REGRESSION

- Regression is study of dependence.
  - It attempts to find relationship/dependence between:
    - A response (dependent) variable Y
    - One or more (independent/indicator/predictor) variables: $X_1, X_2 \ldots X_p$

Categories of Regression:

- If there is only 1 predictor, it is known as Simple Linear Regression.
- If there are more than 1 predictor, it is known as Multiple/Multivariate Regression.

- Given some past data of association between X and Y:
  - Y should be continuous variable
  - X can be continuous or categorical variable

LINEAR REGRESSION

- Linear Regression is a statistical modelling technique that falls under supervised machine learning techniques.
- Linear regression technique is used to predict/forecast various things such as:
  - Predict or forecast sales based on change in product price.
  - Predict crop yield based on rainfall amount.
  - Assessing credit limits for new customer.
  - Predict the onset of a disease for a new patient.
  - Predict TV serial viewership in future.
  - What is the effect of one more year of education on the income of the person?

REAL LIFE EVENTS

YOUR RS 10-LAKH PERSONAL LOAN DISBURSED WITHIN 10 SECONDS

Citi's 'Instant Personal Loan', which is entirely algorithm - driven, accounts for 20% of its monthly personal loan bookings.

Loan will be given to the person based on their credit worthiness.

LINEAR REGRESSION

- This technique concerns with finding and defining a relationship between two variables (dependent variable (DV) and independent variable (IV)) through a linear equation of the form:
(assuming that these variables are linearly related)

$$Y_i = b_0 + b_1 X_i$$

where,

- X_i is the i^{th} datapoint/observation
- Y_i is value of dependent variable for i^{th} observation

- In order to represent the concept graphically, let's draw the Scatter plot diagram between X and Y:

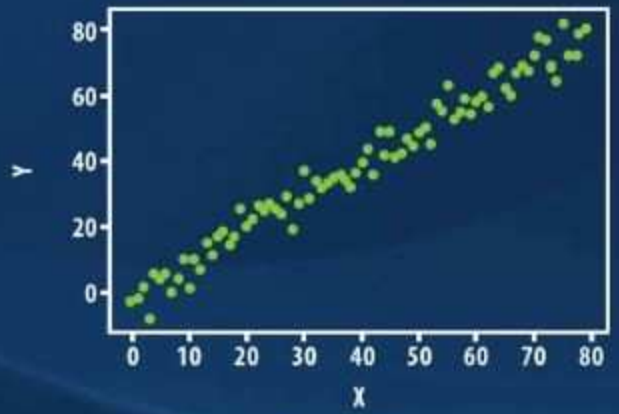


Figure: 3

LINEAR REGRESSION

- **The purpose of linear regression is to find the best fit line that closely represents given dataset.**

e.g. If we regress Y on X, we should get a line such as this:

