MATHS - I : Statistics

- Statistics is the branch of mathematics dealing with the Data Collection, Data Analysis, Interpretation, Data presentation, organizing the Numerical Data.

Statistics = measurement + Analysis.

Statistics

Inferential Statistics

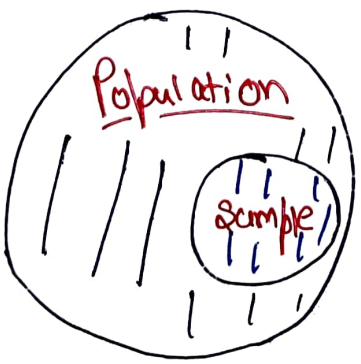
- Many times, a collection of entire data is impossible, Hence a subset of data is collected
- from the subset we get conclusions about the entire population. This is Inferential stats.

Descriptive Statistics.

- These are used to summarize data, such as the mean, standard deviation & etc.
- while applying statistics on data we can find underlying relationships in between the variables.

★ Population:

- Totality of all values or complete list of observations.
- All data points about the subject under study.
  → can be people, vehicles, sales, cats, houses etc.



Taking sample from population using Pandas (Python)

⇒
```
import panda as pd.
df = pd.read_csv ('sales1.csv')
print (df.sample ())
```
sample() is a method. and it always give Random values.

★ **Sample:** A sample is a subset of a population, usually a small portion of the population that is being analysed. It is expensive to perform an analysis on an entire population. So, by analysing sample helps to draw the conclusions about a population.

★ **Variable:** In statistics what we are examining is called variable which is categorized, countable and measurable.

Ex: people have different heights, weights and professions.

→ variable represents the characteristic representation of study.

→ variable is an attribute, has a value, and can store any data type.

⇒ Types of variable

Quantitative DATA

⇒ Data that is measured in Numbers. It deals with numbers that make sense to perform arithmetic calculations

Eg: Height, weight, score.

categorical DATA

Refers to the value that place "things" into different groups or categories.

Eg: Hair colour, Type, letter grade.

① Quantitative variables:

→ Quantitative variables are numeric
→ we can perform arithmetic calculations.
→ Represent a measurable quantity.
Eg: Height, weight etc.

② categorical or Qualitative variables.

→ Stores values which are names or labels.
→ Eg: Type of person: male or female.
Review about food: good, bad, okay.
color of bike: Red, green, blue.

Types of Categorical variable:

Ordinal

(i) Logical order is possible to do analysis.
(ii) Eg: Grades in Exams

Nominal.

(i) There is no logical ordering with respect to actual values.
Eg: Hair colour.

Measures of SPREAD:

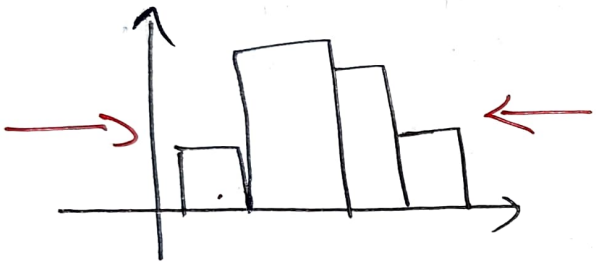
⇒ Types of Quantitative variables:

Discrete variables

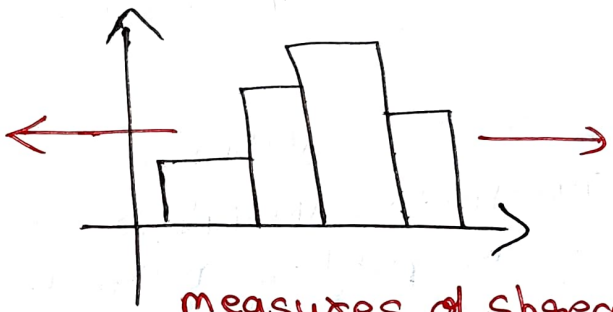
Whole Numbers
- → No. of Pets Owned
- → No. of people in family
- → No. of bikes or cars.

