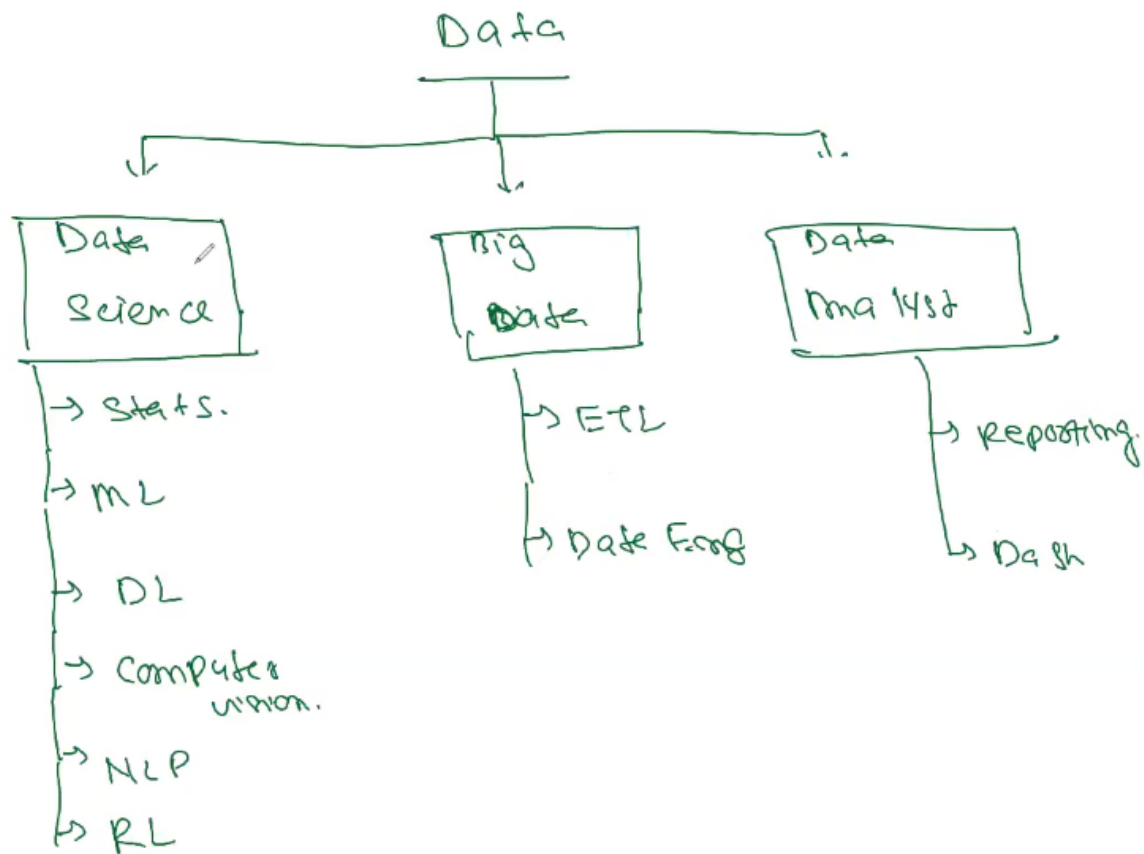
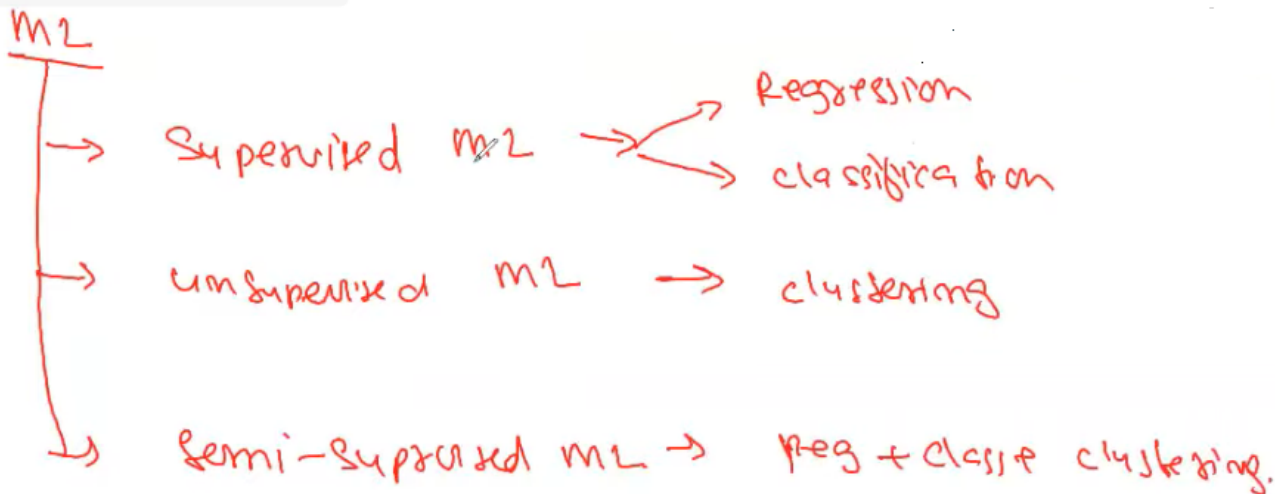


machine learning



Dash=Dashboarding



Supervised ML-Is where we are giving both input and output variables and the machine is going to find out the relationship between them and can predict the out put for the new data based on it.

Regression - Means where we are looking for exact value as the predicted out put

Classification= Means where we are looking for categories as the output. Time series falls under regression.

Unsupervised-Is where we only give the input data system will find the patterns based on distance, size, or properties of the data it tries to group the data into different groups as output.

Semi supervised-

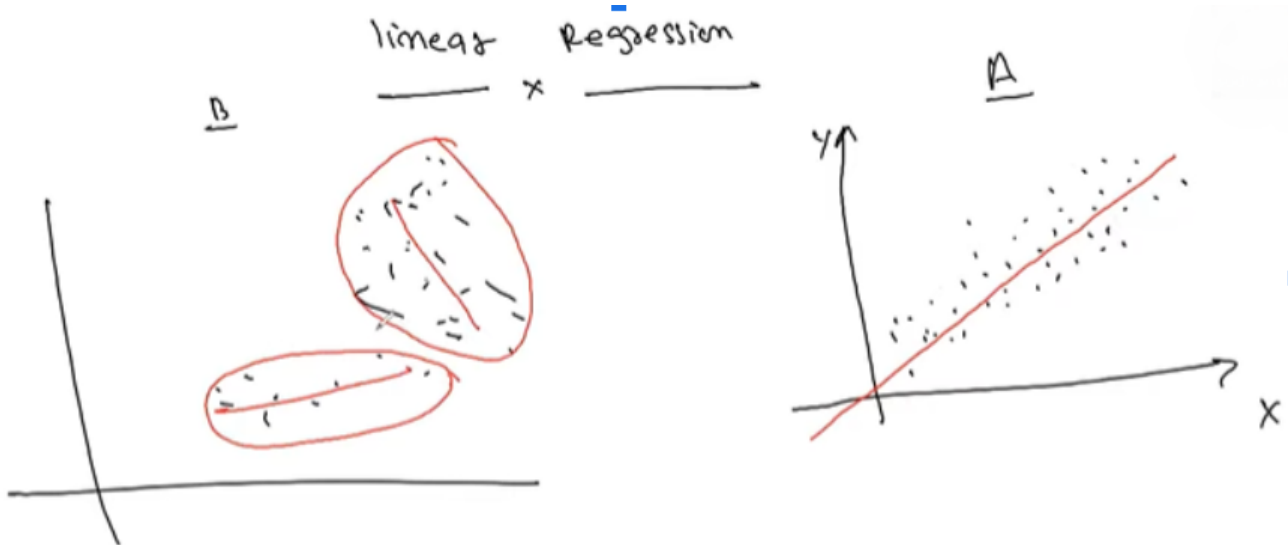
DL- In DL also try to find the patterns as in ML

Datascientist- If you are able to find out the different different patterns such as relationship between the data using either ML, DL, Data analytics then you are a true data scientist.

ML- What ML does is it is going to find out the best fit values for the hyper parameters of the given equation. ML it self does not find the equation, pattern as a data scientist we choose the mathematical equation or algorithms suitable for the data, we Identify the pattern, then we feed these things to the machine machine will be able to find out the best fit parameter and then learns the pattern and based on it predicts output for the new data.

Algorithms

Linear regressions: Falls under supervised ML. Supervised means we provide input and output data at the time of training. Regression if we are trying to find out the exact value based on thenature of the problem statement then we call it as regression. Inside the regression problem statement we are always try to find out the relationship, Here we try to find out the pattern or relation between one variable with other variable, how one variable related to other variable or one data set related to the other dataset. If we are able to establish the relation then we are good and will build an equation based on it. To understand we draw scatter diagram to find out if there is any pattern.



In A we can find the linear relationship between x and y. $y=mx+c$ this is a equation of straight line or slope.

