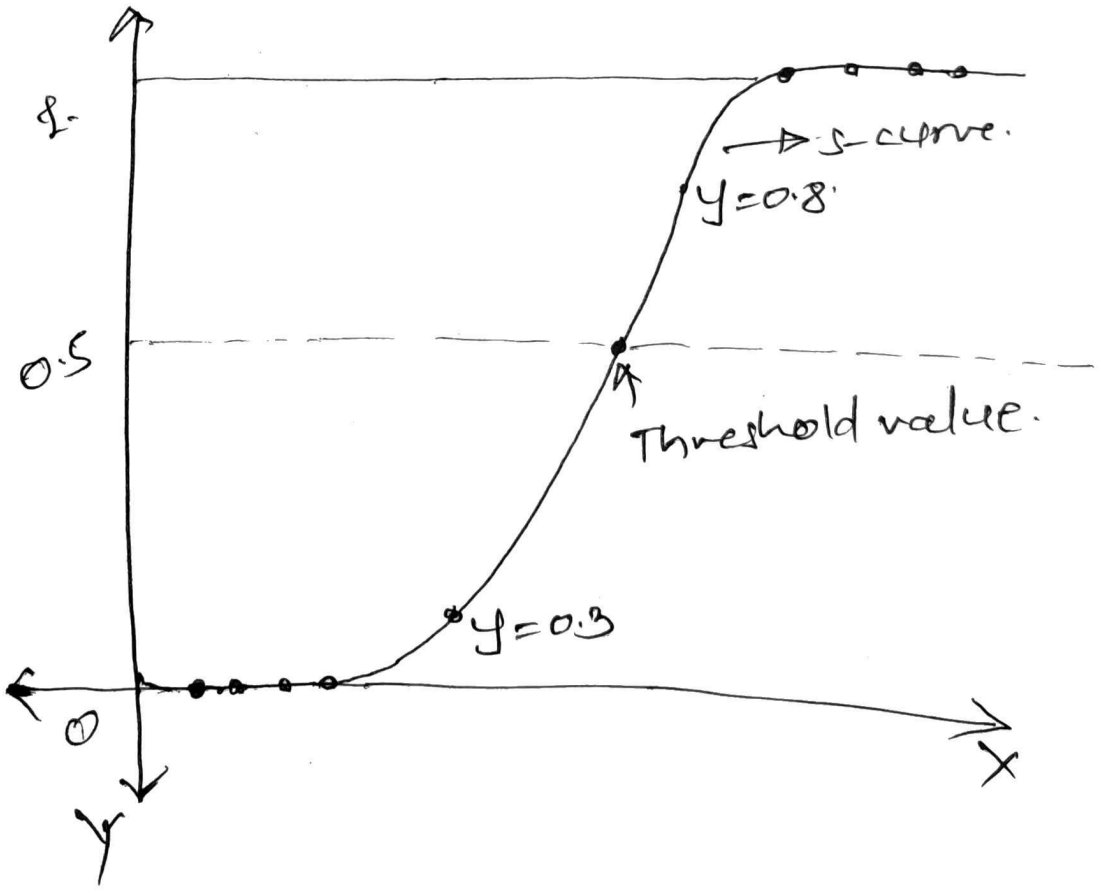
* Logistic Regression.

- Supervised learning technique.
- Used for predicting categorical dependent variables using a given set of independent variables.
- Predicts o/p of categorical dependent variable.
- It is used for solving classification problem.
- It gives probabilistic values which lie between 0 & 1
- We can fit 's' shaped logistic fu which predicts max values (0 & 1)



* Logistic fu^{n}: (sigmoid fu^{n}):

- used to map predicted values to probabilities
- maps real value into another value within a range 0 & 1.
- value of logistic regression must be bet^{n} 0 & 1 which cannot go beyond the limit
- so it forms a curve of 's' shape. & this curve is also called as sigmoid fu^{n}. or logistic fu^{n}
- threshold value is probability of either 0 or 1. values above threshold value tends to 1 & below threshold values tends to 0.

* Assumptions for logistic Regression:

- Dependent variables must be categorical in nature.
- independent variables should not have multi collinearity

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$$

Types: Binomial, multinomial, ordinal.

$$\frac{1}{1+e^{-x}}$$

Appl^n:

credit scoring., Hotel booking, text editing, Medicine, Gaming

Adv & disadv: It constructs linear boundaries

- easier to setup & train - underfitting.
- When outcomes are linearly seperable then it is most efficient
- should not be used if no. of observations are larger than no. of features
- nonlinear problems cannot be solved.

Parameters	Correlation	Regression
	statistical measure that determines the association or co relationship bet two variables.	It describes how to numerically relate an independent variable to dependent variables.
Usage.	Represent linear relationship bet 2 variables.	To fit best line & to estimate one variable based on another
Dependent & independent var.	No difference.	Both are different
Indicates	extent to which two variables move together.	Indicates impact of change & unit on the estimated variable in ref variable (X).
objective	To find a numerical value expressing relationship bet variables.	To estimate values of random variables on the basis of values of fixed variables

1) SST (sum of squares of total):

$$SST = \Sigma (y_i - \bar{y})^2.$$

sum of squared differences bet^n individual data pts (y_i) & mean of response variable ($\bar{y}$).

2) SSR (sum of squares of regression):
sum of squared differences bet^n predicted data pts ($\hat{y}_i$) & the mean of response variable ($\bar{y}$).

$$SSR \ (\Sigma \hat{y}_i - \bar{y})^2.$$

3) sum of squares error (SSE): sum of squared differences bet^n predicted data pts ($\hat{y}_i$) & observed data pts (y_i)

$$SSE = \Sigma (\hat{y}_i - y_i)^2.$$

$$SST = SSR + SSE$$

R squared :- Referred to as coefficient of determination, measure of how well a linear regression model fits a dataset.

- It represents prop. of variance in the response variable that can be explained by predictor var.
- It ranges from 0 to 1.

0 - response var. cannot be explained by predictor var. at all

1 - —11— perfectly —11—

$$R \text{ squared} = \frac{SSR}{SST}$$

e.g. $SSR = 137.5$

$SST = 156$

$R = ?$

Adjusted R squared :-

- It adjusts statistic based on the no. of independent variables in model.
- It tells how well curve/line fits data.

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1} \right]$$

- It adjusts terms in no of line/curve.

n - no of pts in sample.

k - no of independent regressors. no. of variables

- Use need R^2 when working with samples.
- Always less than or equal to R^2
- not necessary when you have population data
  eg. A fund has a sample R squared value close to 0.5 & it is doubtedly offering higher risk adjusted returns with sample size of 50 & 5 predict find adjusted R square value.

  $n = 50$; $k = 5$. R-square = 0.5.

- If you add useless variables adj. R^2 will decrease & if you add useful var. it will increase.