Continous variables

float numbers
- → Weight
- → Salary
- → Bank balance.



measures of Centre
- mode
- median
- mean



measures of spread
- Range
- Standard Deviation.

① **mode** → most frequently observed value.

Eg: 154, 139, 154, 192, 180, 140, 134, 155, 192

mode → 154.

② **median** → value in the middle of an ordered dataset.

⇒ $\frac{n+1}{2}$.

③ **mean** → Average

$$\text{mean} = \frac{\Sigma x_i}{n}$$

$\frac{\text{summation of all values}}{\text{Total number of values.}}$

$$\text{Sample mean} = \bar{x} = \frac{\Sigma x_i}{n}$$

Measures of SPREAD:

★ RANGE:

Range means difference in between minimum values and maximum value. It explains about the data is in between min and max values.

★ Standard Deviation:

The Standard Deviation is a measure of how spread out numbers.

⇒ Square root of variance.

$$s = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n-1}}$$

x_i = observations
$\bar{x}$ = sample mean
n = total no. of observations.

STD tells us how close the values in a Data set are to the mean.

★ Variance:

→ Average of squared differences from the mean.
→ Variance is the average of squared differences from the mean.
→ How far the data points are in a population from the population mean.

$$s^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1}$$

Outliers

A value which is very far from common Data.
→ can be largest value or the smallest.

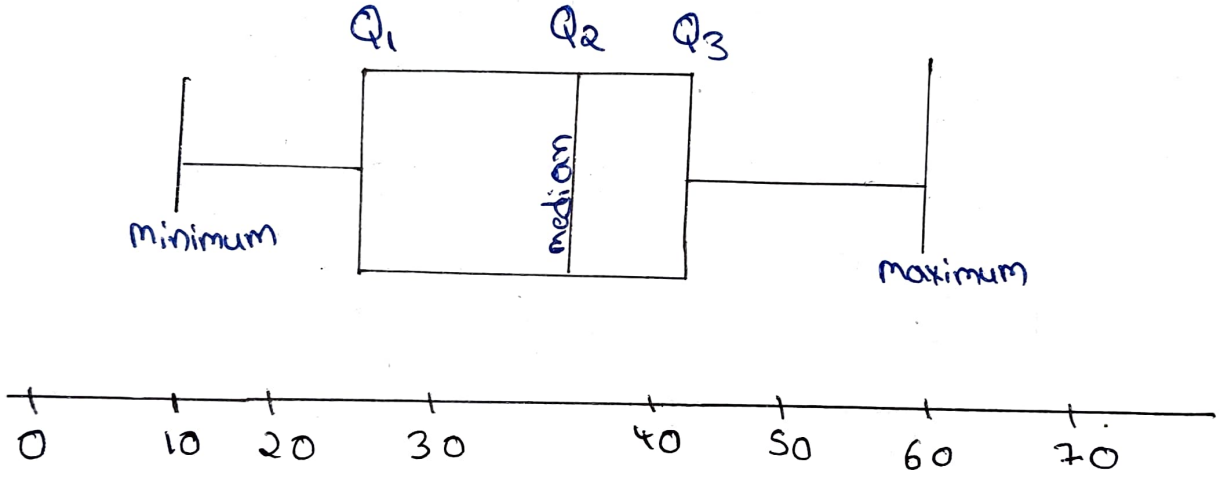
The five Numbers

1, 2, 3, 4, ---- 25 ---- 50 ---- 98, 99, 100.

minimum — 1st Quartile — median — 3rd Quartile — maximum

25%. 25%.

✡ Box-Plot:



✡ IQR: Inter quartile Range:

$IQR = Q_3 - Q_1$.

① Box plot shows the maxi, mini, medium, first Q and third Quartile of the DataSet.

② gives the overall statistical information about data distribution.

③ Good for detecting the Outliers.

Eg:

```
import matplotlib.pyplot as plt
data = [10, 20, 99, 30, 40, 50]
plt.boxplot(data)
plt.show()
```

* A Data value is considered to be an outlier if:

Data value $< Q_1 - 1.5(IQR)$ — Small Outlier

OR

Data value $> Q_3 + 1.5(IQR)$ — Large Outlier.

