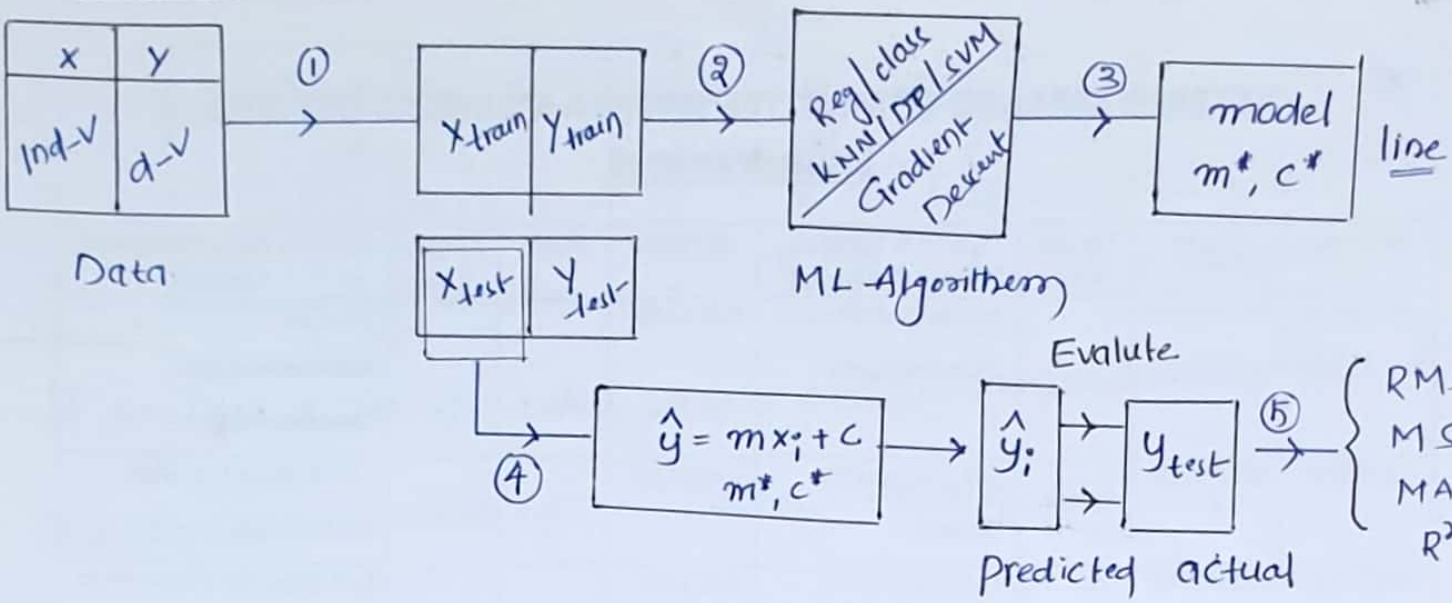
Graphical Representation :-



Simple linear Regression (S.L.R) :-

it is a Statistical method that allows us to Summarize and Study Relationships b/w two Continious variables Simply if we have one Independent variable available in data we called as Simple linear Regression.

Eqn - $y = mx + c$

dep_var ↳ Ind-var

Multiple linear Regression :- it is a Statistical technique that uses several exploratory variables to predict the ~~model~~ Outcome of response variable Simply we can say, if we have more than one Independent variables available in data we called as multiple linear regression.

The goal of MLR is to model the linear relationship b/w Independent variables and dependent variables.

$y = m_1x_1 + m_2x_2 + c$ → slope, → Intercept

$y = m_1x_1 + m_2x_2 + m_3x_3 + m_4x_4 + c$

↳ features (x_1, x_2, x_3, x_4)

In MLR we have some problems that is multicollinearity

x_1	x_2	x_3	x_4	x_5	x_6	y
		2	4			
		4	16			
		5	25			
		7	49			

$X = (x_1, x_2, x_3, x_4, x_5, x_6)$

'X' is Independent variable

'y' is dependent variable

if we have a data set which contain multiple Indipendent variables & One dependent variable (y), if Indepent variables any two variable correlate to each other ie depend on other variable lets take (x_3, x_4) variable, here x_4 is depend on x_3 variable (ie Square of it) then there is facing problem called as "multicolliniority"

In this we have to do check which variablee are dependent and resolve it, make all the variables to Independent, otherwise we are face problem called as "Interpretable model". ie we lost the Interpretability of a model.