Hieght and wieght,

Temperature and sales of icecreame

One is increasing other is also increasing.

$y= mx+c$,

m=means slope or tangent or tan Theeta = y/x or dy/dx . m is trying to define relationship or proportional or coefficient or rate of change of y with respect to x

c=means intersection, or intercept or error

If $c=0$, $y=mx$, $m= y/x$, $c=0$ means my line is starting from origin.

If $m=0.5$

$$y=0.5x$$

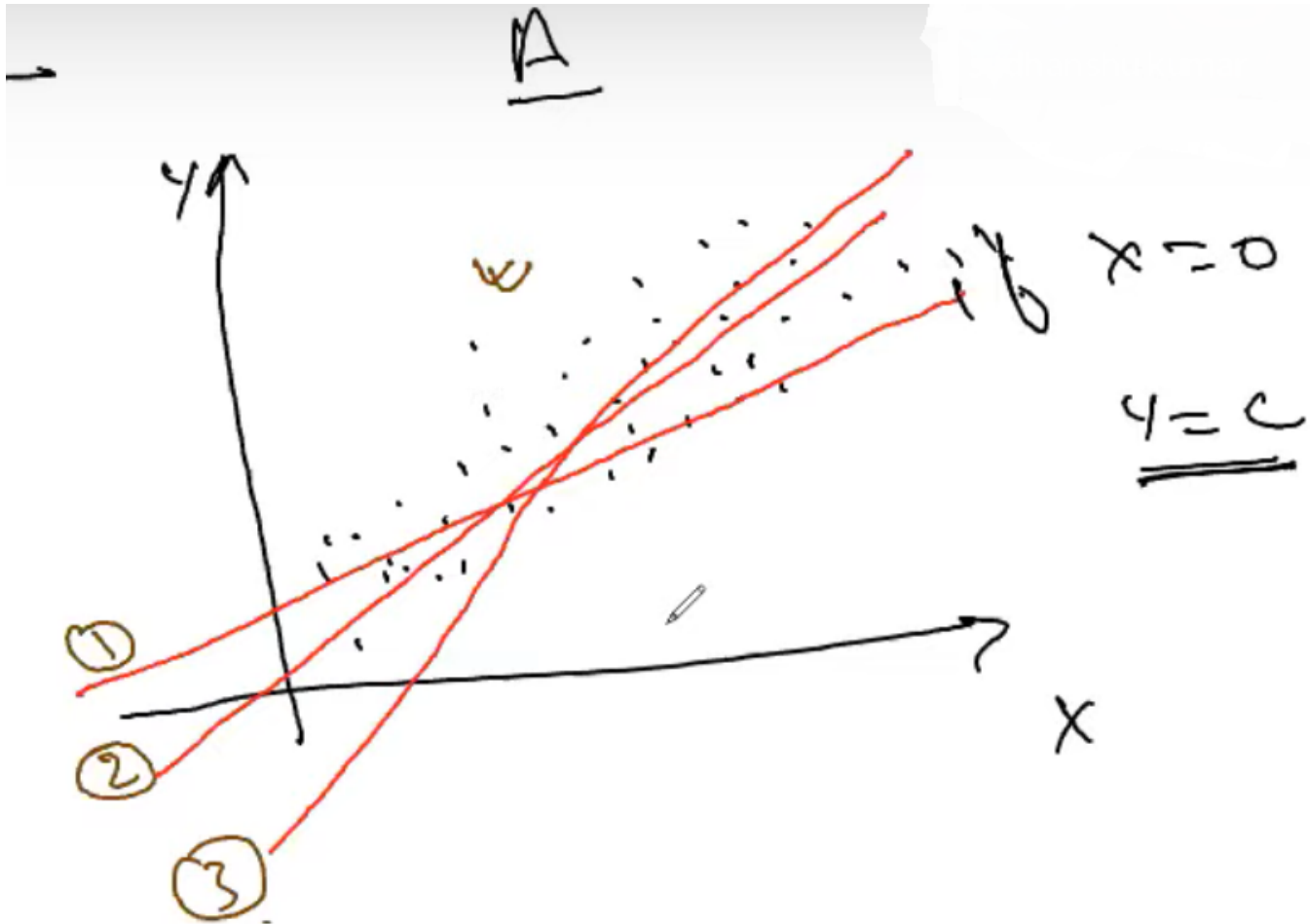
If one unit change in x then y is going to change by 0.5 unit.

What is c , if $x=0$, $y=c$. i.e. whenever x is zero then $y=c$. X will be zero at the origin the $c=y$ or $y=c$.

I.e. in no change of x then y is affected by c unit

Any change in the m and c value you will get the different line or slopes. m and c draws the line.

But we will look for only one line. The line which is called as best fit line. Best fit line is the one which can generalise the entire data with less error. The best fit line will be provided by best value of m and c .



How to find the value of m and c: Let us consider the data of weight and height as below.

As we know $y = mx + c$, here our x is H and y is w . So $w = mH + e$. Now we have value for H and W so we can find out the value of m and e . Let us take a two observations and form 2 equations by substituting the value of H and W

$$50 = m * 5.1 + e \text{ -----(1)}$$

$$49 = m * 5.2 + e \text{ -----(2)}$$

Now solve the above equations

$$m * 5.1 + e - 50 = 0$$

$$m * 5.2 + e - 49 = 0$$

There fore

$$m * 5.2 + e - 49 = m * 5.1 + e - 50$$

$$m * 5.2 - m * 5.1 = e - 50 - e + 49$$

$$m(5.2 - 5.1) = -1$$

$$m * .1 = -1$$

$$m * (1/10) = -1$$

$$\underline{m = -10}$$

By substituting m in any one of the above equation we get value of e

$$m * 5.2 + e - 49 = 0$$

$$-10 * 5.2 + e = 49$$

$$e = 49 + 52$$

$$\underline{e = 101}$$

So we get $m_1 = -10$ and $e_1 = 101$ by using first two observations, so this gives one line in the graph.

Similarly if we solve for other pairs of observations we get $m_2 = 25.55$ $e_2 = -82.44$, so this gives one line in the graph. Similarly if you try to solve the equations for different combinations pairs of the H and W we get different m and e value, so as many m and e value we get those many lines will be obtained in the graph. But which is the best fit line? How to identify it?

	H	w	$w = mH + e$
✓	5.1	50	$m =$
✓	5.2	49	$e =$
	5.3	53	$50 = m \times 5.1 + e \quad \text{--- (I)}$
	6.2	26	$49 = m \times 5.2 + e \quad \text{--- (II)}$
	6.1	24	
	5.9	—	

$$m = -10$$

$$e = 101$$

	H	w	$w = mH + e$
✓	5.1	50	$m_1 = -10 \quad m_2 = 25.55$
✓	5.2	49	$e_1 = 101 \quad e_2 = -82.44$
	5.3	53	$50 = m \times 5.1 + e \quad \text{--- (I)}$
	6.2	26	$49 = m \times 5.2 + e \quad \text{--- (II)}$
	6.1	24	$w = -10 \times 6.1 + 101 \quad \text{--- A ✓}$
Ⓢ	5.9	—	$w = 25.55 \times 5.9 - 82.44 \quad \text{--- B ✓}$
			$w_{1A} = 42$
			$w_{2B} = 68.305$

We are able to find multiple value for m and e hence we will have multiple lines. But which line to accept which we are not able to find out.

How to find out the best fit line: As we know each value of m and e will lead to one line. Now we have calculated m and e values for 3 pairs of data as below.

$m_1 = -10$, $m_2 = 25.55$, $m_3 = 24.61$ and

$e_1 = 101$, $e_2 = -82.44$, $e_3 = -72.4$

To find out the best fit line we need validate these values. For that let us form an equation now for 3 pairs of data again.

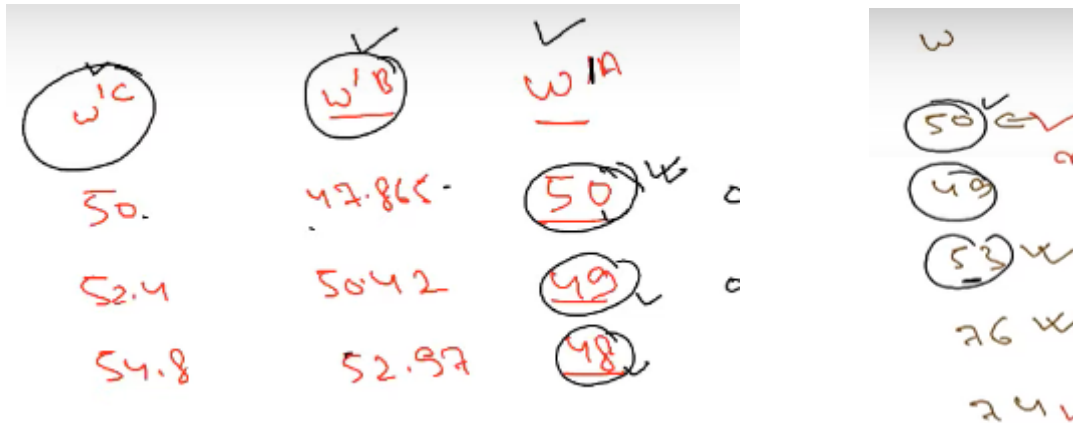
$w = m \cdot H + c$, let us consider now that w is unknown to us now and we will try to find out the value of w (call it as w') by substituting the value of m and c for each value of H (here we will check for 3 values of H).