Eg:

Q_1	Q_2	Q_3
10 25	33	36 59

10, 11, 12, 25, 25, 27, 31, 33, 34, 34, 35, 36, 50, 59

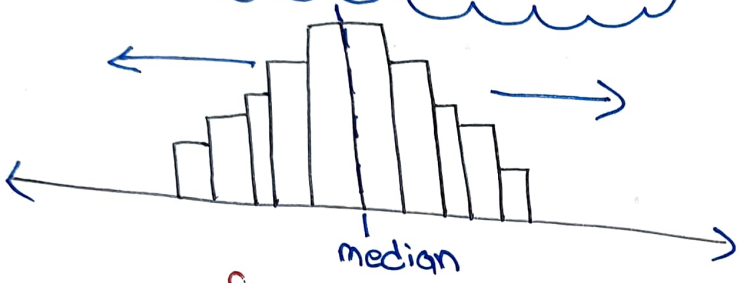
① $IQR = Q_3 - Q_1 = 36 - 25 = 11$

② Lower whisker $= Q_1 - 1.5(IQR)$
$= 25 - 1.5(11)$
$= 8.5$

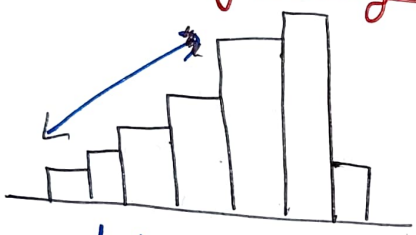
Upper whisker $= Q_3 + 1.5(IQR)$
$= 36 + 1.5(11)$
$= 52.5$

∴ Any value less than 8.5 & Largen than 52.5 is an outlier

Symmetry and Skewness:

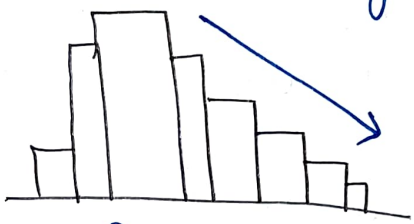


Symmetry — Same shape on both sides of the median.



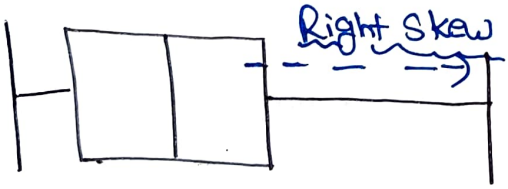
Left skewed

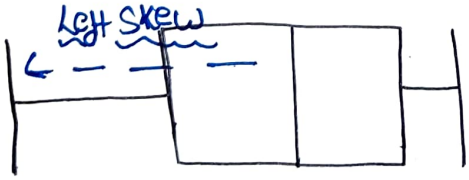
Long tail that trails towards left

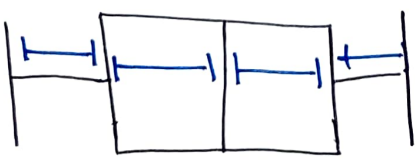


Right skewed

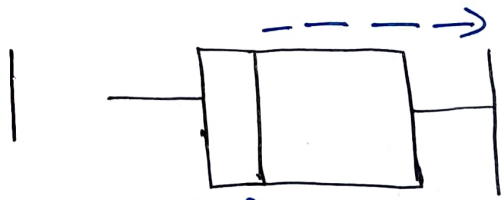
Long tail that trails toward Right

① 

② 

③ 

Symmetric



∴ Right skewed box plot

→ If Right side of the box is large so Right skew box plot.

→ If box is equal then, go for the large whiskers.

