Simple Linear Regression

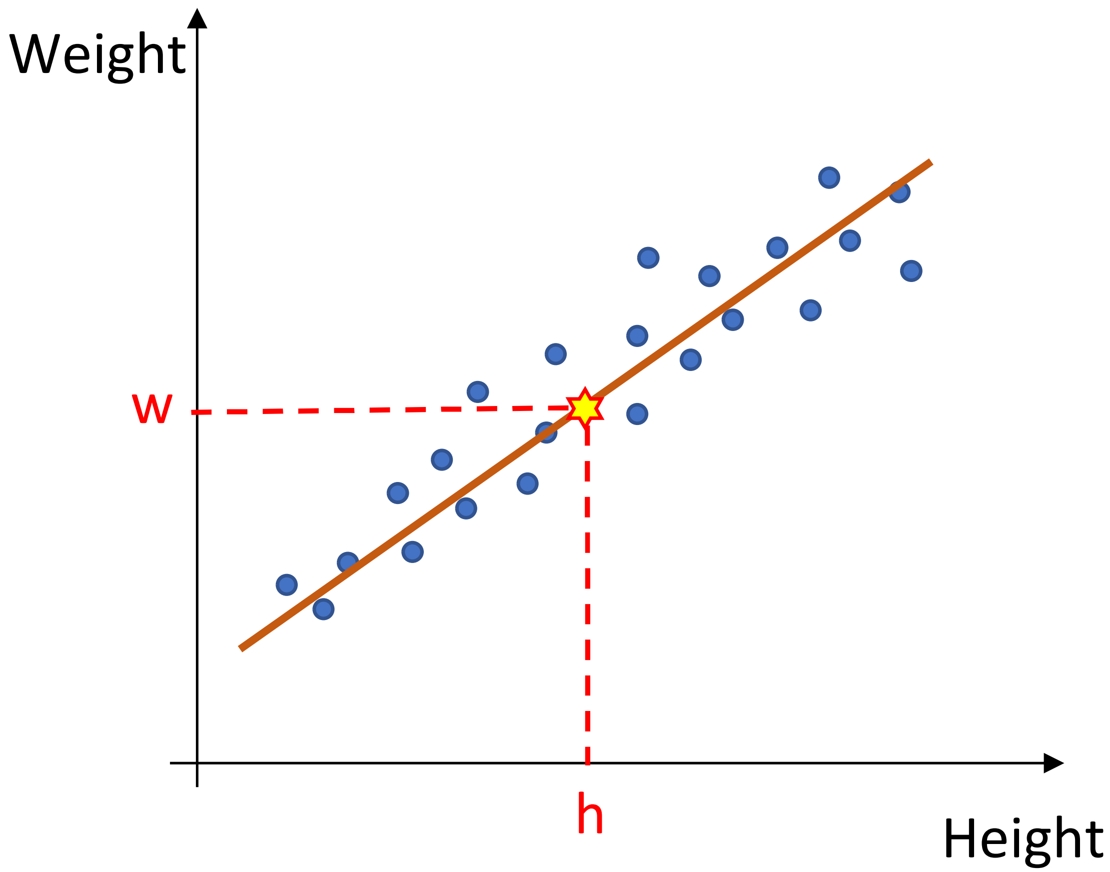
Sourav Karmakar

souravkarmakar29@gmail.com

OUTLINE

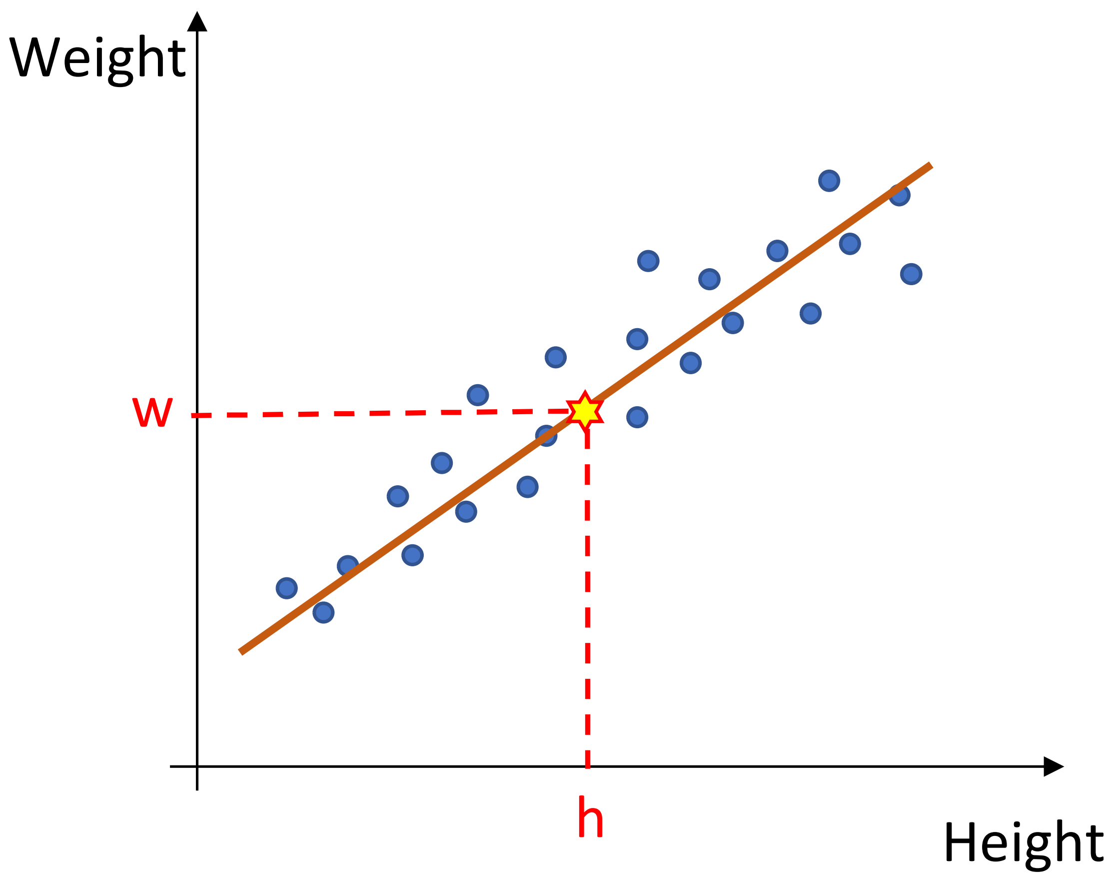
- Simple Linear Regression: Intuition
- Hypothesis function for Simple Linear Regression
- Mean Square Error Loss / Cost function
- Intuition of Cost Function
- Ordinary Least Square Regression

Simple Linear Regression : Intuition



- Consider the scatter plot of the Weight vs. Height of adults as shown beside.
- The trend or the form of the relationship is strongly positive.
- Now suppose we wish to estimate the **weight** of a person just by knowing his/ her **height**.
- In order to do so we first fit a straight line through our data points.
- Then from the graph, knowing the height we can find the weight of the corresponding person.
- Hence, we are intending to find out the **equation of the straight line** that ***best*** describes the relationship between Weight and Height.

Simple Linear Regression : Intuition



- There is only one predictor/input variable (Height) and one target variable (Weight) and we are intending to find out a relationship of the form:

$$\boldsymbol{y} = \boldsymbol{\theta_0} + \boldsymbol{\theta_1}\boldsymbol{x}$$

$$Here, y\ is\ the\ target\ variable\ and$$
$$x\ is\ the\ predictor\ variable$$

