What is Regression Analysis?

Regression Notes

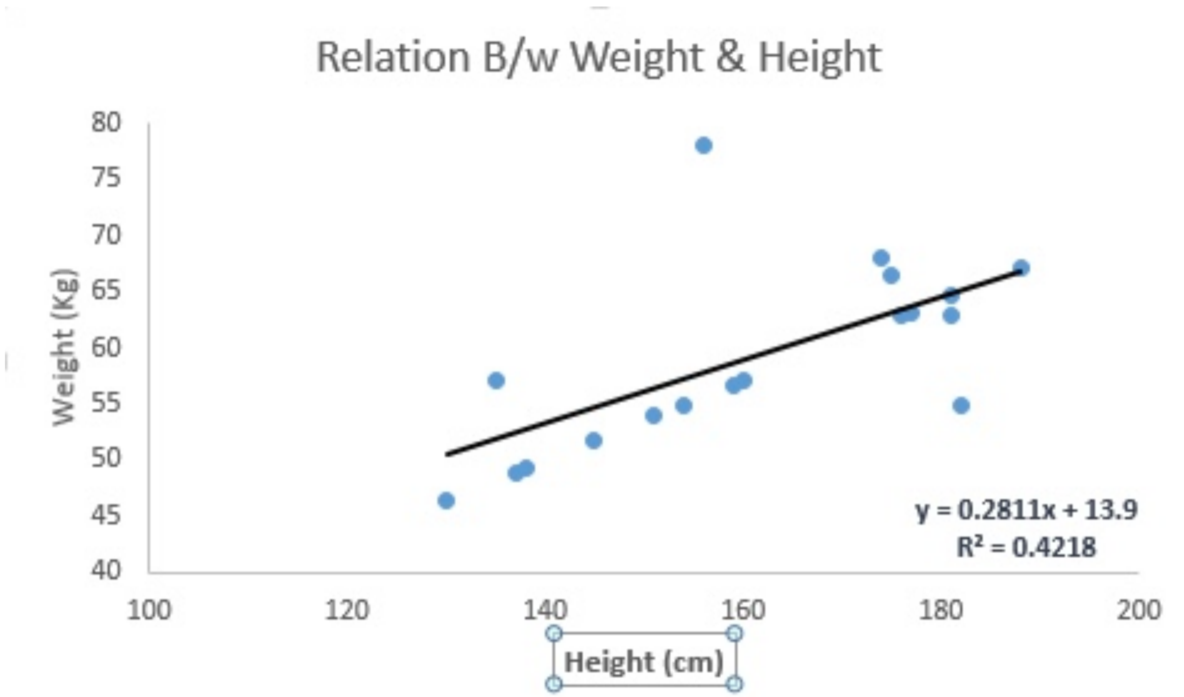
Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor).

Example–Let's say, you want to estimate growth in sales of a company based on current economic conditions. You have the recent company data which indicates that the growth in sales is around two and a half times the growth in the economy. Using this insight, we can predict future sales of the company based on current & past information.

Linear Regression

Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line).

- It is represented by an equation Y=a+b*X + e, where a is intercept, b is slope of the line and e is error term.
- This equation can be used to predict the value of target variable based on given predictor variable(

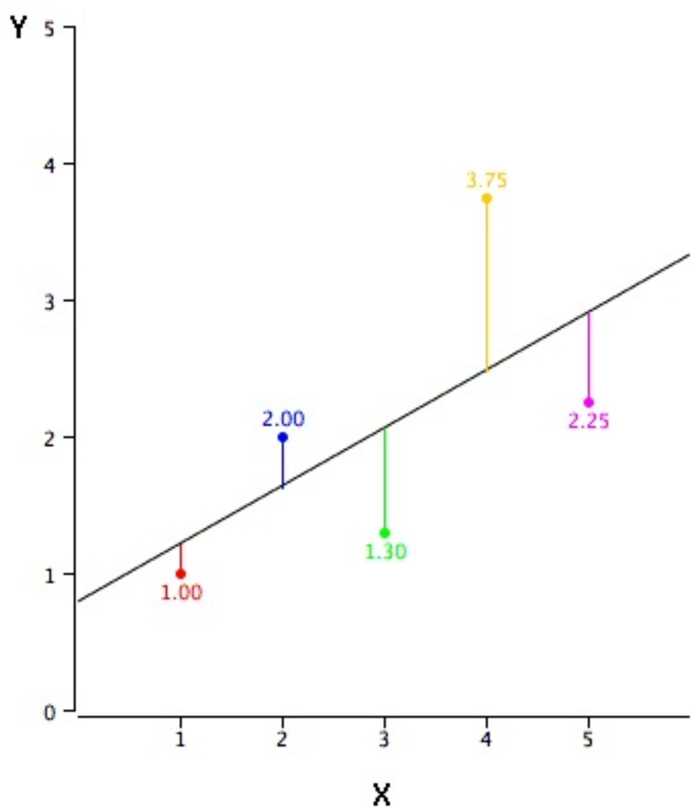


The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable.