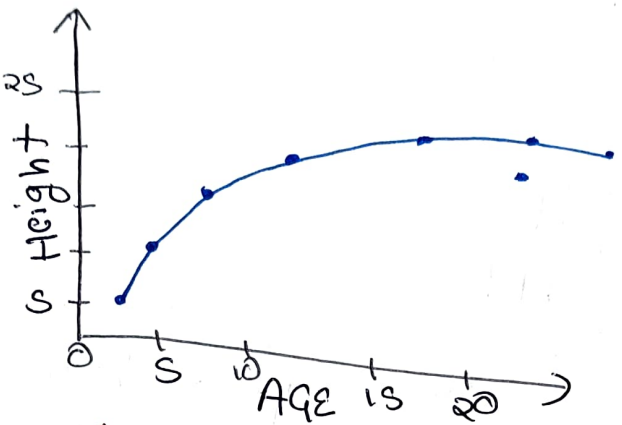
VARIABLES:

(i) If

(

Responsible variable

(measures the Out come of study)



Explanatory variable

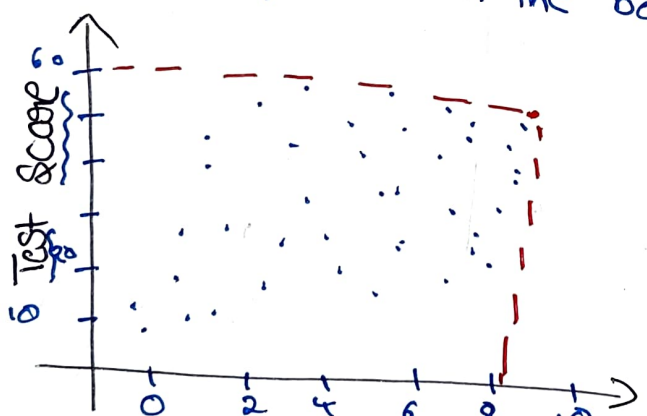
(Explains the outcome of study)

Scatter Plot:

: rate of growth / loss on the basis of one variable.

Scatter plot is a graph which explains

Eg:

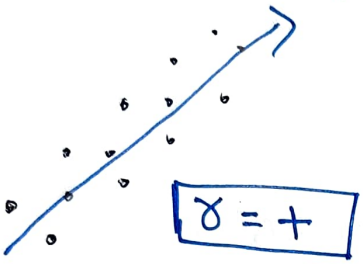


⇒ If two variables have any relation then we can find this relationship by using Scatter plot.

Correlation:

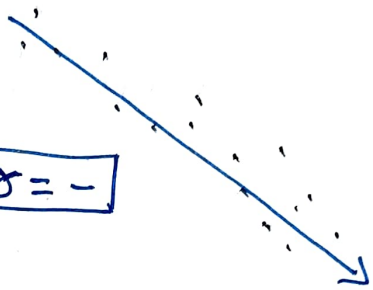
Explains about Direction and Strength of Relationship between two variables.

(i) If upward than Correlation is positive



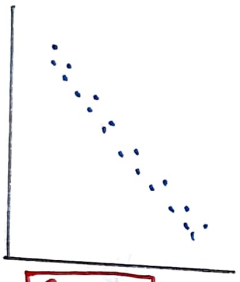
$r = +$

(2) Downwards than Correlation is negative.