- We have to find out $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_1}$, such that the straight line $\boldsymbol{y} = \boldsymbol{\theta_0} + \boldsymbol{\theta_1}\boldsymbol{x}$ fits into our dataset **best**.

- This is called Simple Linear Regression, because it has only one predictor variable and the relationship among target and predictor variable is linear.

Simple Linear Regression : Hypothesis

Simple Linear Regression Model with Single Predictor

$$\underbrace{\boldsymbol{y}}_{\text{Target Variable}} = \underbrace{\boldsymbol{\theta}_0}_{\text{Intercept}} + \underbrace{\boldsymbol{\theta}_1}_{\text{Slope}} \underbrace{\boldsymbol{x}}_{\text{Predictor Variable}} + \underbrace{\boldsymbol{\varepsilon}}_{\text{Random Residual Error}}$$

$\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are called model parameters

- We use our sample data to find estimates for the coefficients/ model parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ i.e.: $\widehat{\boldsymbol{\theta}_0}$ and $\widehat{\boldsymbol{\theta}_1}$.
- We can then **predict** what the value of $\boldsymbol{y}$ should be corresponding to a particular value for $\boldsymbol{x}$ by using the Least Squares Prediction Equation (also known as our **hypothesis function**):

$$\widehat{\boldsymbol{y}} = \widehat{\boldsymbol{\theta}_0} + \widehat{\boldsymbol{\theta}_1}\boldsymbol{x}$$