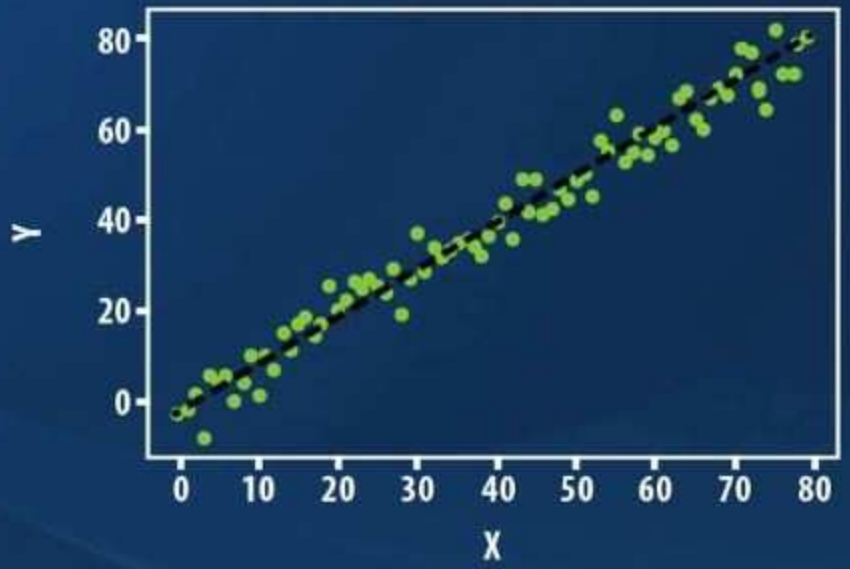
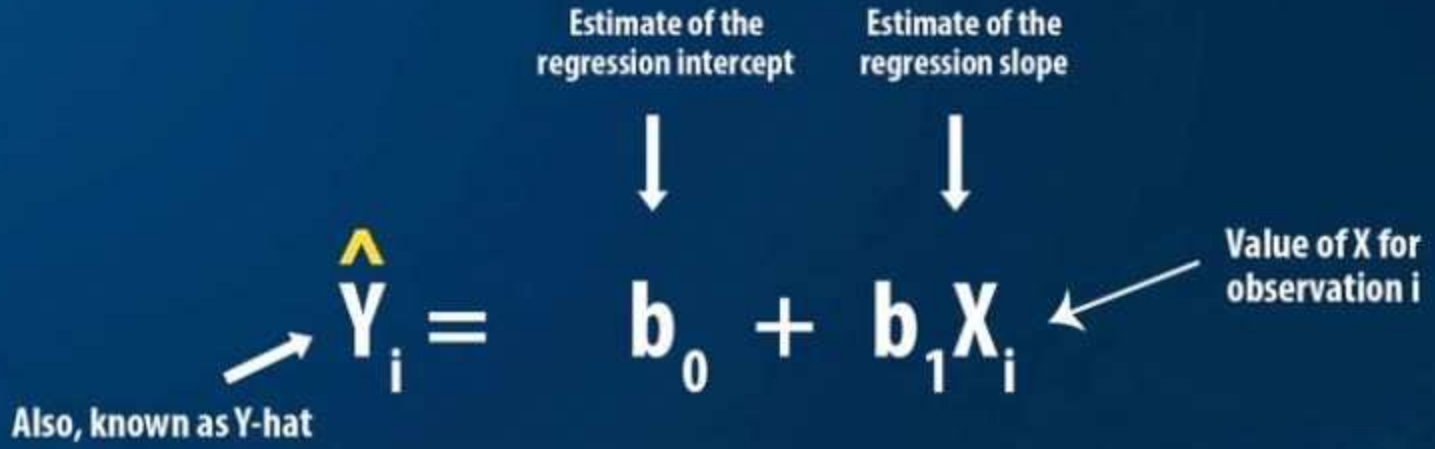


Figure: 4

LINEAR REGRESSION

- So, linear Regression equation provides an *estimate* of regression line (or best fit line).



- b_1 effectively tells how one unit of change in X impacts or develops how much change in Y.

Note:The individual random error terms e_i have a mean of zero.

LINEAR REGRESSION

- Parameters of the line i.e b_0 and b_1 are estimated by minimizing the sum of squared errors/residuals (ordinary least squares) between Y (real) and y-hat (estimated) values as represented below:

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

- Y_i are true values at X_i (i.e. past data)
- b_0 and b_1 are also known as coefficients