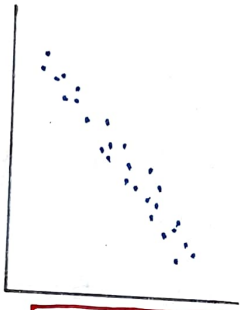
$r = -$

Direction & Strength

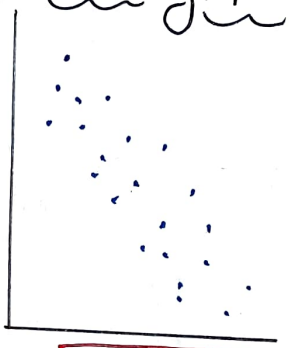


$r = -1$

Perfectly Negative



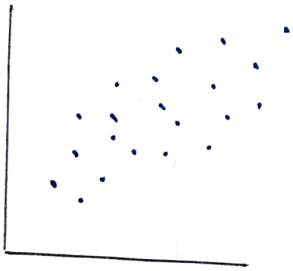
$r = -0.7$



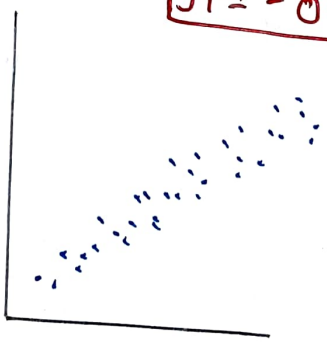
$r = -0.4$

No relation

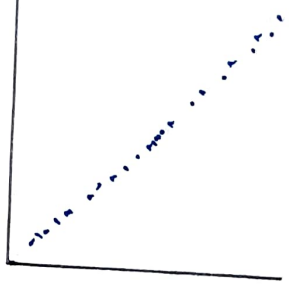
$r = 0$



$r = 0.4$



$r = 0.7$



$r = 1$

Perfectly Linear Positive Relation.

$$r = \frac{1}{(n-1)s_x s_y} \sum (x_i - \bar{x})(y_i - \bar{y})$$

Understanding Correlation:

$$r = \frac{1}{(n-1)S_x S_y} \Sigma (x_i - \bar{x})(y_i - \bar{y})$$

- $n \rightarrow$ Number of observations.
- $\Sigma \rightarrow$ Standard Deviation (for x and y)

Eg: A teacher wants to Determine the Corr between the numb of Hours Spent Studying and test scores.

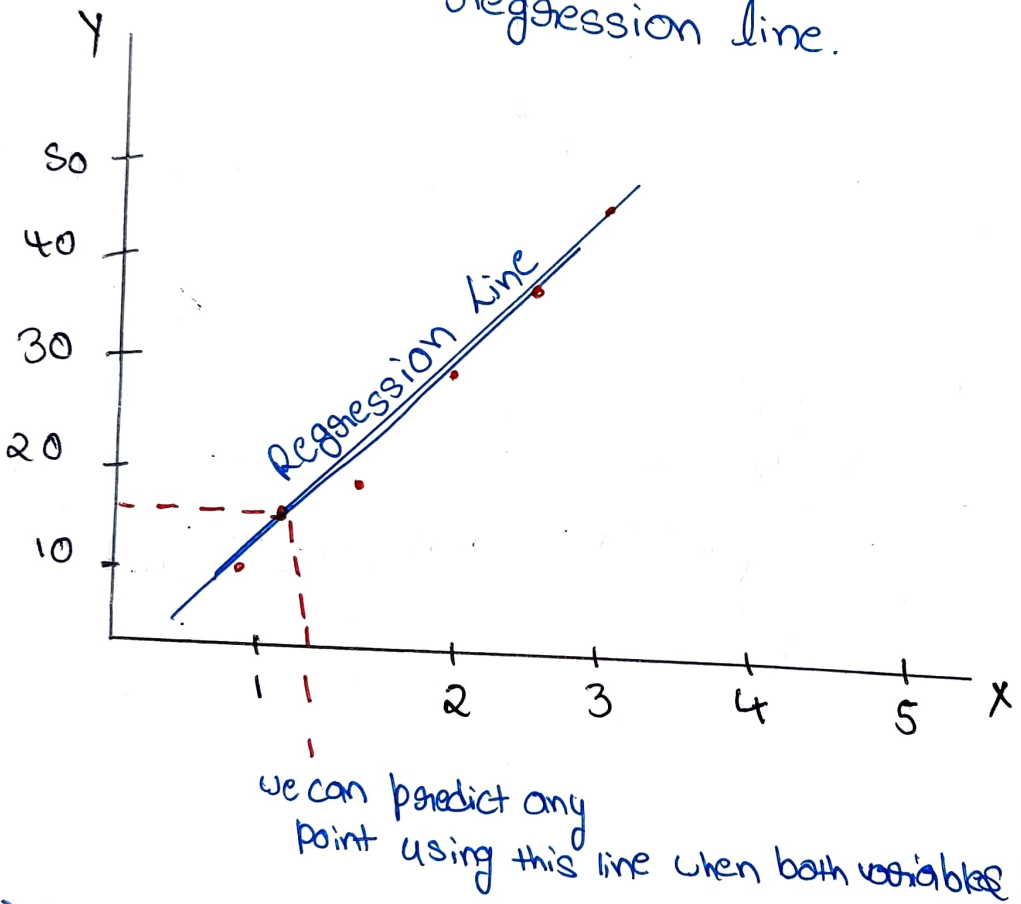
Student	Time spent x_i	Score (100) y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
John	13	53	13 - 12.5 = 0.5	53 - 68 = -15	(0.5)(-15) = -7.5
Allie	15	69	15 - 12.5 = 2.5	1	2.5
mark	7	92	-5.5	24	-132
Samantha	3	10	-9.5	-58	551
Jessica	10	85	-2.5	17	-42.5
Joseph	27	99	14.5	31	449.5
	$\bar{x} = 12.5$	$\bar{y} = 68$			Sum = 921
	$S_x = 8.28$	$S_y = 32.91$			