A $\rightarrow m_1 = -10, c_1 = 101, H = (5.1, 5.2, 5.3)$ for each value of H we need to calculate w'
 i.e. $w'_1 = 5.1 \cdot -10 + 101, w'_2 = 5.2 \cdot -10 + 101, w'_3 = 5.3 \cdot -10 + 101$ gives result as (50, 49, 48)

B $\rightarrow m_2 = 25.55, c_2 = -82.44, H = (5.1, 5.2, 5.3)$
 $w'_1 = 5.1 \cdot 25.55 - 82.44, w'_2 = 5.2 \cdot 25.55 - 82.44, w'_3 = 5.3 \cdot 25.55 - 82.44$ gives result as (47.865, 50.42, 52.97)

C $\rightarrow m_3 = 24.61, c_3 = -72.4, H = (5.1, 5.2, 5.3)$
 $w'_1 = 5.1 \cdot 24.61 - 72.4, w'_2 = 5.2 \cdot 24.61 - 72.4, w'_3 = 5.3 \cdot 24.61 - 72.4$ gives result as (50, 52.4, 54.8)

So by solving above equations we get following w' values



Now in order to find the best fit line we need to compare the results of $w'A, w'B, w'C$ values with corresponding values of actual w we had. The one which gives results most similar to actual w that will become the best fit m and c values. For that we need to find the difference as below

For $w'A$

$$\frac{|w_1 - w'_1| + |w_2 - w'_2| + |w_3 - w'_3|}{3} = \frac{|50 - 50| + |49 - 49| + |53 - 48|}{3} = \frac{5}{3} = 1.667$$

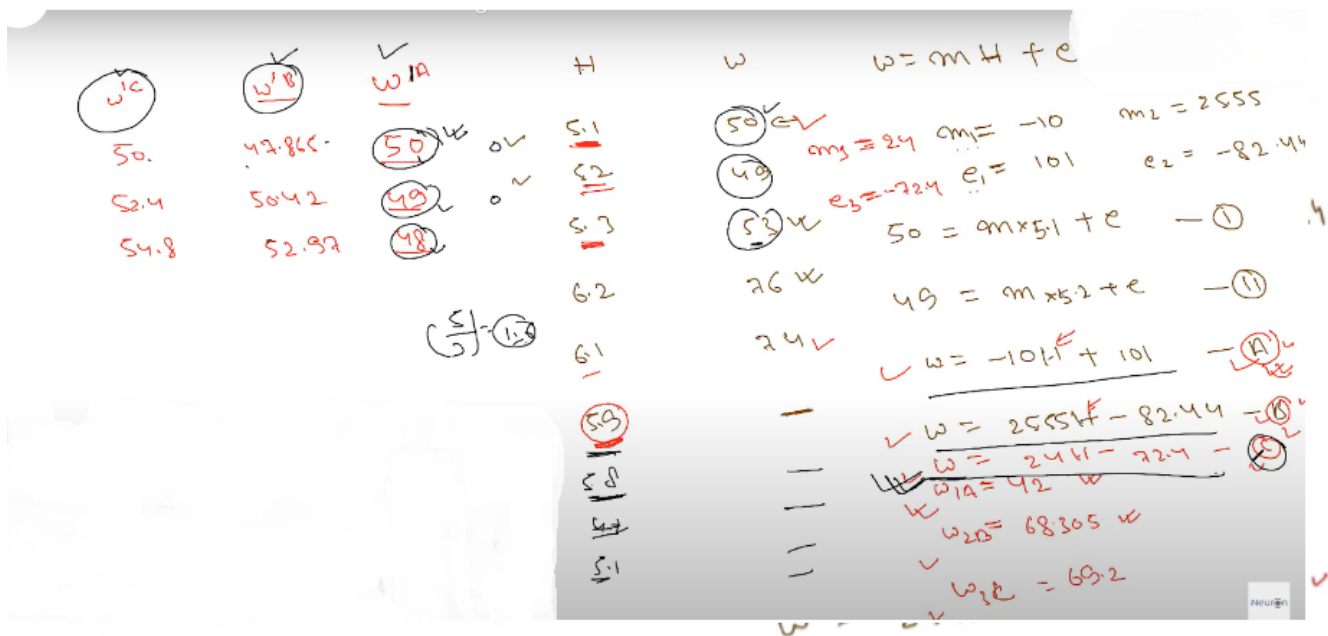
For $w'B$

$$\frac{|w_1 - w'_1| + |w_2 - w'_2| + |w_3 - w'_3|}{3} = \frac{|50 - 47.865| + |49 - 50.42| + |53 - 52.97|}{3} = \frac{2.135 + 1.42 + 0.03}{3} = \frac{3.585}{3} = 1.195$$

For $w'C$

$$\frac{|w_1 - w'_1| + |w_2 - w'_2| + |w_3 - w'_3|}{3} = \frac{|50 - 50| + |49 - 52.4| + |53 - 54.8|}{3} = \frac{0 + 3.4 + 1.8}{3} = \frac{5.2}{3} = 1.733$$

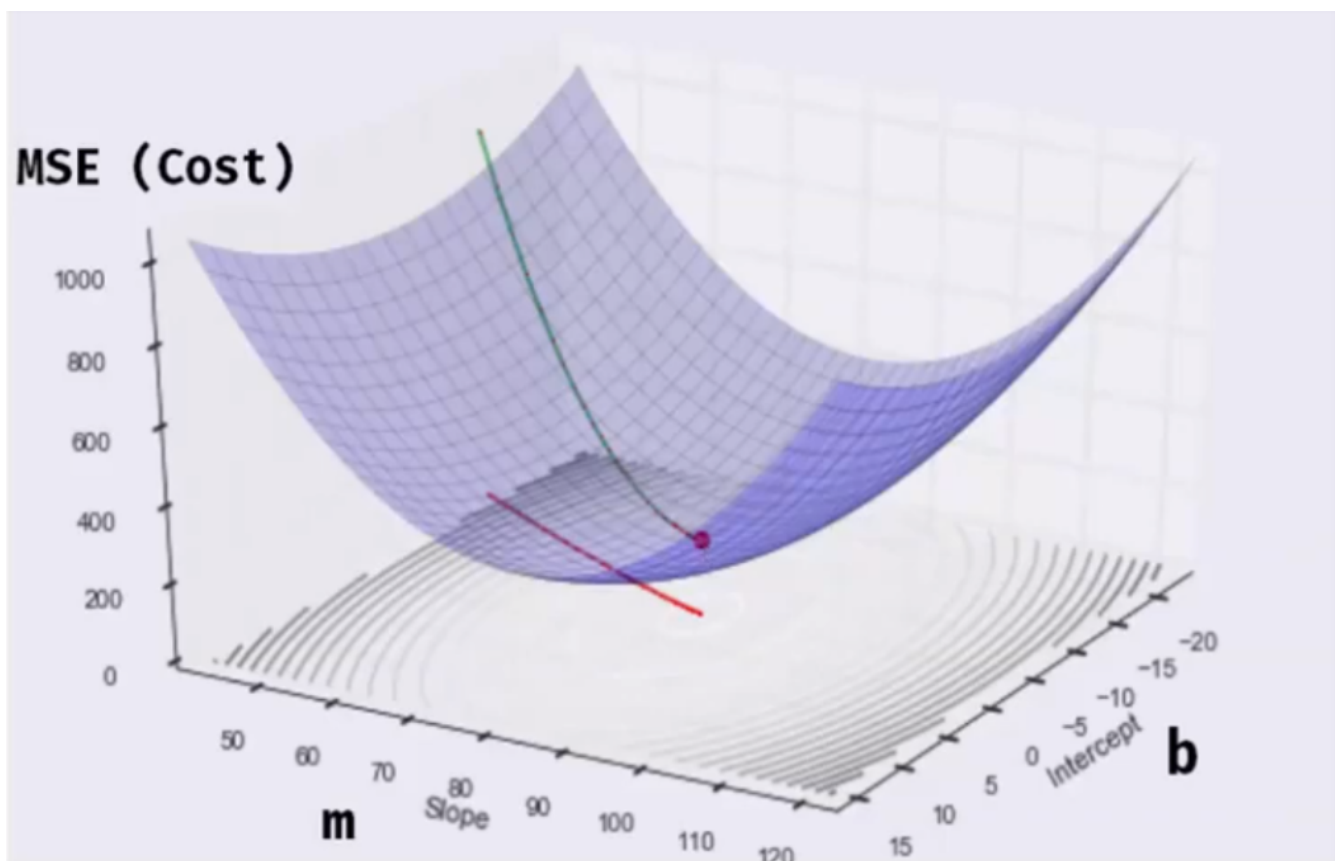
The one which has less difference is $w'B$ one 1.195 so the best fit line is the one which is produced by $m_2 = 25.55, c_2 = -82.44$. The difference ratio we found out by $w - w'/n$ is called as error function or loss function or cost function as well. The lower the loss function or cost function value that will lead to best fit line which can generalise the new data with less error.