Now, the question is "How do we obtain best fit line?".

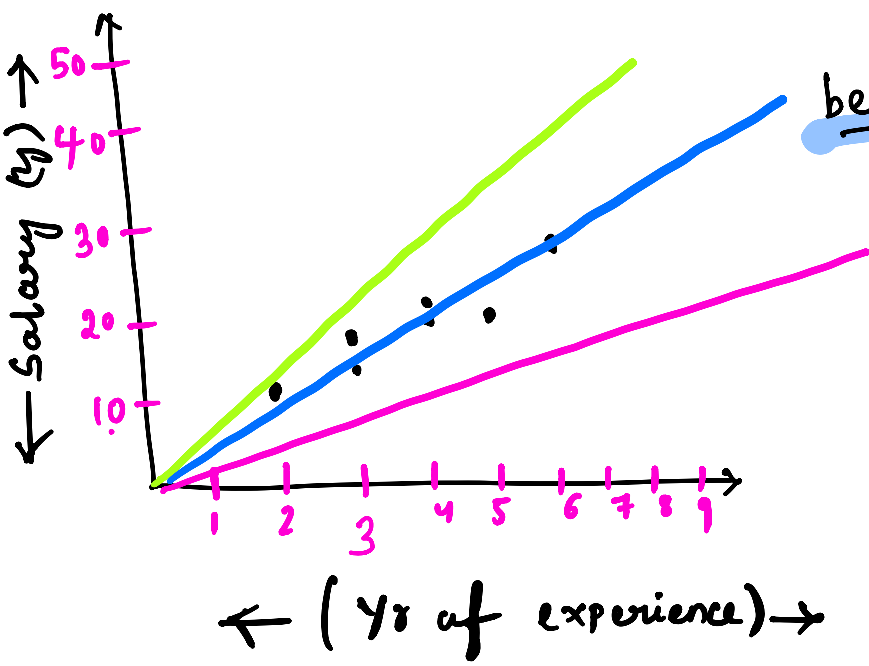
How to obtain best fit line (Value of a and b)?

- This task can be easily accomplished by Least Square Method or using Gradient Descent Method.
- Least Square is the most common method used for fitting a regression line. It calculates the best-fit line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line.



- Because the deviations are first squared, when added, there is no cancelling out between positive and negative

Year of experience	Salary
(X)	(Y)
2	10
3	15
4	18
5	20
4	20
3	14
6	26



best fit line

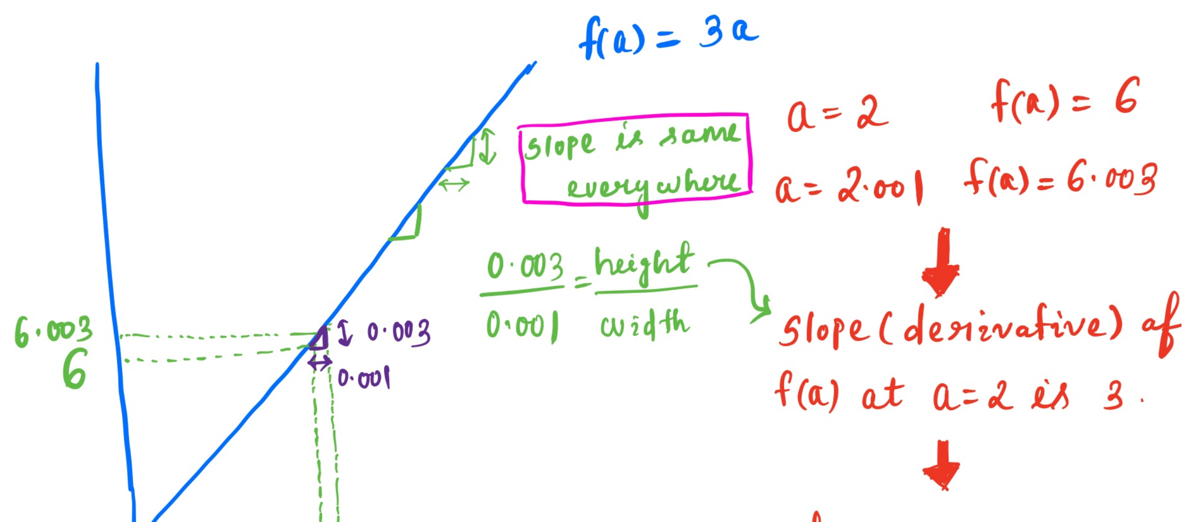
↓

why ?

↓

most closest line to all points.

Intuition about Derivatives



$$\frac{d\, f(a)}{d\, a} = 3$$

Costfunction

↳ A cost function basically tells us how good our model is at making Predictions for a given value of m and b.

Note- Loss function computes the error from a single training exampe, while cost function is the average of the loss functions for all the training example.

$$\boxed{\text{cost} = \frac{1}{N} \sum_{i=1}^{N} (y' - y)^2}$$

Goal $\rightarrow$ Minimize the cost function.

$\downarrow$

why ?