$$r = \frac{1}{(6-1)(8.28)(32.91)} (921) \Rightarrow \boxed{r = 0.602}$$

$\Rightarrow$ The correlation is +0.6, its upward Direction.

Regression

- Regression explains about how we can draw line in between the points
- The line represents the pattern of the data.
- Line is called regression line.



→ Regression line predicts the change in Y when X increases by one unit.

→ Here change in Y can be increase / decrease

Regression line:

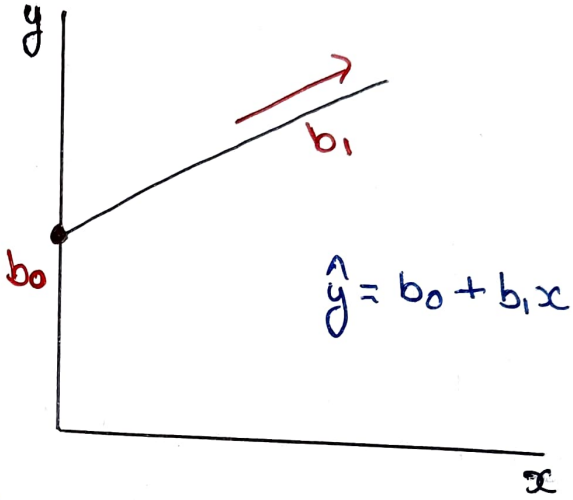
$$\hat{y} = b_0 + b_1 x$$

- $\hat{y}$: Predicted value of y
- b_0 : Y intercept
- b_1 : Slope
- x : Any value of x.