So now what is machine learning means: As we now know how to find the best fit line manually. But if we have huge data we will not be able to perform above calculations for each and every data manually. So, how if we make the machine to calculate it, so in ML we are making the machine to perform these calculations automatically for entire given data and find the optimal or best fit m and c value. The way in which the machine can find the optimal value of m and c is **called machine learning**. The whole objective of linear regression algorithm is to find out the value of m and c where we have less error or less loss function or less cost function.

So why there is an error, the error is because of wrong value of m and c . So our objective will be to find out the value of m and c where error is zero. If we are able to find out the position where the error is zero then we will get corresponding right value of m and c .

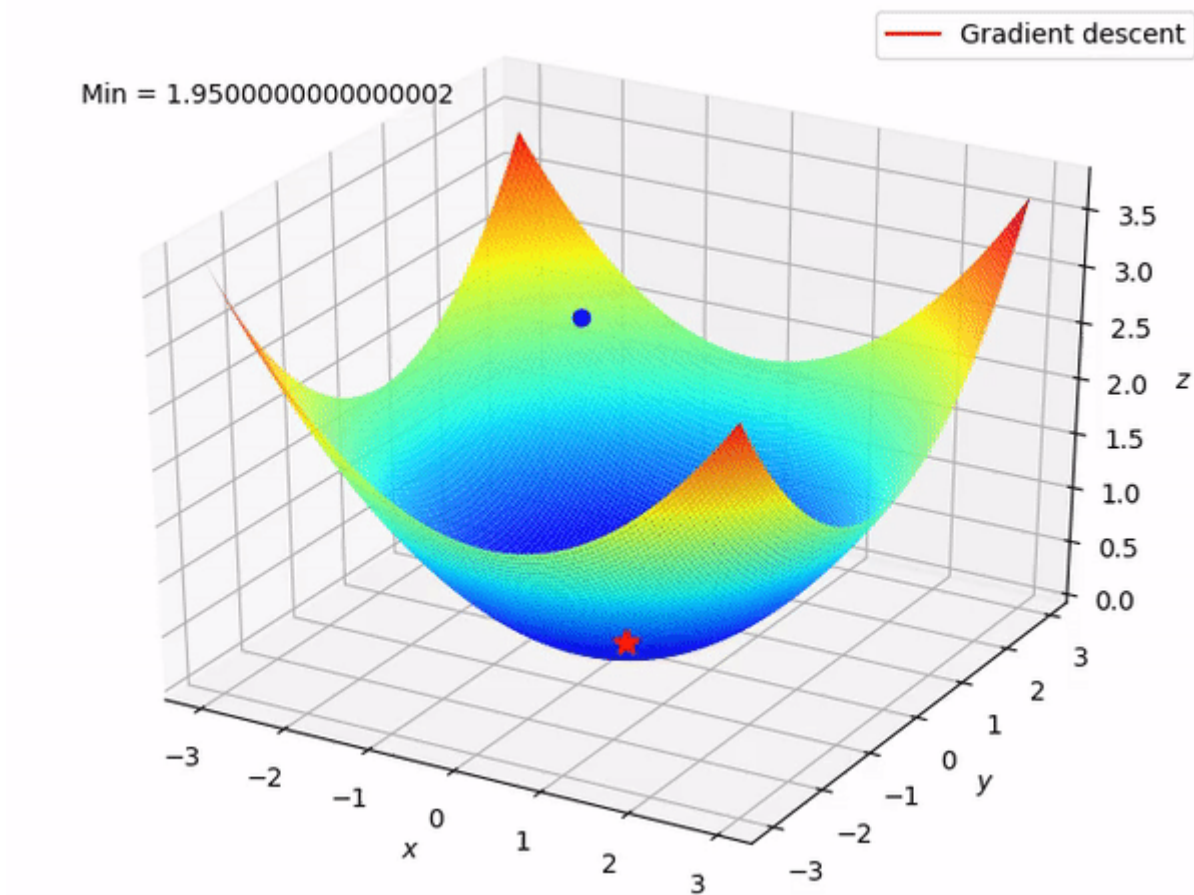
Machine learning means trying to find out the point or position where the error is zero with respect to the introduced parameters inside your equation (here m and c are the introduced parameters in your current equation. These parameters will change as the equation will change for different algorithms or approaches). Finding out such point is a tedious job so we expect the machine to do it.



Let us consider our 3 parameters error, m and c , as we know from the error or cost function or loss function equation error, m and c are related to one another i.e. if there is a change in one parameter there will be change in the other two. Now let us consider these e, m and c in the 3d graph, Let us have m in x axis c in y axis and e error in z axis as shown in the above graph.

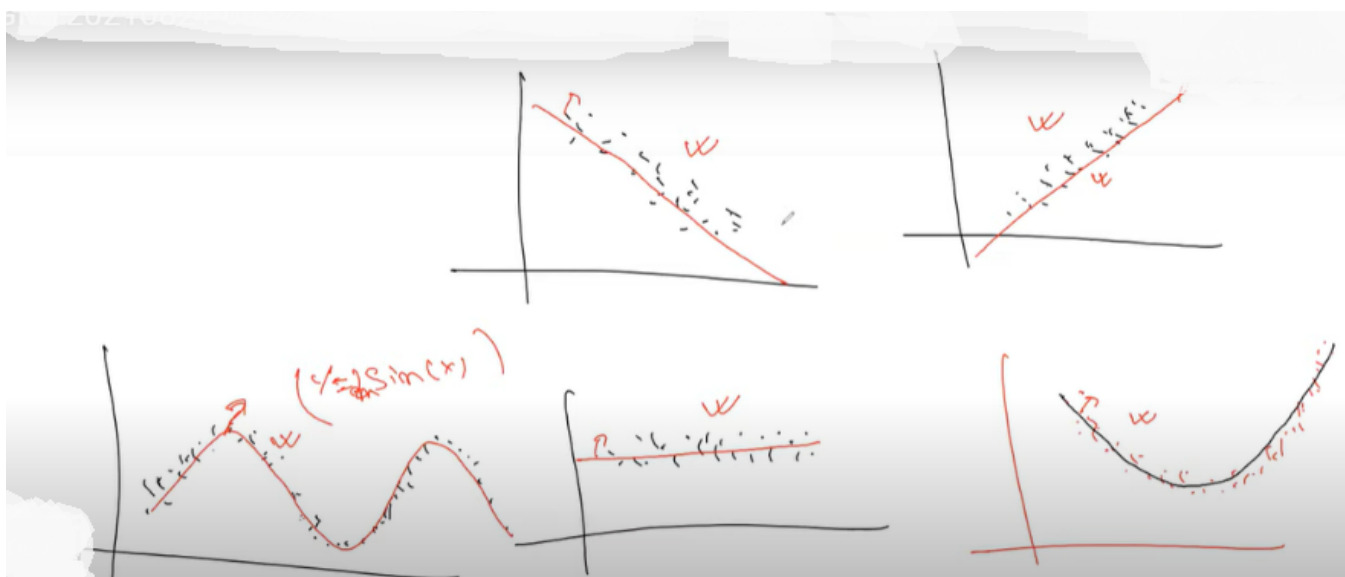
Now we know we found some error value (1.13 for example), if we point this error in the 3d graph, then it will be some where in the middle of the graph. If we move this error point on the graph then corresponding m and c value will also change. If we move the error point such a way that it touches the floor surface, then the value of the error will be zero, and corresponding m and c becomes the best fit values. The process of moving the error point to reach it's floor surface is called as Gradient Descent. The point where it touches the floor surface is called as Global minima.

Note: Here Gradient descent is not only for linear regression equation, it works on the error and the introduced parameters of any mathematical equations. The aim of the Gradient descent is to move the error point to reach at its global minima.

