values after iteration 1 with Linear regression hypothesis.**

$$h_\theta = \begin{bmatrix} \theta_0 & \theta_1 \end{bmatrix} \begin{bmatrix} x_0 & x_0 & x_0 & x_0 \\ x_1 & x_2 & x_3 & x_4 \end{bmatrix}$$

$$= \begin{bmatrix} 0.005 & 0.026 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 6 & 5 & 7 \end{bmatrix} = \begin{bmatrix} 0.057 & 0.161 & 0.135 & 0.187 \end{bmatrix}$$

Now, similar to iteration no. 1 performed above we will again calculate Cost function and update θj values using Gradient Descent.
We will keep on iterating until Cost function doesn't reduce further. At that point, model achieves best θ values. Using these θ values in the model hypothesis will give the best prediction results.