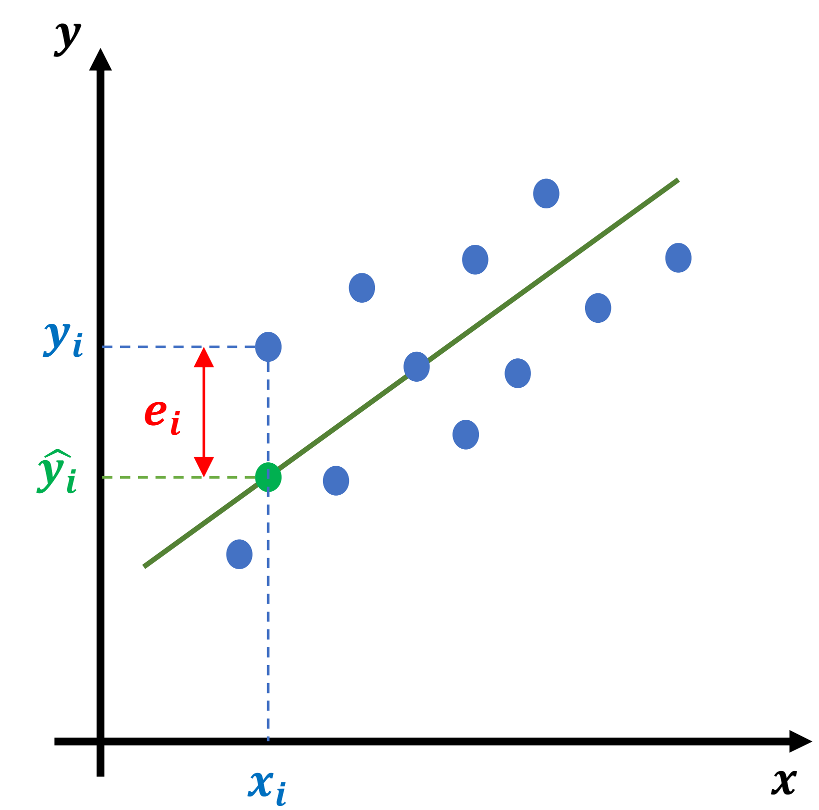
Where $\widehat{\boldsymbol{y}}$ is our prediction for $\boldsymbol{y}$

Simple Linear Regression : Cost Function

Residuals and Residual Sum of Squares:

- For i^{th} sample $\langle \boldsymbol{x_i}, \boldsymbol{y_i} \rangle$ the predicted value of $\boldsymbol{y_i}$ is $\widehat{\boldsymbol{y}}_{\boldsymbol{i}}$, Which we obtain from the equation $\widehat{\boldsymbol{y}}_{\boldsymbol{i}} = \widehat{\boldsymbol{\theta_0}} + \widehat{\boldsymbol{\theta_1}}\, \boldsymbol{x_i}$

- Then, $\boldsymbol{e_i} = \boldsymbol{y_i} - \widehat{\boldsymbol{y}}_{\boldsymbol{i}}$ (actual – predicted) represents the i^{th} residual.



- We define **Residual Sum of Squares** (RSS) as:

$$RSS = \sum_{i=1}^{m} \boldsymbol{e_i}^2 = \sum_{i=1}^{m} (\boldsymbol{y_i} - \widehat{\boldsymbol{y}}_{\boldsymbol{i}})^2 = \sum_{i=1}^{m} (\boldsymbol{y_i} - (\widehat{\boldsymbol{\theta_0}} + \widehat{\boldsymbol{\theta_1}}\, \boldsymbol{x_i}))^2$$

There are total m no. of samples

Simple Linear Regression : Cost Function

Mean Square Error Cost Function:

- We can define the cost function as:

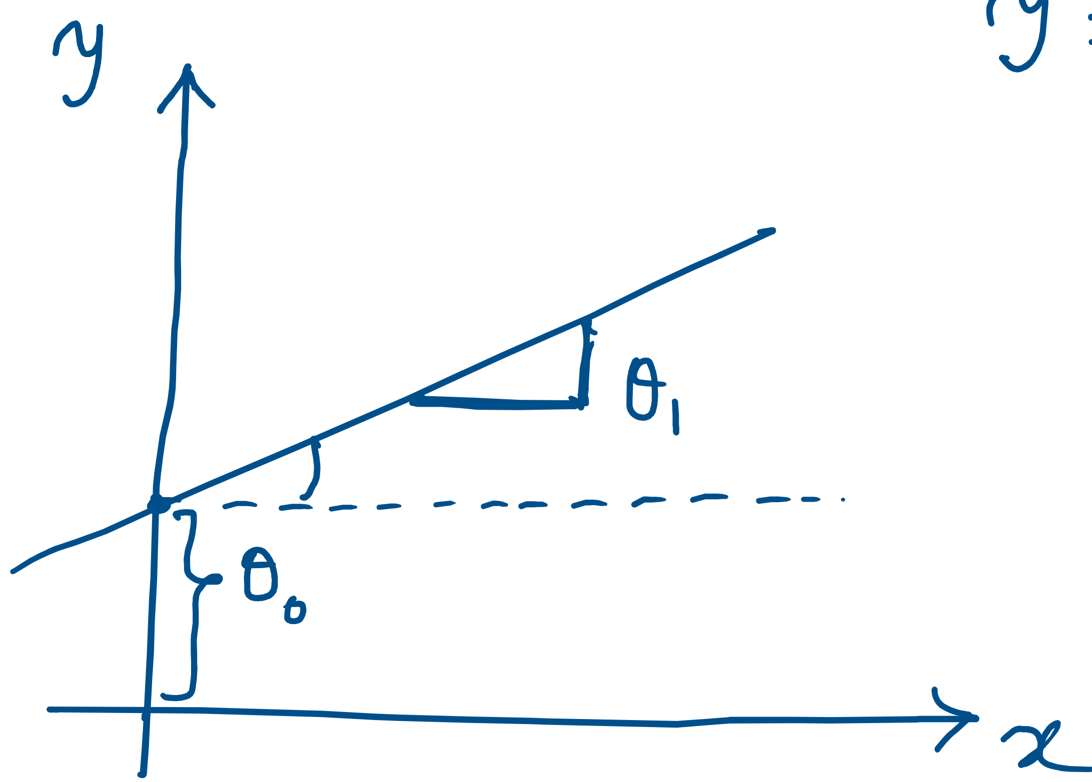
$$J(\widehat{\theta_0}, \widehat{\theta_1}) = \frac{1}{2}\frac{RSS}{Number\ of\ training\ samples} = \frac{1}{2m}\sum_{i=1}^{m}(y_i - (\widehat{\theta_0} + \widehat{\theta_1}\, x_i))^2$$