LINEAR REGRESSION – WAY TO FIND BEST FIT LINE

Objective: Minimize SSE for each datapoint

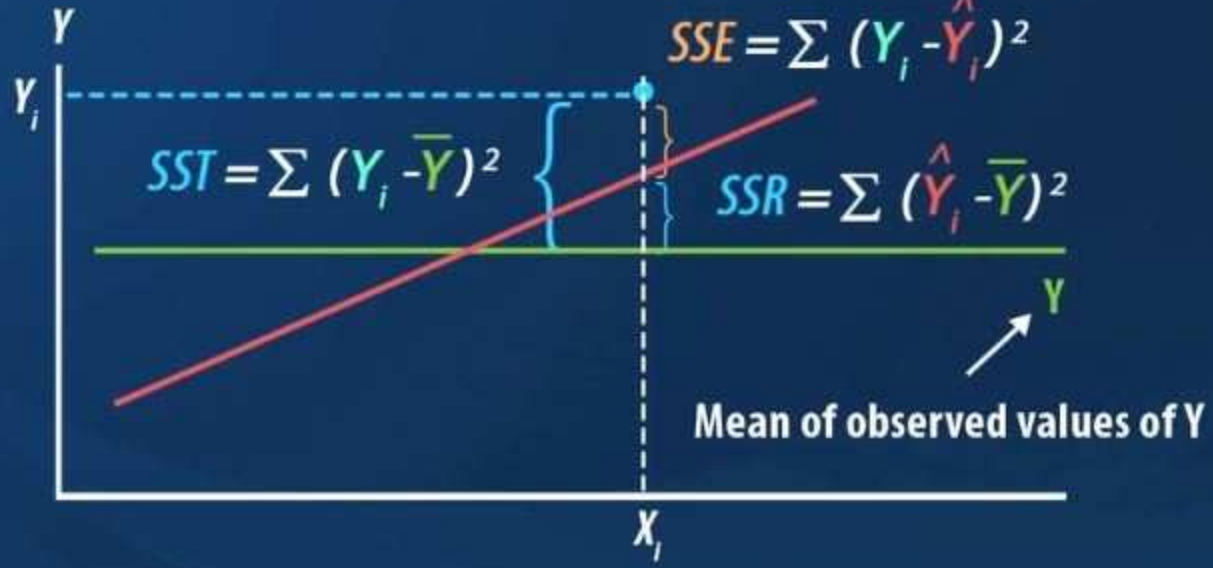
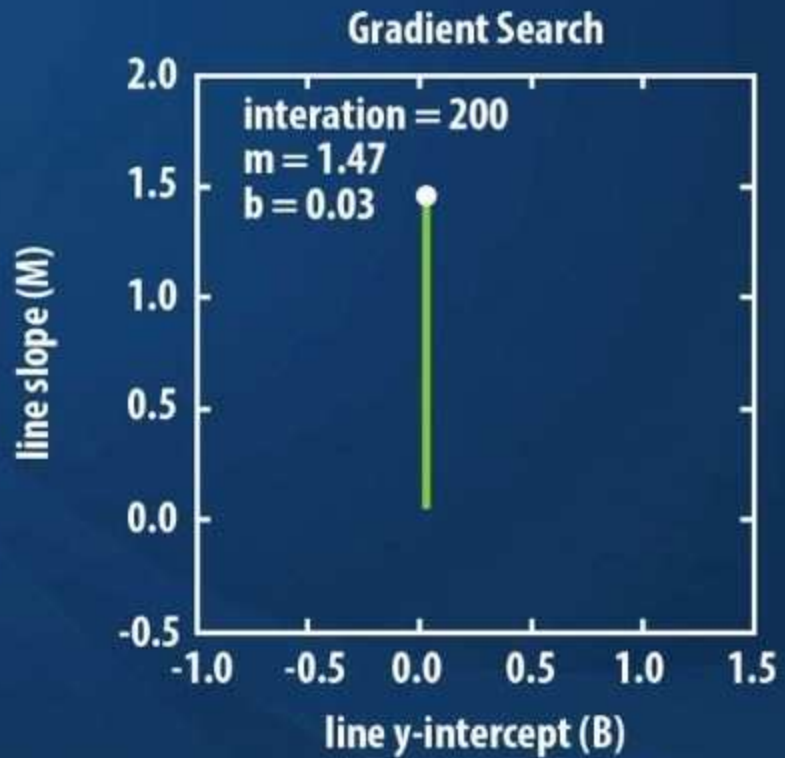