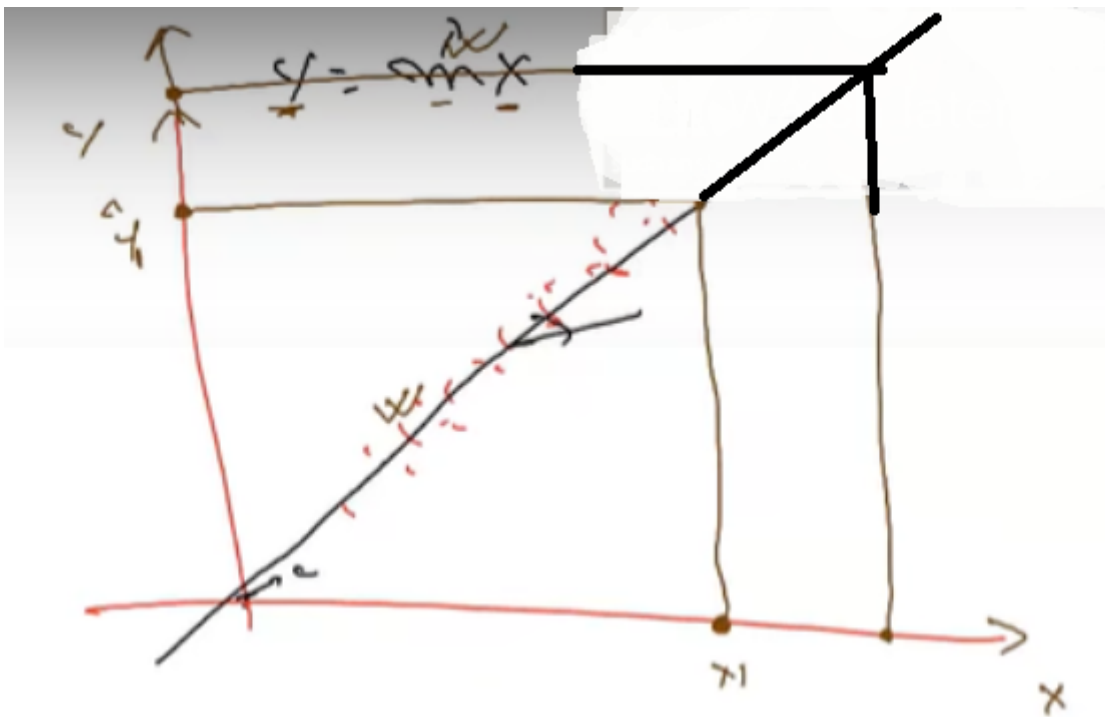
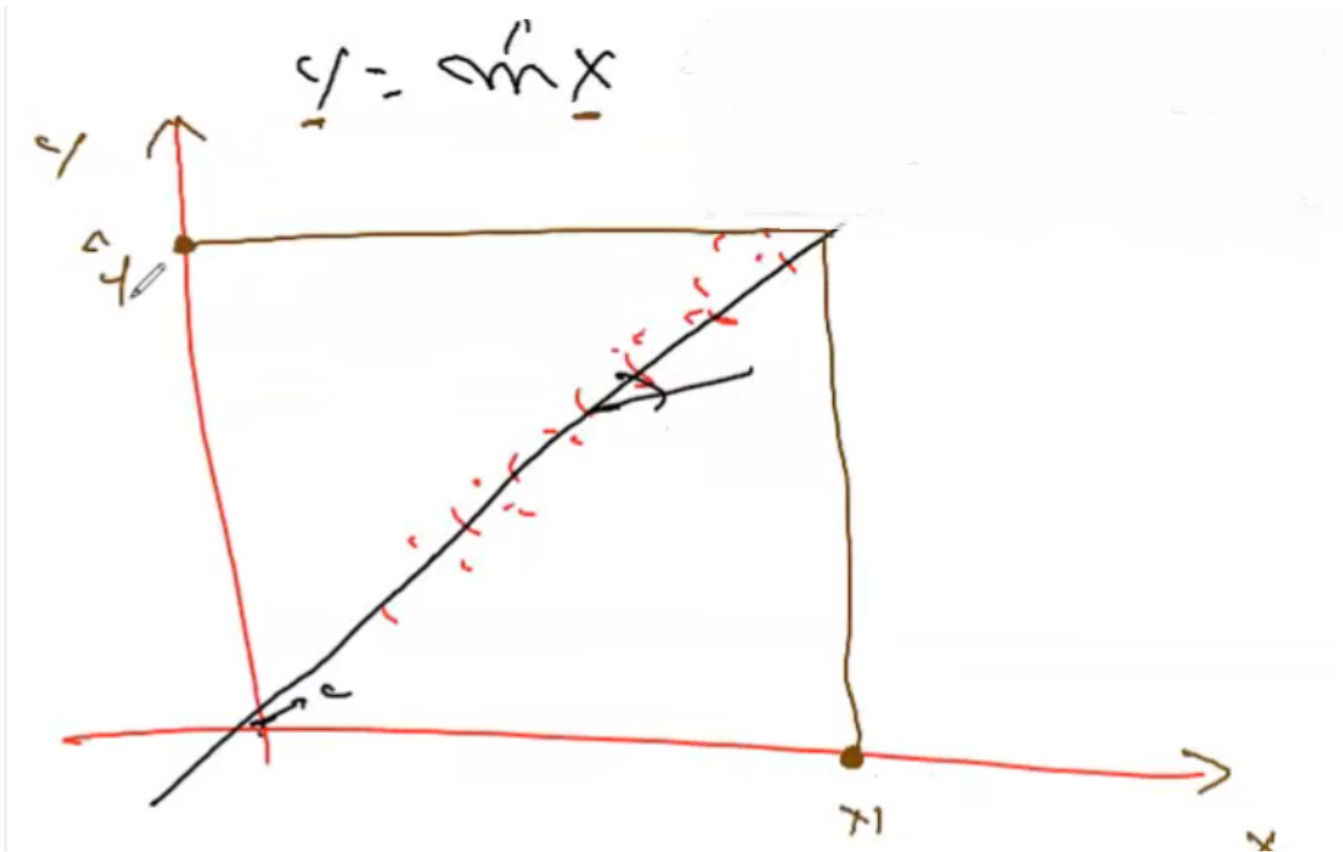


So we can say that we need to find a point where rate of change of error with respect to m (x axis) should be zero as well as rate of change of error with respect to c (y axis) should be zero.
 I.e. $de/dm = 0$ and $de/dc = 0$.

In machine learning we try to find the line which can best fit and able to generalise the data accurately. So for each kind of relation ship between x and y , the data might have been distributed in different different ways. So for all of them there will be a separate equation which defines the line which we can draw to best fit. In each case y formula will be different.



By drawing the best fit line, now we can find out the value of y for a given x as below.
 For example for the given x_1 point when we draw a straight line from x_1 to till the best fit line and from that point if we draw a horizontal line reaching y axis we will be able to find out the y_1 , since this is a predicted y_1 based on our best fit line so we call it as y_1^{\wedge} ,



However we are able to find out y^{\wedge} for a given x point, but now the quest is how confident you are on the predicted y^{\wedge} . What is the accuracy of the predicted value given by our best fit line? Without finding the accuracy you will not be able to find out the confidence score of this best fit line.

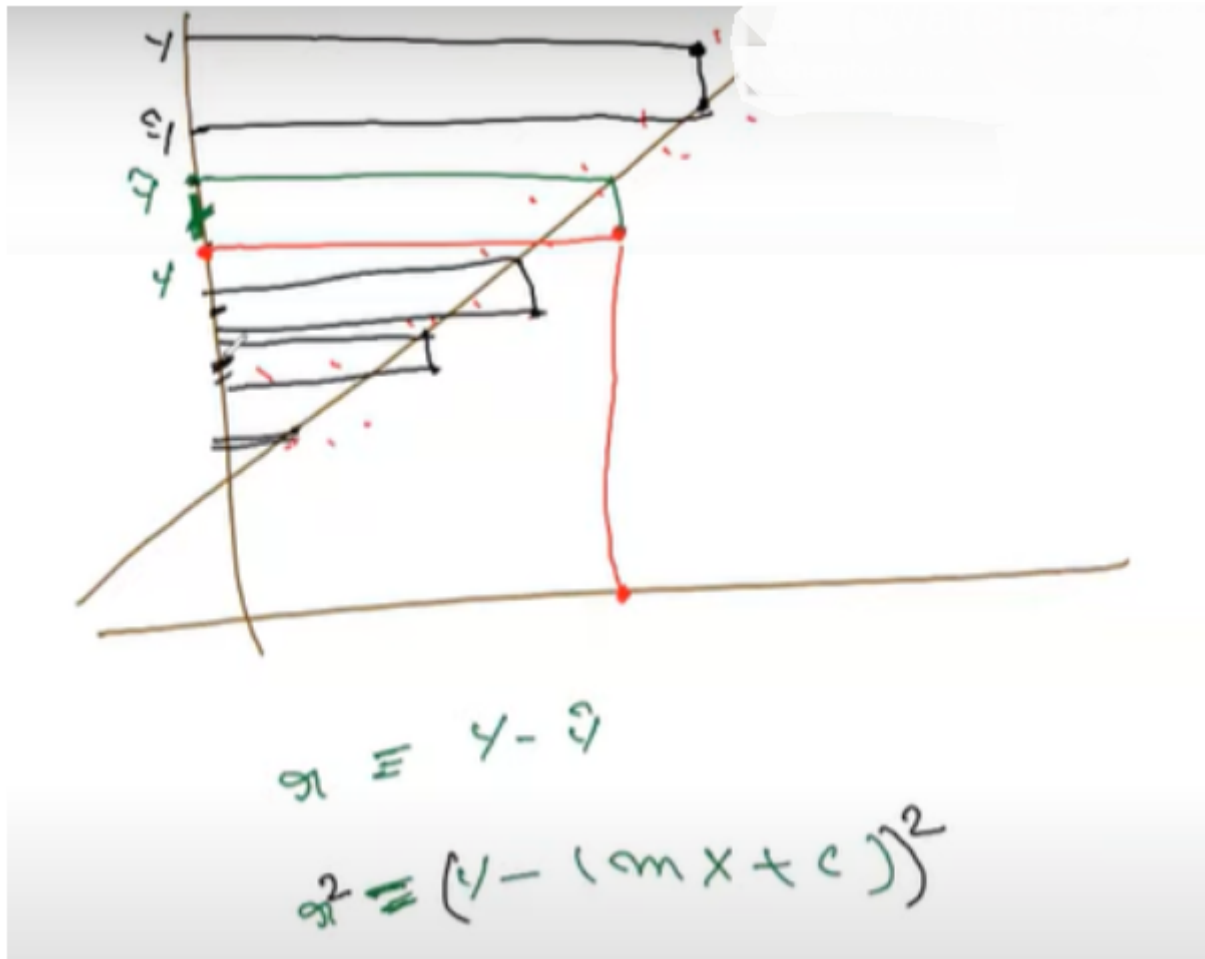
How to find out the accuracy: In order to find out the accuracy we need to find out the residue?