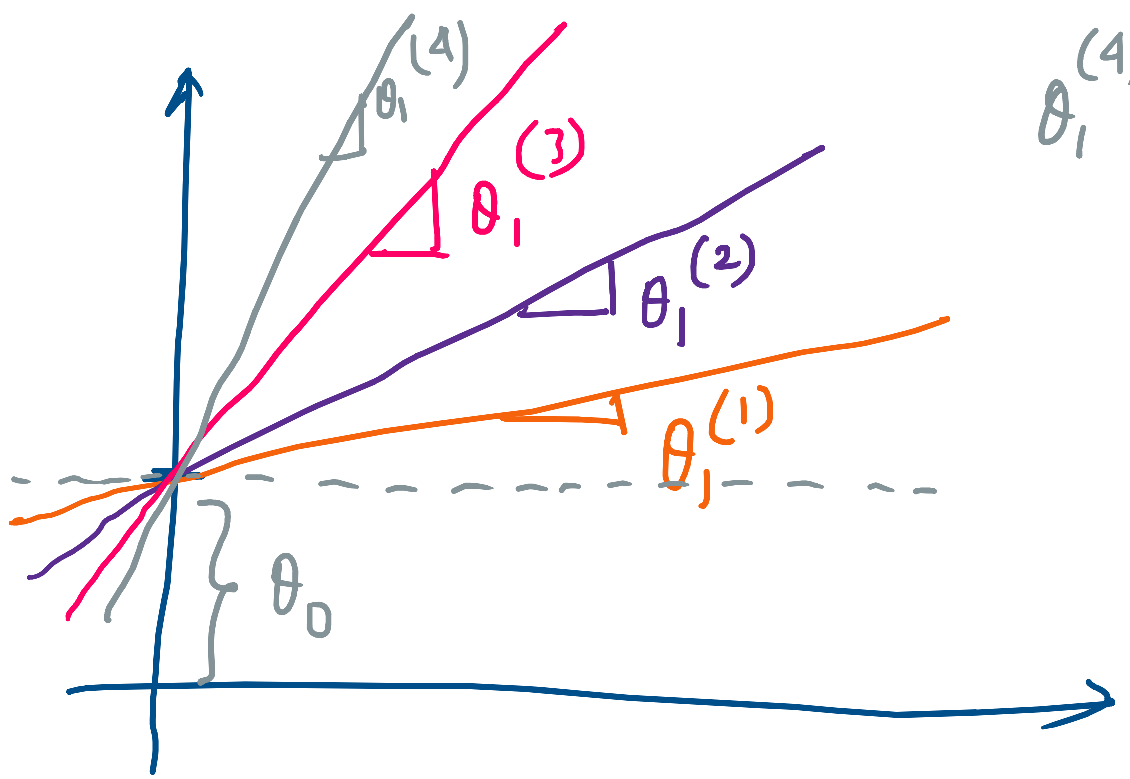
Here a factor $\frac{1}{2}$ is multiplied just for computational simplicity. Otherwise, the cost function $J(\widehat{\theta_0}, \widehat{\theta_1})$ is nothing but **mean or average of the Residual sum of squares.** (also known as **Mean Square Error (MSE)**).