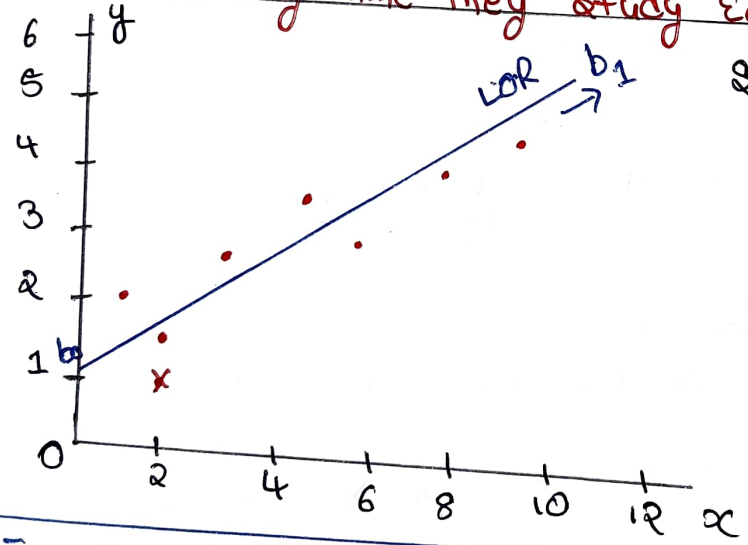
$\hat{y} = b_0 + b_1 x$

$b_0 = \bar{y} - b_1 \bar{x}$

$b_1 = r \frac{s_y}{s_x}$



Q. Suppose A Researcher wants to Predict A Student's GPA from the Amount of time they Study Each week.



Study Time(x_i)	GPA(y_i)
1	2.0
2	1.5
3	2.5
5	3.5
6	3.0
8	4.0
10	4.5

$\therefore \bar{x} = 5$, $\bar{y} = 3$, $s_x = 3.26$, $s_y = 1.08$, $r = 0.94$

we know, $\hat{y} = b_0 + b_1 x$

$b_0 = \bar{y} - b_1 x$

$b_1 = r \frac{s_y}{s_x}$

$b_0 = 1.45$

$b_1 = 0.311$

$\therefore \boxed{\hat{y} = 1.45 + 0.311x}$

So, As study time increases by one unit, we predict a student's GPA to increase by 0.311

R (r)

→ Has values between -1 & 1

→ measures the Linear relation -ship between two quantitative variables with respect ~~and~~ to direction and strength.

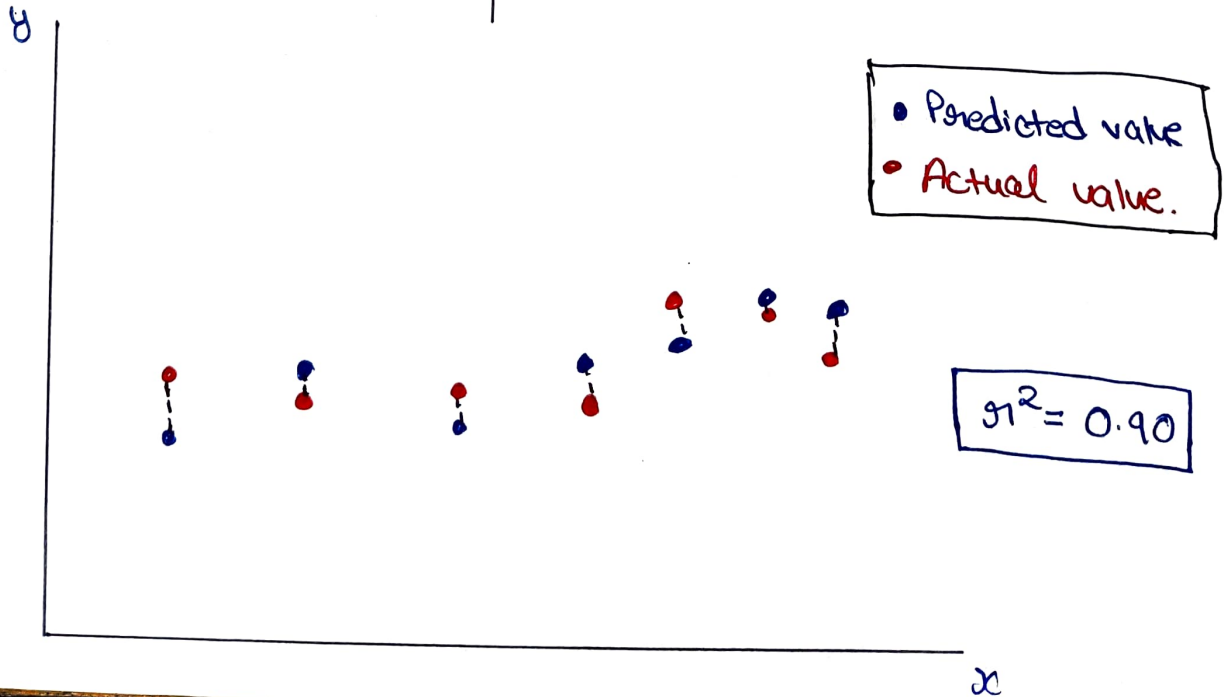
R^2 (r^2)

→ Has values between 0 & 1

→ Is a measure of how close each data points fits to the Regression line.

→ Tells us How well the regression line predicts Actual values.