Now we know how to calculate the predicted value manually. Now let us understand how the machine learning calculates the same?

When to apply Linear regression algorithm? Following assumptions must be satisfied by your equation between y and x . Then only we can accept the value of m and c .

- 1) First you need to find out the relationship between x and y and that to you need to observe if there is a linear kind of relation between x and y
- 2) Mean of the residual should be equal to zero
- 3) Error terms are not supposed to be correlated
- 4) Independent variables or x variables and residuals are of uncorrelated
- 5) Error term must show case cost and variance
- 6) There should not be any multicollinearity
- 7) Error terms supposed to be normal distributed

What is Residual?



Residual is the difference between y (actual) and \hat{y} (predicted)

$$r = y - \hat{y} = y - (mx + c)$$

Since sometimes we get r as positive and sometimes as negative value depending on y is greater than or less than \hat{y} . So we square the r .

$$Q^2 = (y - (mx + c))^2$$

Since there are multiple y and y^{\wedge} values we are going to find the summation of residuals as

$$\sum_{i=1}^M Q^2 = \sum_{i=1}^M (y - (mx + c))^2$$

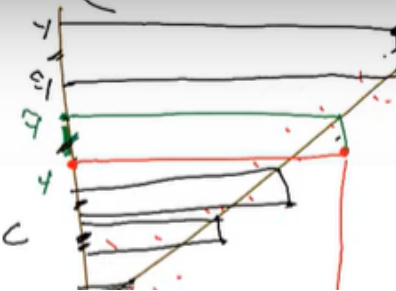
So we try to find out line where the summation of residual is going to be minimum or close to zero.
If we expand the summation equation of residual square.

093221 Recording 1920x1018FS

$$\sum_{i=1}^M Q^2 = \sum_{i=1}^M (y - (mx + c))^2$$

$$(a-b)^2 = (a^2 + b^2 - 2ab)$$

$$= y^2 + (mx + c)^2 - 2y(mx + c)$$

$$Q^2 = y^2 + m^2 x^2 + c^2 + 2mxc - 2yx - 2yc$$


Now residual is nothing but a error we are finding here.

As we know that the best fit line we can find when we find the the point of m and c where error tending to touch surface.

I.e. $de/dm = 0$ and $de/dc = 0$

So here

$dR/dm = 0$, $dR/dc = 0$

So if we do partial derivation

$$Q^2 = y^2 + m^2 x^2 + c^2 + 2mxc - 2yx - 2yc$$

$$dR/dm = 0 + 2mx^2 + 0 + 2xc - 2yx - 0 = 2mx^2 + 2xc - 2yx = 2x(mx + c - y) = 0$$

$$dR/dc = 0 + 0 + 2c + 2mx - 2y = 2(c + mx - y) = 0$$

$$\frac{dR}{dm} = 0 + 2mx^2 + 0 + 2xc - 2yx - 0$$

$$= 2mx^2 + 2xc - 2yx$$

$$= 2x(mx + c - y)$$

$$\frac{dR}{dc} = 2(b + mx - y) \quad - \textcircled{1}$$

$$\frac{dR}{dm} = 2x(mc + c - y) \quad - \textcircled{2}$$

Since we have multiple r_2 so equation becomes summation equation from $i=1$ to M

$$\frac{dR}{dc} = \sum_{i=1}^M 2(b + mx - y) \quad - \textcircled{1}$$

$$\frac{dR}{dm} = \sum_{i=1}^M 2x(mc + c - y) \quad - \textcircled{2}$$

Since we look for a point where dR/dc and dR/dm is zero so we are going to equate above equations to zero

$$\frac{dR}{dc} = \sum_{i=1}^M 2(b + mx - y) = 0 \quad - \textcircled{1}$$

$$\frac{dR}{dm} = \sum_{i=1}^M 2x(mc + c - y) = 0 \quad - \textcircled{2}$$

Here $b=c$ there was typo

If we expand above equations we get below equations

$$\sum_{i=1}^m 2b + \sum_{i=1}^m 2mx_i - \sum_{i=1}^m 2y_i = 0 \quad - (1)$$

$$\sum_{i=1}^m 2x_i mc + \sum_{i=1}^m 2x_i c - \sum_{i=1}^m 2x_i y_i = 0 \quad - (2)$$

Since 2 is common if we divide both the equations by 2 we will get below one

$$\sum_{i=1}^m b + \sum_{i=1}^m mx_i - \sum_{i=1}^m y_i = 0 \quad - (1)$$

$$\sum_{i=1}^m x_i mc + \sum_{i=1}^m x_i c - \sum_{i=1}^m x_i y_i = 0 \quad - (2)$$

Now we need to find the best value of m and c satisfying to reach global minima or error to zero. So how we need to find is we keep on trying with each possible value of m and c to find the best fit line. So in what rate we are suppose to try with different different value of m and c is

$$m_{new} = m_{old} - \eta \frac{1}{m} \left(\sum_{i=1}^m (y_i - \hat{y}_i) \right)$$

$m_{new} = m_{old} - \eta \frac{1}{m} (\text{summation of } i=1 \text{ to } m (y - \hat{y}))$

$$m_{new} = m_{old} - \eta \frac{1}{m} \left(\sum_{i=1}^m (y_i - \hat{y}_i) \right) = (3)$$

$$c_{new} = c_{old} - \eta \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)$$

$$\left(\sum_{i=1}^m (y_i - \hat{y}_i) \right)$$

Since $\left(\sum_{i=1}^m (y_i - \hat{y}_i) \right) = \text{residual } r$

Therefore we can write mnew and cnew as below

$$m_{\text{new}} = m_{\text{old}} - \eta \nabla E_m$$

$$C_{\text{new}} = C_{\text{old}} - \eta \nabla E_C$$

Here eta is used to control the trial rate from mold to mnew, we do not want to move fast with finding the new values based on the change in error. Eta helps to move smoothly to reach global minima. Eta is a constant parameter or hyper parameter it can range from 10 to 0.00001 according to google. 0.001 is the good one.



$$\nabla E_C$$

means a value of C where error is zero

$$\nabla E_C = \frac{dR}{dC} = \left(\sum_{i=1}^m 2(b + mx_i - y_i) \right) = 0$$

Example:

- w
- ① ~~60~~ 60
 - ② ~~62~~ 62
 - ③ ~~65~~ 65
 - ④ 72
 - ⑤ 40

11 c/-

5.1 $m = 10$

5.3 $c = 9$

5.5 $\nabla E_m = x^2 m + xc - xy$

5.5 $\nabla E_c = c + mx - y$

6.1 $\nabla E_m =$

5.1 $\nabla E_c =$

For $w = 65$

$$\nabla E_m = -5.5$$

$$\nabla E_c = -1$$

$$m_{\text{new}} = m_{\text{old}} - \eta \nabla E_m$$

$$c_{\text{new}} = c_{\text{old}} - \eta \nabla E_c$$

$\eta = 0.001$

$$m_{\text{new}} = 10.0055$$

$$c_{\text{new}} = 9.001$$

For $w = 65$ new values are 5.328,

$$\nabla E_m = -5.5, -5.328$$

$$\nabla E_c = -1, -0.9482$$

The learning or adopting a new value of m and c for having error as zero is called as machine learning. The process is called as gradient descent.

In reality the delta will be calculated as summation of residual change with respect to m and c by a machine. Here we have considered for only one value of x .