Our Objective:

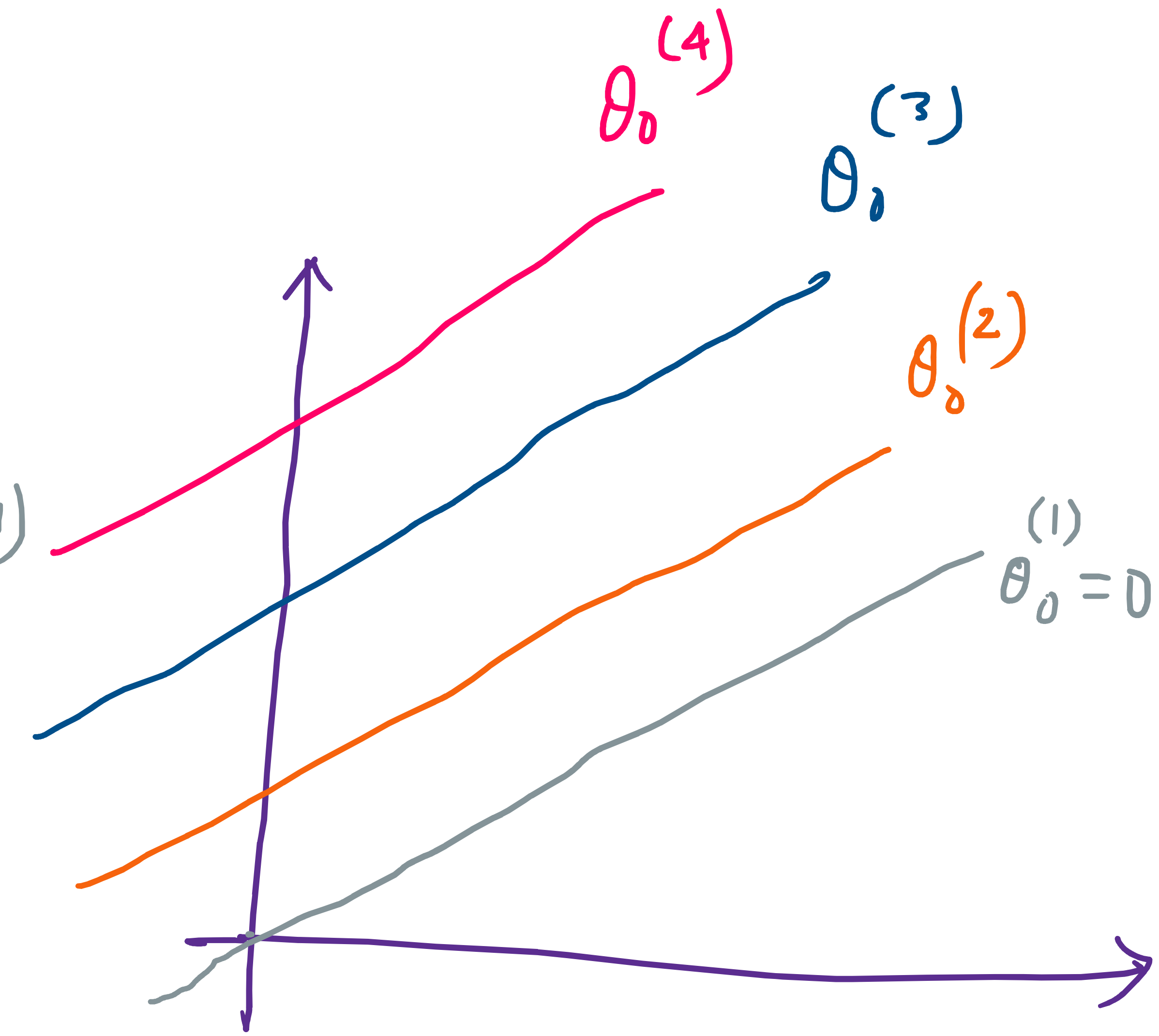
- To find the suitable values of $\widehat{\theta_0}$ and $\widehat{\theta_1}$ such that the cost function $J(\widehat{\theta_0}, \widehat{\theta_1})$ is minimized, in other words the Residual Sum of Square **(RSS)** is minimized. Then the straight line $\hat{y} = \widehat{\theta_0} + \widehat{\theta_1}\, x$ will fit our data **best**. This is called **least squares fit**.

$$y = \underbrace{\theta_0}_{\text{intercept}} + \underbrace{\theta_1}_{\text{slope.}} x$$