Eg:

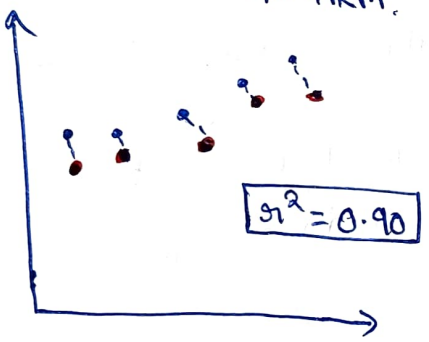


Case - 1

R squared value is high

Explains the actual values and predicted values are close together.

→ Actual & predicted values having very less distance b/w them.

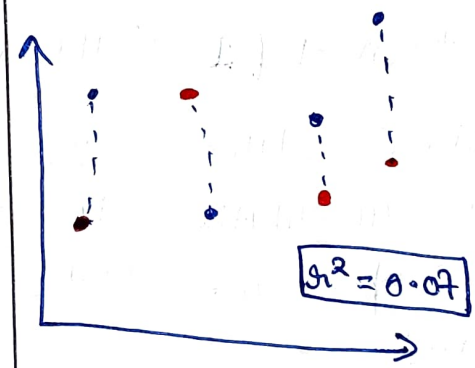


Case - 2

R squared value is low

Explains the regression line which the data points are does not fit well.

→ values having large distance b/w them.

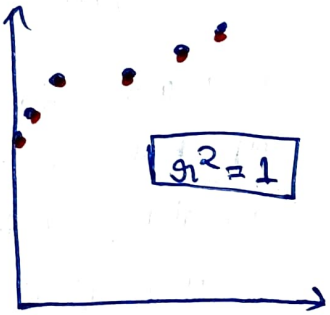


Case - 3

R squared = 1

Perfect prediction

→ means we can predict value of y for any given value of x.



NOTE: When prediction (score value, threshold value, score) >90+, than its a good model | prediction.

& score <55 or so is a bad model.

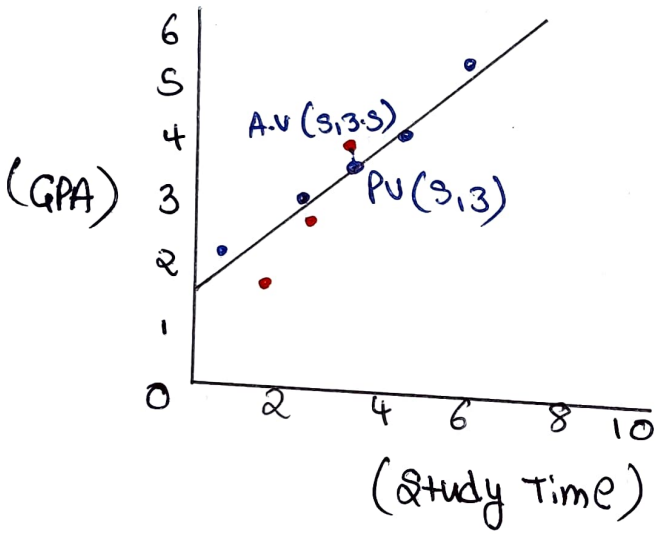
⇒ Residual:

How far the Predicted value is from the actual value, tells us the error in prediction.

Residual = Actual value of Y − Predicted value of y

$$R = y_i - \hat{y}$$

Eg:



residual = $3.5 - \hat{y}$

$\hat{y} = 1.45 + 0.311x$

$= 2.072$

① residual = $3.5 - 2.072$

$= 1.5$ (Approx) (positive)

② residual for Another value $= 1.5 - 2.072$

$= -0.572$

(negative).

⇒ Residual positive ⟶ Actual value is above & prediction is below.

⇒ Residual negative ⟶ prediction is high & A.V. is low.

NOTE ⟶ In maths it's called Residual but in M.L it is called Cost function (Loss/Error).

MATRIX:

Scalar

24

Vector

row $\begin{bmatrix} 2 & -8 & 7 \end{bmatrix}$

column $\begin{bmatrix} 2 \\ -8 \\ 7 \end{bmatrix}$

Matrix

$$\begin{bmatrix} 6 & 4 & 24 \\ 1 & -9 & 8 \end{bmatrix}$$

rows x columns.