Figure:5

FINDING BEST FIT LINE – LIBRARY METHOD INTERNAL



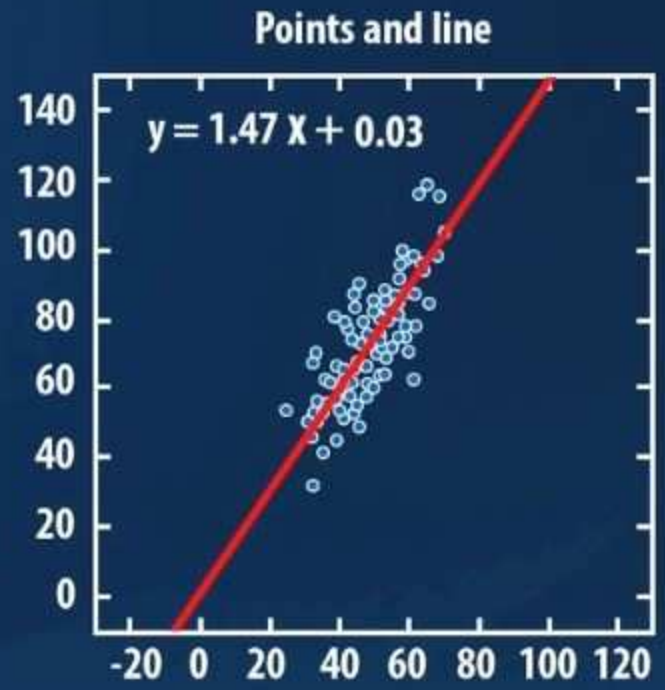


Figure:6

LINEAR REGRESSION EXAMPLE

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet).
- A random sample of 50 houses is selected where:
  - House prices (in Lakhs) is considered Dependent Variable (Y)
  - Size (in square feet) is considered Independent Variable (X)
- What is its linear regression equation? (Find using prebuilt library methods.)

LINEAR REGRESSION EXAMPLE: SAMPLE DATASET

House Price in 1000 (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
378	1910
364	1850
319	1425
...	...

Table:1

LINEAR REGRESSION EXAMPLE: FINDING BEST FIT LINE

- Regression line found for this data set is:

b_0 b_1 Our model

house price = 57.438 + 0.2136 * (square feet)

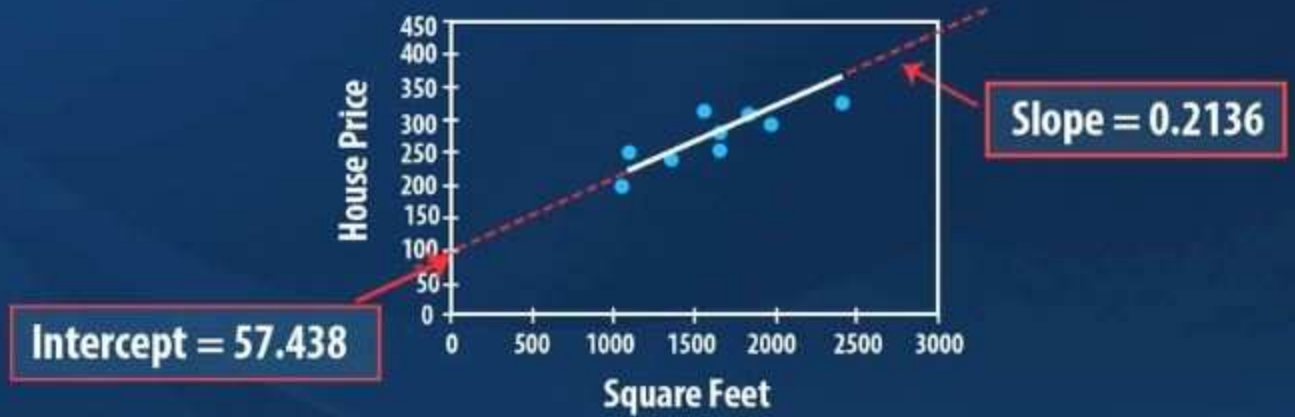


Figure:7

LINEAR REGRESSION EXAMPLE – COEFFICIENT MEANING

- Here, b_0 being = 57.438 indicates that, for houses within the range of sizes observed, Rs 57438 is the portion of the house price not explained by square feet.
- And b_1 being = .2136 means that the average value of a house increases by .2136 (1000) = ~214 /-, for each additional one square ft increase in size.
- This is known as model is interpretable.

LINEAR REGRESSION EXAMPLE – PREDICTING

- Lets predict the price of house when size of house is 2100 sq ft.

$$\text{house price (estimate)} = 57.438 + 0.2136 * (\text{square feet})$$
$$= 57.438 + 0.2136 * (2100)$$
$$= 506 * (1000)$$

Note: Do not extrapolate beyond the range of Xs when doing regression analysis