$$\theta_1^{(4)} > \theta_1^{(3)} > \theta_1^{(2)} > \theta_1^{(1)}$$



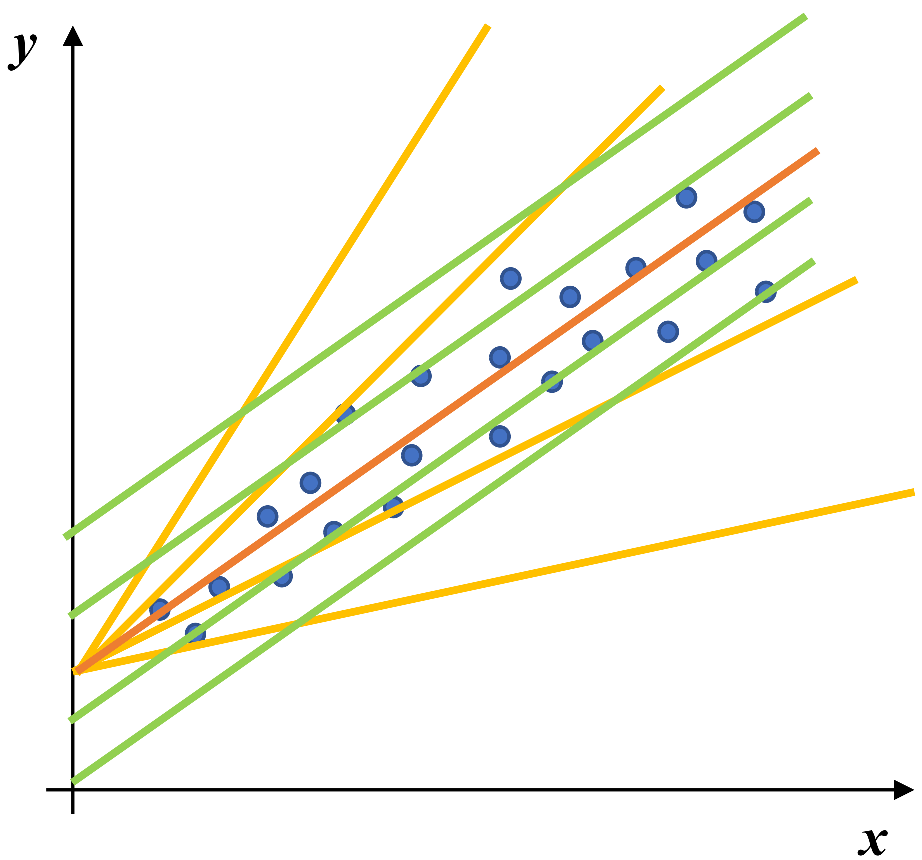
$$\theta_0^{(4)} > \theta_0^{(3)} > \theta_0^{(2)} > \theta_0^{(1)}$$

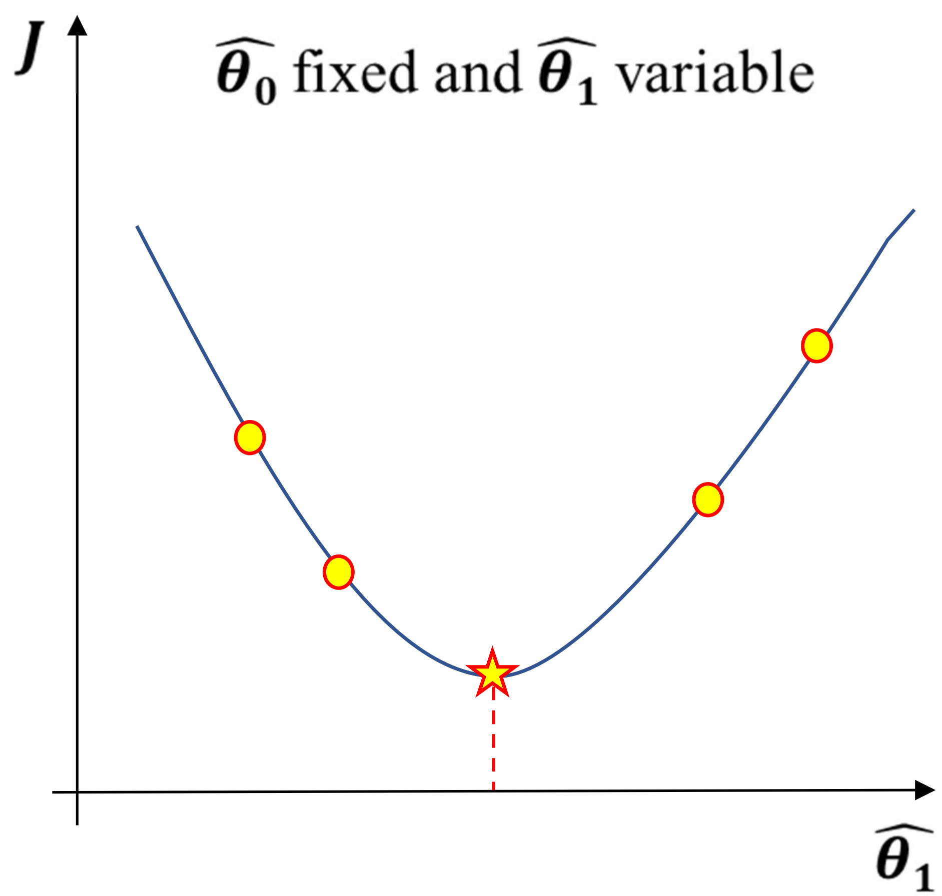
Simple Linear Regression : Cost Function