Implimantation of Simple linear Regression :-

1. Import libraries
2. load the data set
3. Exploratory Data-Analysic [find the linear Relationship should Exist]
4. split the data into train & test
5. Do Normalization (or) Standardization
6. train the model Using X_{train}, Y_{train} (ie Apply / fit Algorithem)

Intro

(a) while train data we have-to take some assumptions of linear Reg

linear relationship should exist between input & Output

(b) Errors / Residuals on training data should follow Normal distribution

ie mean is zero and varience is anything E of traindata $\sim N(0, \sigma^2)$

(c) Errors / Residuals on training data (X_{train}, Y_{train}) expected to be

Independent of each other

(d) Errors / Residuals on training data expected to be "HOMOSCEDASTICITY"

* Residuals also called as Errors

Note :- linear Regression task can be Solved by two approaches—

(1) Geometric way approach (2) probability way approach

In modern days we ase using Geometric. w. approach to Solve linear reg. task

then we use only assumption (a) if follows or not

Statistitions Generally using probability way of approach-to Solve linear R. task

in this case we have-to check if follows 4 assumption or not

(7) predict on X_{test} - Using the model $(m^*, c^*) \Rightarrow Y\text{-test-pred}$.

(8) Evaluate the model - Y_{test} (vs) $Y\text{-test-pred}$ by using matrices we

Calculate RMSE, MSE, MAE etc.

Note :- Accuracy in only Classification task

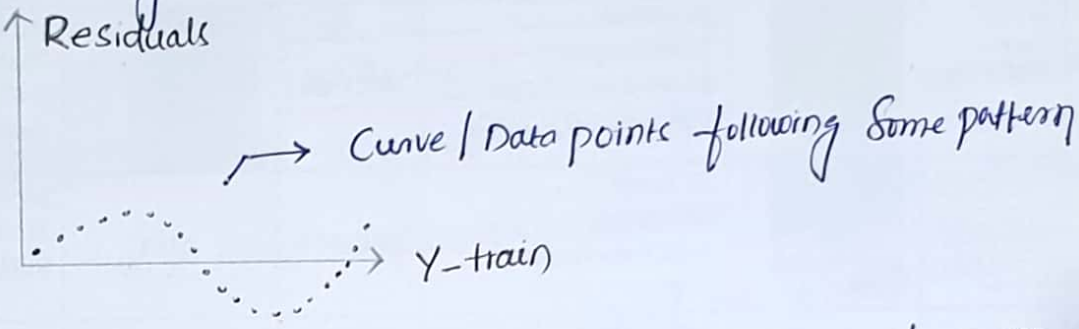
No Accuracy for Regression problems.

Assumption (b, c, d) can also known as Residual Analysis.

Residual Analysis :- it is nothing but assumptions of linear Regression (b,c,d) brief explanation.

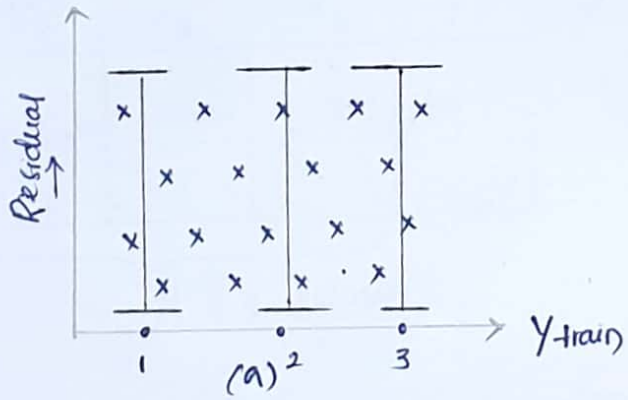
Ass (b) - tells us residual of training data, when plot - follow normal dist ie bell shape Curve with '0' mean.