$\downarrow$

lower error $\longrightarrow$ signifies that the algorithm has done a good job in learning.

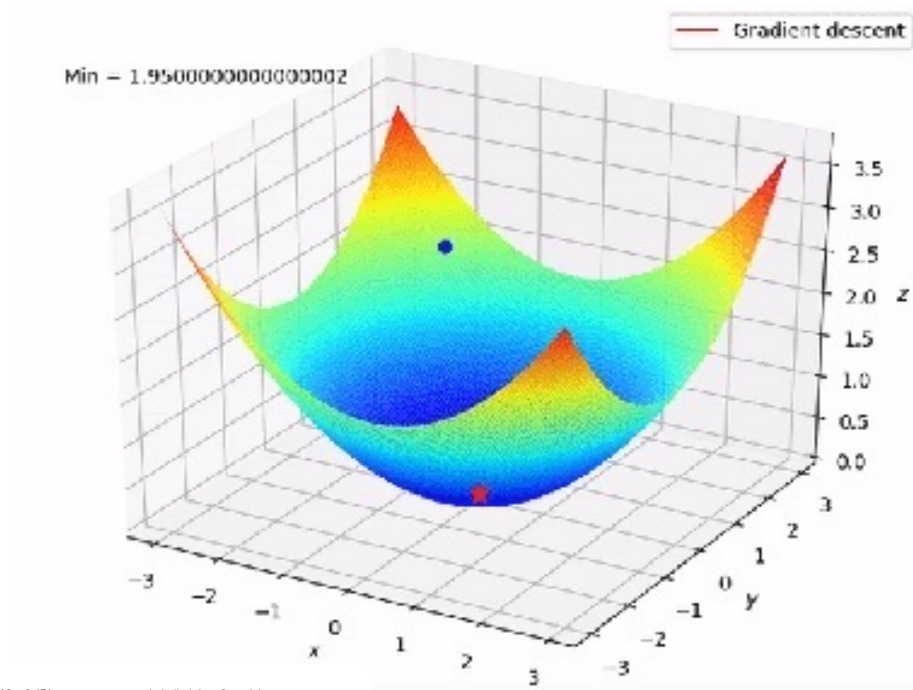
$\downarrow$

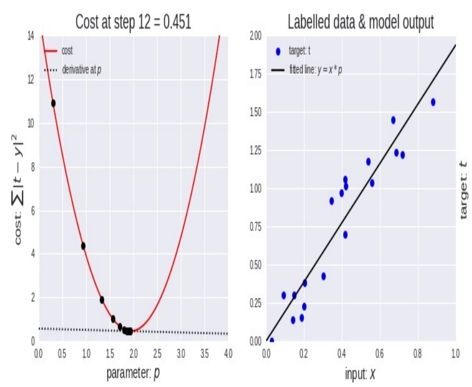
ultimately we want m, b values which gives the smallest possible error.

How to minimize ?

↓

Gradient Descent





↓

In higher dimensions, we need an algorithm to locate the minima, that algorithm is called Gradient Descent.

$$\text{cost} = \frac{1}{N}\sum_{i=1}^{N}(y_i' - y_i)^2$$

$$J_{m,b} = \frac{1}{N}\sum_{i=1}^{N}(\text{Error}_i)^2$$

Find the derivative of the cost function w.r.t both m and b.

$$\frac{\delta J}{\delta m} = 2 \cdot \text{Error} \cdot \frac{\delta}{\delta m}\,\text{Error}$$

$$\frac{\delta J}{\delta b} = 2 \cdot \text{Error} \cdot \frac{\delta}{\delta b} \cdot \text{Error}$$

Let's calculate the gradient of error w.r.t m & b.

$$\frac{\delta}{\delta m}\,\text{Error} = \frac{\delta}{\delta m}(y' - y) = \frac{\delta}{\delta m}(mx + b - y)$$

constants

$$\boxed{\frac{\delta}{\delta m}(\text{Error}) = x}$$

$$\frac{\delta}{\delta b} Error = \frac{\delta}{\delta b}(y' - y)$$

$$= \frac{\delta}{\delta b}(mX + b - y)$$

$$\boxed{\frac{\delta}{\delta b} Error = 1}$$

Plugging the values back in the cost function and multiplying it with learning rate,

$$\frac{\delta J}{\delta m} = 2 \cdot Error * X * LearningRate$$

$$\frac{\delta J}{\delta b} = 2 \cdot Error * LearningRate$$

updating slopes / weights

$$m = m - \delta m$$

$$\boxed{m' = m^0 - Error * X * LearningRate}$$

updating b

$$\boxed{b' = b^0 - Error * LearningRate}$$

Important Points:

- There must be linear relationship between independent and dependent variables
- Multiple regression suffers from multicollinearity, autocorrelation, heteroskedasticity.
- Linear Regression is very sensitive to Outliers. It can terribly affect the regression line and eventually the forecasted values.
- Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable
- In case of multiple independent variables, we can go with forward selection, backward elimination and step wise approach for selection of most significant independent variables.