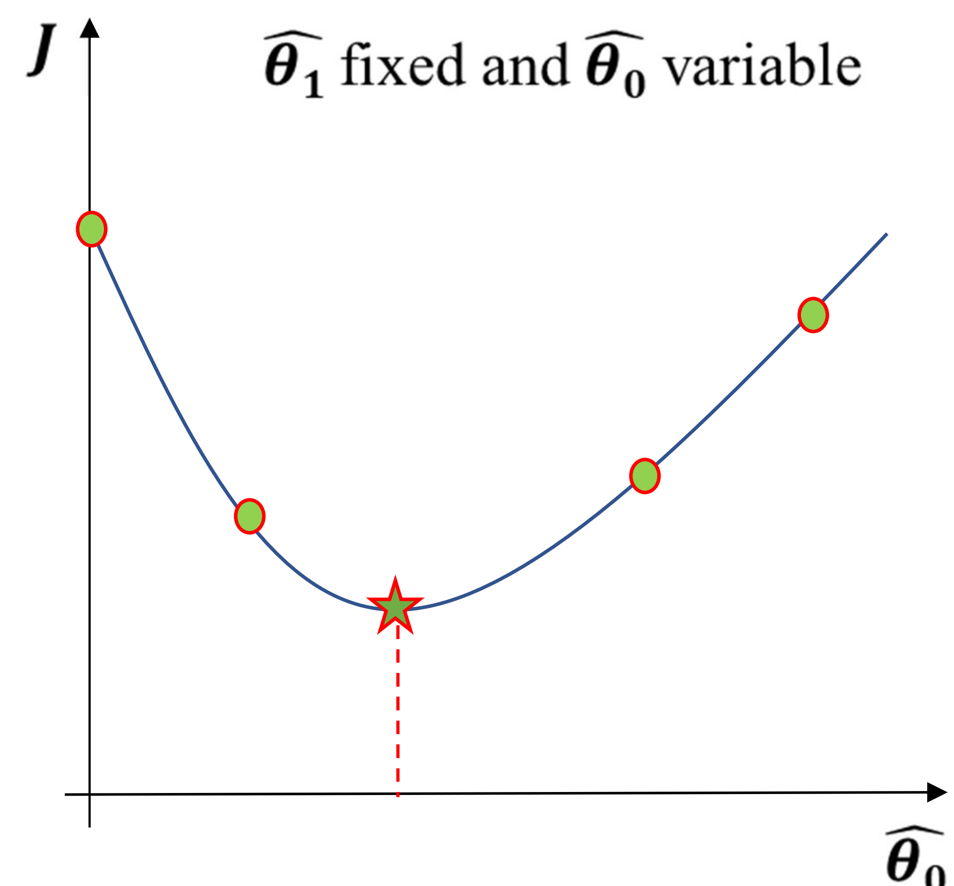
Intuition of Cost Function:

Consider the example of single predictor variable where the hypothesis function is $\boldsymbol{\hat{y}} = \boldsymbol{\widehat{\theta_0}} + \boldsymbol{\widehat{\theta_1}}\,\boldsymbol{x}$ and the cost function is $\boldsymbol{J}\left(\boldsymbol{\widehat{\theta_0}}, \boldsymbol{\widehat{\theta_1}}\right) = \frac{\mathbf{1}}{\mathbf{2}\boldsymbol{m}}\sum_{i=1}^{m}(\boldsymbol{y_i} - (\boldsymbol{\widehat{\theta_0}} + \boldsymbol{\widehat{\theta_1}}\,\boldsymbol{x_i}))^2$.

Now we keep one parameter fixed and vary other. Let's see how $\boldsymbol{J}\left(\boldsymbol{\widehat{\theta_0}}, \boldsymbol{\widehat{\theta_1}}\right)$ varies.







- Our objective is to find the values of the parameters for which the cost function is minimized.

Simple Linear Regression : OLS fit

Solving for the best fit: Ordinary Least Squares (OLS) Regression:

- We have to Minimize $\boldsymbol{RSS}$ or $\boldsymbol{J}\left(\widehat{\boldsymbol{\theta}_0}, \widehat{\boldsymbol{\theta}_1}\right)$ with respect to $\widehat{\boldsymbol{\theta}_0}$ and $\widehat{\boldsymbol{\theta}_1}$
- Hence we have to do, $\frac{\partial}{\partial \widehat{\boldsymbol{\theta}_0}}(\boldsymbol{RSS}) = \mathbf{0}$ and $\frac{\partial}{\partial \widehat{\boldsymbol{\theta}_1}}(\boldsymbol{RSS}) = \mathbf{0}$
- By solving the above two equations we get the following value of $\widehat{\boldsymbol{\theta}_1}$ and $\widehat{\boldsymbol{\theta}_0}$:

$$\widehat{\boldsymbol{\theta}_1} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x})^2} = r_{xy}\frac{\sigma_y}{\sigma_x} \quad and \quad \widehat{\boldsymbol{\theta}_0} = \bar{y} - \widehat{\boldsymbol{\theta}_1}\bar{x}$$

