Short & Sweet (Episode 1)

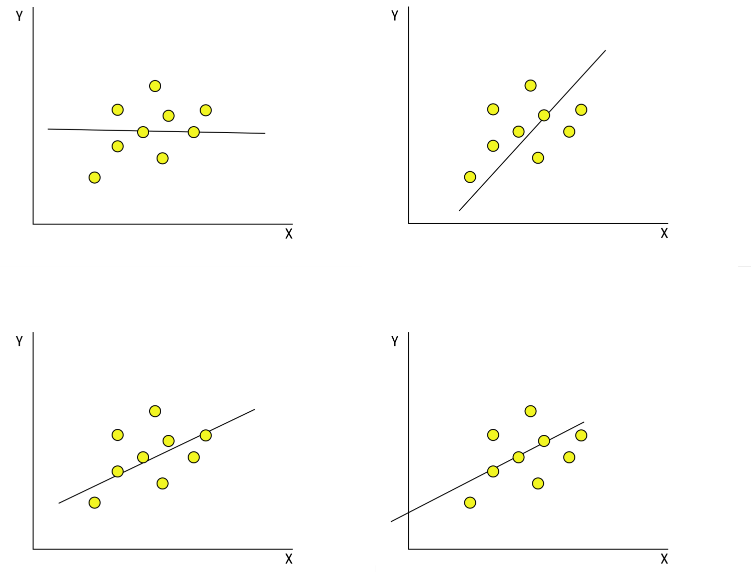
Linear Regression

Linear Regression is a **Machine Learning Algorithm** which helps to establish linear relationship between **independent variables** and **target variable** and can predict the outcome through unseen data comes to it. **Why Linear Regression??**

- Simple to implement.
- It is less complex as compared to other algorithms.
- It is susceptible to overfitting, but it can be avoided using dimensionality reduction and other techniques.

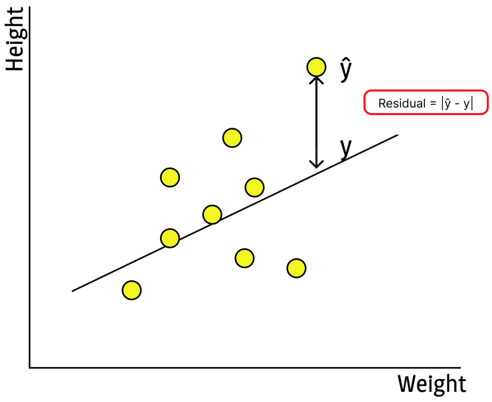
How It Works

As it is by name, we'll try to fit a line / build some linear relationship between features and target. For this, we'll fit a line such that it coincides with majority of datapoints. **"More point with which our line intersects, more the line is accurate and best fitted."**



We can see that line of **bottom two figures** looks like fitted good. So that's we simply need it.

- Now, try different lines such that it intersects with majority datapoints.
- By different lines, I mean changing magnitude of parameters of equation of line i.e. **"y = wx + b".**
- Here, **w** stands for **slope/weight** and **b** stands for **intercept/bias**.
- **Weights (or slope)** determine the contribution of each feature to the predicted output. Larger weights indicate a more significant influence of the corresponding feature on the prediction.
- **Bias (or intercept)** adjusts the overall prediction, allowing the model to make accurate predictions even when all input features are zero.
- So we'll try to vary the values of weights and bias so that line rotates, move upward and downward. But how we analyze we reach to that best line mathematically.
- We'll calculate difference between actual datapoints and datapoints which is predicted by line. Each time when we vary the value of weights and bias, we calculate the difference for each datapoints which is called **residuals** and sum it up.



- Every time we iterate, we'll try to reduce it, more the sum is less, more the fitted line is good.
- There are many methods through which we can sum up the residuals which are:

  - **Mean Squared Error**

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  - **Root Mean Squared Error**

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

  - **Mean Absolute Error**

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

  - **R squared**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Mainly we used **Root Mean Squared Error** and **R squared** for estimation of best fitted line. Mainly these functions called **cost function**.

- Now, how do we reach to optimal weights and bias quickly , so here we reach to concept of **Gradient Descent**.

Update parameters quickly using Gradient Descent

Gradient Descent is a way to plot the relationship between **weights** with **cost function** to update the weight parameter and **bias** with **cost function** to update the bias parameter.

- Formula to update weights parameter.

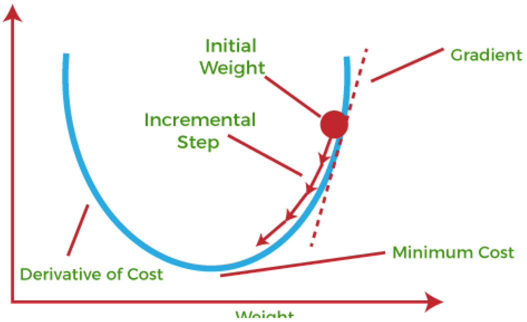
$$w_j = w_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}$$

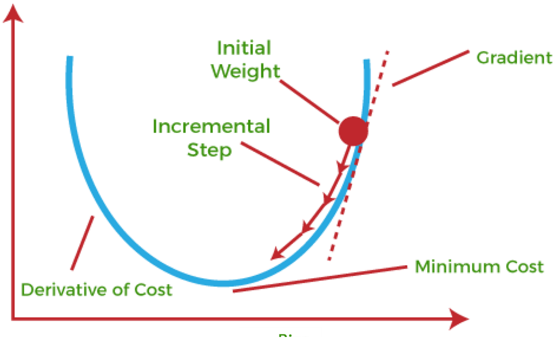
- Formula to update bias parameter.

$$b = b - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})$$

Here,

- **W_j** is the weight which we are going to update.
- **α** is the learning rate.
- **h_θ** is the cost function that we’ve discussed before.
- **x(i)** is the independent feature on which we are calculating residuals and then cost function.
- **y(i)** is the actual datapoints.
- **b** stands for bias.





- Until now, we are talking about only one independent feature and one target feature but in real life, we have more crucial independent features. In that case, number of weights parameters increases. Like, **“y = $w^1x^1 + w^2x^2 + w^3x^3$ +....... + b”.** Regarding this, our **cost function** and **Gradient descent optimization technique** becomes more intense.

Assumptions for Linear Regression

- There exists a linear relationship between the independent variable, x, and the dependent variable, y.
- The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
- The residuals have constant variance at every level of x.

- The residuals of the model are normally distributed.