- 1) First you need to find out the relationship between x and y and that to you need to observe if there is a linear kind of relation between x and y
- 2) Mean of the residual should be equal to zero
Meaning is when you calculating the new m and c the Delta or new residual values should be in a such a way that it cancels out each other, other wise you are not in a right way of approaching the global minima.
- 3) Error terms are not supposed to be correlated with your y. i.e. given an error value; we cannot predict the next error value.
- 4) Independent variables or x variables and residuals are of uncorrelated. generalises that in no way should the error term be predicted given the value of independent variables. This is called as **exogeneity. -IMP INTRV**
- 5) Error term must show case constant variance, i.e there should be change in the delta value you keep on calculating with respect to new m and c values. This is called as **homoscedasticity. IMP-INTRV**
- 6) There should not be any multicollinearity
Means if in case of multi columned data the relation ship between x1 and x2 is called as **multicollinearity**. There should not be such correlations in the data feature you select. You must select a independent X variables. IMP-INTRV. If there is multicollinearity, the precision of prediction by the OLS model decreases.
- 7) Error terms supposed to be normal distributed. I.e. Meaning is when you calculating the new m and c the Delta or new residual values should be in a such a way that it cancels out each other, other wise you are not in a right way of approaching the global minima.

How accurate is my model?

Your model is the learning built on the equation $y = mx+c$ in this case.

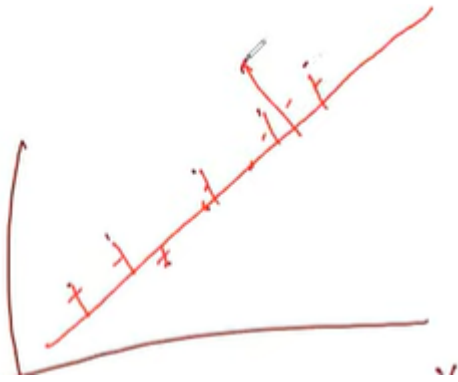
You will be able to define the accuracy of your model by using r square statistics. R2 statistics coming from the residual but here the formula is different.

$$r^2 = 1 - \frac{RSS}{TSS}$$

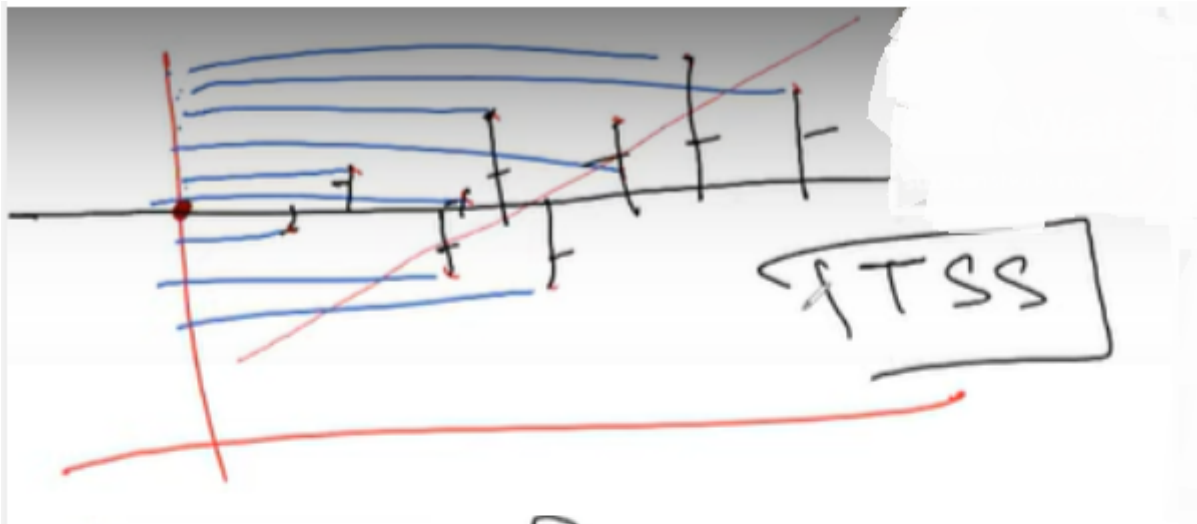
RSS= Residual summation of square = obsolete summation of $y-y^{\wedge}$ i.e. the distance between actual and predicted one and square of it. If Rss is less the R square will be good one. R square range between 0 to 1.

TSS= Total summation of square

Below one is for calculating RSS



Below one is for calculating TSS- First take all the actual y values and take a average of actual y s. Then draw a horizontal line at the point of average y value. Now calculate the distance between each y actual with respect to its average. If you square it this gives the TSS. Your TSS is going to be constant it does not depend on your model.



$$\sum (y_i - \bar{y})^2 = RSS$$

$$\sum (y_i - \bar{y})^2 = TSS$$

Some more Info from canvas
Example for multicollinearity:

Salary= $a \cdot (\text{years of experience}) + b \cdot (\text{age}) + c$ (A typical Linear Regression equation)

Where, a and b are coefficients and c is a constant.



Here, age and years of experience are correlated, since for a person as the years of experience increases, his/her age also increases. Mathematically,

Age=years of experience + d

(where d is a constant, the age when person started the job)

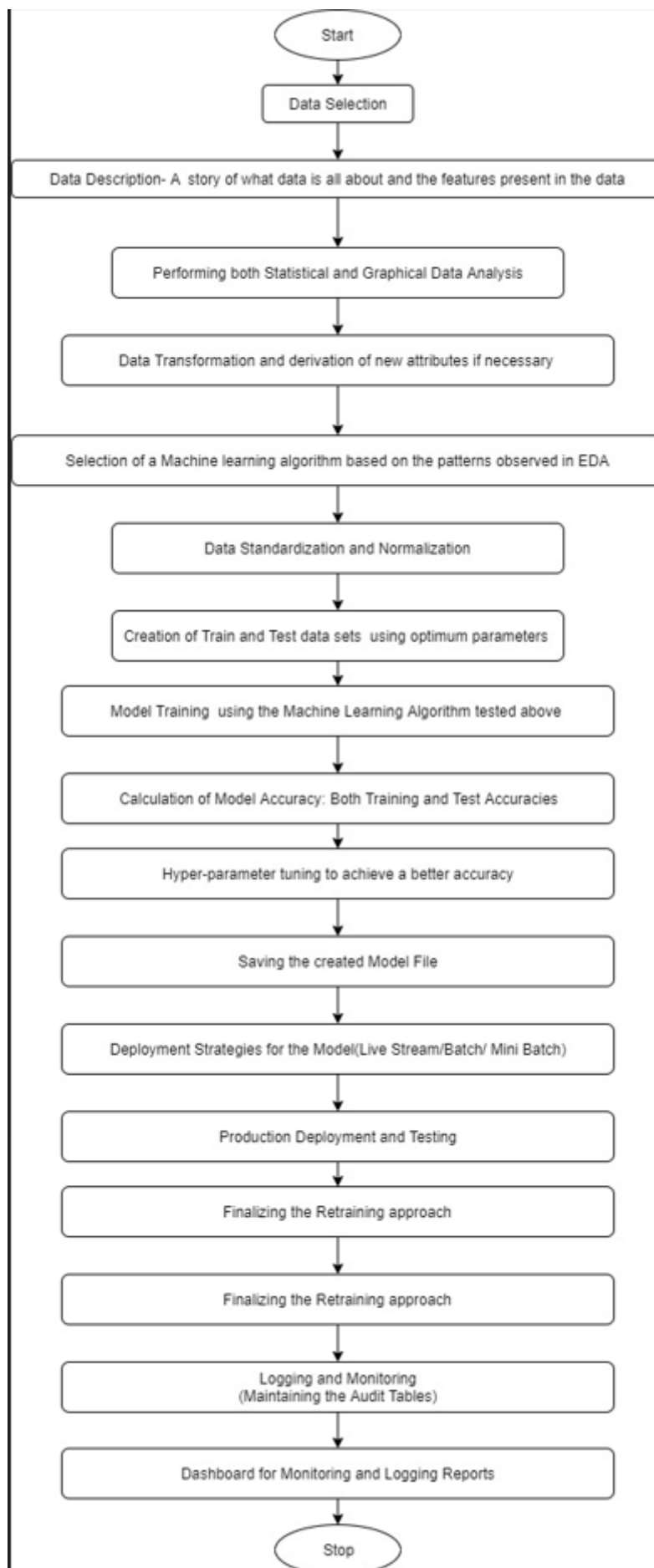


So, our Linear Regression equation which was only supposed to have one linear relation i.e., $Y = mx_1 + nx_2 + c$ now has one more relation $\rightarrow x_2 = ax_1 + d$



More than one Linear relation in an equation, hence the name Multi-Collinearity

ML application flow



R² statistics

The R-squared statistic provides a measure of fit. It takes the form of a proportion—the proportion of variance explained—and so it always takes on a value between 0 and 1. In simple words, it represents how much of our data is being explained by our model. For example,