$where, \bar{x}\ is\ the\ mean\ of\ predictor\ variable\ x\ and\ \bar{y}\ is\ the\ mean\ of\ target\ variable\ y$

$\sigma_x\ is\ the\ standard\ deviation\ of\ x\ and\ \sigma_y\ is\ the\ standard\ deviation\ of\ y$

$and\ r_{xy}\ is\ the\ \boldsymbol{correlation\ coefficient}\ between\ x\ and\ y.$

Calculation of Optimum values of model Parameters $\hat{\theta}_0$ & $\hat{\theta}_1$:-

$$J = \frac{1}{2m}\sum_{i=1}^{m}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \frac{1}{2m}\sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

$$\hat{y}^{(i)} = \hat{\theta}_0 + \hat{\theta}_1 x^{(i)}$$

$$\therefore J = \frac{1}{2m}\sum_{i=1}^{m}\left(\hat{\theta}_0 + \hat{\theta}_1 x^{(i)} - y^{(i)}\right)^2$$

$$\frac{\partial J}{\partial \hat{\theta}_0} = \frac{1}{2m}(\cancel{2})\sum_{i=1}^{m}\left(\hat{\theta}_0 + \hat{\theta}_1 x^{(i)} - y^{(i)}\right)$$

$$= \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_0 + \hat{\theta}_1 \underbrace{\boxed{\frac{1}{m}\sum_{i=1}^{m} x^{(i)}}}_{\bar{x}} - \underbrace{\boxed{\frac{1}{m}\sum_{i=1}^{m} y^{(i)}}}_{\bar{y}}$$

$$\frac{\partial J}{\partial \hat{\theta}_0} = \hat{\theta}_0 + \hat{\theta}_1 \bar{x} - \bar{y}$$

$$\frac{\partial J}{\partial \hat{\theta}_0} = 0 \Rightarrow \hat{\theta}_0 + \hat{\theta}_1 \bar{x} - \bar{y} = 0$$

$$\Rightarrow \boxed{\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}}$$

$$J = \frac{1}{2m}\sum_{i=1}^{m}\left(\bar{y} - \hat{\theta}_1\bar{x} + \hat{\theta}_1 x^{(i)} - y^{(i)}\right)^2$$

$$J = \frac{1}{2m}\sum_{i=1}^{m}\left(\hat{\theta}_0 + \hat{\theta}_1 x^{(i)} - y^{(i)}\right)^2$$

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1\bar{x}$$

$$= \frac{1}{2m}\sum_{i=1}^{m}\left(\bar{y} + \left(x^{(i)} - \bar{x}\right)\hat{\theta}_1 - y^{(i)}\right)^2$$

$$\therefore \frac{\partial J}{\partial\hat{\theta}_1} = \frac{1}{m}\sum_{i=1}^{m}\left(\bar{y} + \left(x^{(i)} - \bar{x}\right)\hat{\theta}_1 - y^{(i)}\right)\left(x^{(i)} - \bar{x}\right)$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left[\left(x^{(i)} - \bar{x}\right)\hat{\theta}_1 - \left(y^{(i)} - \bar{y}\right)\right]\left(x^{(i)} - \bar{x}\right) = \underbrace{\boxed{\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \bar{x}\right)^2}}_{Var(x)}\hat{\theta}_1 - \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \bar{x}\right)\left(y^{(i)} - \bar{y}\right)$$

$$= Var(x)\cdot\hat{\theta}_1 - \underbrace{\boxed{\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \bar{x}\right)\left(y^{(i)} - \bar{y}\right)}}_{Cov(x,y)}$$

$$\frac{\partial J}{\partial \hat{\theta}_1} = Var(x) \cdot \hat{\theta}_1 - Cov(x,y) = 0$$

$$\Rightarrow \hat{\theta}_1 = \frac{Cov(x,y)}{Var(x)}$$

$$r_{xy} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y} = \text{Correlation coefficient between } x \,\&\, y$$

$$\Rightarrow Cov(x,y) = r_{xy} \cdot \sigma_x \cdot \sigma_y$$

$$Var(x) = \sigma_x^2$$

$$\hat{\theta}_1 = \frac{r_{xy} \cdot \sigma_x \cdot \sigma_y}{\sigma_x^2}$$

$$\boxed{\hat{\theta}_1 = r_{xy} \cdot \frac{\sigma_y}{\sigma_x}} \quad \& \quad \boxed{\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}}$$

Thank You