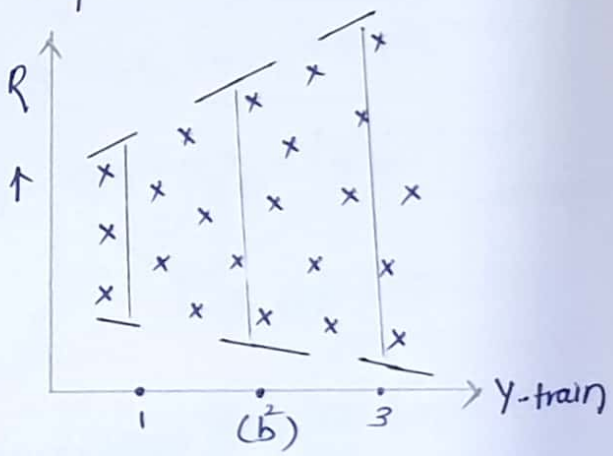
Ass (c) - tells us training data expected to be Indipendent of each other.



By the above fig. shown as the train data is not Indipent when we plot graph on (Residuals - Y train) it follow Some pattern. (dependent)

Ass (d) - tell us residuals on training data expected to be "homoscedasity"





If we observe figure (a) - lets take point '1' on y train the varience or Spread at '1' point Same as '2' point as well '3' point also if we Observe the fig(a) maintain Constant varience / spread at the end. this is called as "HOMOSCEDASTICITY", if the plots are not Constant varience [fig(b) shown] is called as "HETROSCEDASTICITY"

Simple linear Regression :- it is a regression model that estimates the relationshipe between One Indipendent Variable and One dependent-Variable. Using Straight line, both variables Should be Quandtatave.

X	Y
Height	Weight
Ind-var	Dep-var

$\Rightarrow y = mx + c$ → Intercept, slope

→ ∈R

$$w_1 x_1 + w_2 x_2 + w_0 = 0$$

Multiple linear Regression :- it is a statistical technique that estimates the relationship blw two or more exploratory variables (Independent) and aresponsible Variable (dep-var) both variables should be Quandtatave.

Multiple regression is the extension of Ordinary least-Squares (OLS) Regression because it includes more than One exploratory Variable.

X_1	X_2	X_3	Y
Ind-Var	I-Var	I-Var	dep-Var

→ ∈R

X (3-Ind-Var)

$$w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 = 0$$

$y = m_1 x_1 + m_2 x_2 + m_3 x_3 + c = 0$ → Intercept

→ 3-diff Slopes

Note :- ① All the four Residual analysis (Assumptions) hold on Simple linear Regression as well as Multiple linear Regression.

② The Indipendent variables (X) must not be exist- "multicolliniority" if multicolliniority exist then we have to drop One of the column.

We Solve this regression problems by ① Geometric way
② by using Prob & Statistic way.

Recursive feature Elemination (R.F.E) :-

if we have a data set which Consist of multiple Indipendent variables and one dependent variable if there are multiple variables then there high chances to get multicolliniority problem (One Column maintain Some relation with other column or two are more Columns derive other Column).

if there is multicolliniority problem then we loose the better Interprettability. So In Order to get best interpretability we have to Select features that are most Important to our model and remove unwanted features.

In this process there are multiple diffrent ways to Select the features, in this the basic methode is to "Recursive feature elemination" there are different kind of feature elimination methods available.

they are → ① Automatic way → R.F.E

② Manual way → ① forward feature Selection.

② Backward feature Selection.

③ Mixed way approach → In this we do both Automatic and manual way approach.

VIF (Varience inflation factor) :- VIF is a tool to help identify the degree of multicollinearity in a set of multiple Regression variables.

$$VIF = \frac{\text{Overall model Varience}}{\text{Varience of a Single Indipendent variable.}}$$

Assumptions of linear Regression :-

if we have a data that Contain Some i/p variable (X) and o/p variable (Y) and if the o/p variable is Continious ($\in R \rightarrow$ Real Number) Then we have to perform regression task

Procedure to Complete task :-

Step-1 We have to Split the data into trained data, test-data

Step-2 take trained data and passed to machine learning algorithem (KNN, linear Reg., logistic Reg, DT, SVM, etc) and it gives us to a model

Step-3 if we use linear Reg we get model as (m^*, c^*) which is a line in step-2 we have to use GD (Gradient descent) to Optimize model

Step-4 we have to do predictions on test data (X_{test}) that gives Ypred values $(\hat{y}) = mx_i + c$

Step-5 finally we have to evaluate the y-Pred values with Yact values here (Yact = Y-test) in this we use to matrices to Caliculate performance of model, there Some matrices are used for regression task are-

RMSE, MSE, MAE, R2.