R

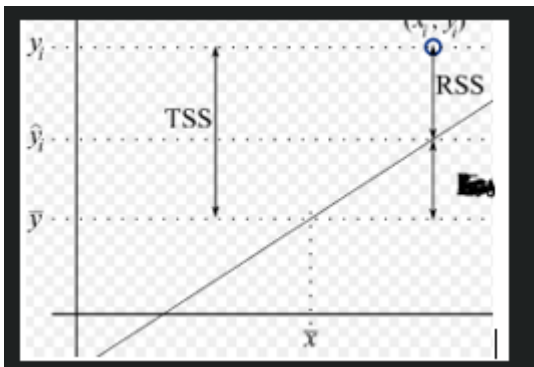
2

R² statistic = 0.75, it says that our model fits 75 % of the total data set. Similarly, if it is 0, it means none of the data points is being explained and a value of 1 represents 100% data explanation.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

$$TSS = \sum (y_i - \bar{y})^2$$



Adjusted R^2 statistics

As we increase the number of independent variables in our equation, the R^2 increases as well. But that doesn't mean that the new independent variables have any correlation with the output variable. In other words, even with the addition of new features in our model, it is not necessary that our model will yield better results but R^2 value will increase. To rectify this problem, we use Adjusted R^2 value which penalises excessive use of such features which do not correlate with the output data. Let's understand this with an example:

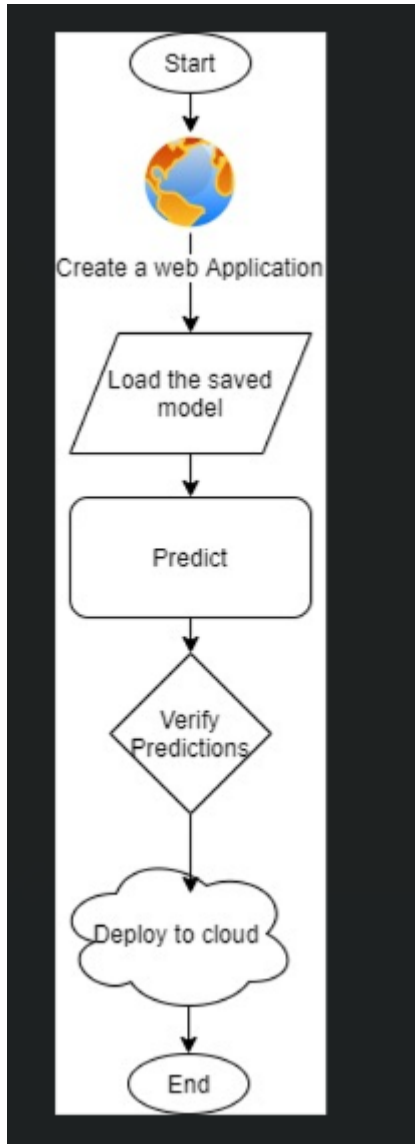
We can see that R^2 always increases with an increase in the number of independent variables. Thus, it doesn't give a better picture and so we need Adjusted R^2 value to keep this in check. Mathematically, it is calculated as:

$$R^2_{\text{adjusted}} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where
 R^2 = sample R-square
 p = Number of predictors
 N = Total sample size.

In the equation above, when $p = 0$, we can see that adjusted R^2 becomes equal to R^2 . Thus, adjusted R^2 will always be less than or equal to R^2 , and it penalises the excess of independent variables which do not affect the dependent variable.

Testing flow



What is Regression Analysis?

Regression in statistics is the process of predicting a Label(or Dependent Variable) based on the features(Independent Variables) at hand. Regression is used for time series modelling and finding the causal effect relationship between the variables and forecasting. For example, the relationship between the stock prices of the company and various factors like customer reputation and company annual performance etc. can be studied using regression.

The use of Regression

Regression analyses the relationship between two or more features. Let's take an example:

Let's suppose we want to make an application which predicts the chances of admission a student to a foreign university. In that case, the

The benefits of using Regression analysis are as follows:

- It shows the significant relationships between the Label (dependent variable) and the features (independent variable).
- It shows the extent of the impact of multiple independent variables on the dependent variable.
- It can also measure these effects even if the variables are on a different scale.

These features enable the data scientists to find the best set of independent variables for predictions.

Linear Regression

Linear Regression is one of the most fundamental and widely known Machine Learning Algorithms which people start with. Building blocks of a Linear Regression Model are:

- Discrete/continuous independent variables
- A best-fit regression line
- Continuous dependent variable. i.e., A Linear Regression model predicts the dependent variable using a regression line based on the independent variables. The equation of the Linear Regression is:

$$Y = a + b \cdot X + e$$

Where, a is the intercept, b is the slope of the line, and e is the error term. The equation above is used to predict the value of the target variable based on the given predictor variable(s).

Simple Linear Regression

Simple Linear regression is a method for predicting a **quantitative response** using a **single feature** ("input variable"). The mathematical equation is:

$y =$

β

0

$+$

β

1

x

Model Confidence

Question: Is linear regression a low bias/high variance model or a high bias/low variance model?

Answer: It's a High bias/low variance model. Even after repeated sampling, the best fit line will stay roughly in the same position (low variance), but the average of the models created after repeated sampling won't do a great job in capturing the perfect relationship (high bias). Low variance is helpful when we don't have less training data!

If the model has calculated a 95% confidence for our model coefficients, it can be interpreted as follows: If the population from which this sample is drawn, is **sampled 100 times**, then approximately **95 (out of 100) of those confidence intervals** shall contain the "true" coefficients.

Multiple Linear Regression

Till now, we have created the model based on only one feature. Now, we'll include multiple features and create a model to see the relationship between those features and the label column. This is called **Multiple Linear Regression**.

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

Each x represents a different feature, and each feature has its own coefficient. In this case:

$$y = \beta_0 + \beta_1 \times TV + \beta_2 \times Radio + \beta_3 \times Newspaper$$

Let's use Statsmodels to estimate these coefficients

Feature Selection

How do I decide **which features have to be included** in a linear model? Here's one idea:

- Try different models, and only keep predictors in the model if they have small p-values.
- Check if the R-squared value goes up when you add new predictors to the model.

What are the **drawbacks** in this approach? -If the underlying assumptions for creating a Linear model(the features being independent) are violated(which usually is the case),p-values and R-squared values are less reliable.

- Using a p-value cutoff of 0.05 means that adding 100 predictors to a model that are **pure noise**, still 5 of them (on average) will be counted as significant.
- R-squared is susceptible to **model overfitting**, and thus there is no guarantee that a model with a high R-squared value will generalise. Following is an example:

Selecting the model with the highest value of R-squared is not a correct approach as the value of R-squared shall always increase whenever a new feature is taken for consideration even if the feature is unrelated to the response.

The alternative is to use **adjusted R-squared** which penalises the model complexity (to control overfitting), but this again generally [under-penalizes complexity](#).

a better approach to feature selection is **Cross-validation**. It provides a more reliable way to choose which of the created models will best **generalise** as it better estimates of out-of-sample error. An advantage is that the cross-validation method can be applied to any machine learning model and the scikit-learn package provides extensive functionality for that.

Multi- Collinearity

Origin of the word: The word multi-collinearity consists of two words:Multi, meaning multiple, and Collinear, meaning being linearly dependent on each other.

For e.g., Let's consider this equation $a + b = 1 \Rightarrow b = 1 - a$

It means that 'b' can be represented in terms of 'a' i.e., if the value of 'a' changes, automatically the value of 'b' will also change. This equation denotes a simple linear relationship among two variables.

Definition: The purpose of executing a Linear Regression is to predict the value of a dependent variable based on certain independent variables.

So, when we perform a Linear Regression, we want our dataset to have variables which are independent i.e., we should not be able to define an independent variable with the help of another independent variable because now in our model we have two variables which can be defined based on a certain set of independent variables which defeats the entire purpose.

- Multi-collinearity is the statistical term to represent this type of a relation amongst the independent variable- when the independent variables are not so independent 😊.
- We can define multi-collinearity as the situation where the independent variables (or the predictors) have strong correlation amongst themselves.

Why Should We Care About Multi-Collinearity?

- The coefficients in a Linear Regression model represent the extent of change in Y when a certain x (amongst X1,X2,X3...) is changed keeping others constant. But, if x1 and x2 are dependent, then this assumption itself is wrong that we are changing one variable keeping others constant as the dependent variable will also be changed. It means that our model itself becomes a bit flawed.
- We have a redundancy in our model as two variables (or more than two) are trying to convey the same information.
- As the extent of the collinearity increases, there is a chance that we might produce an overfitted model. An overfitted model works well with the test data but its accuracy fluctuates when exposed to other data sets.
- Can result in a Dummy Variable Trap.

Remedies for Multicollinearity

- **Do Nothing:** If the Correlation is not that extreme, we can ignore it. If the correlated variables are not used in solving our business question, they can be ignored.
- **Remove One Variable:** Like in dummy variable trap
- **Combine the correlated variables:** Like creating a seniority score based on Age and Years of experience
- Principal Component Analysis