Analytics Using PySpark

upGrad

Course:
Data Science-Data Engineering

Lecture on:
Linear Regression using PySpark

Contents

upGrad

1. Basic EDA using Spark ML library
2. Linear Regression
3. Logistic Regression
    a. Case Study: CTR Prediction
    b. Hands-On Coding
4. K-Means
    a. Case Study
    b. Hands-On Coding

Machine Learning: Quick Recap

MACHINE LEARNING

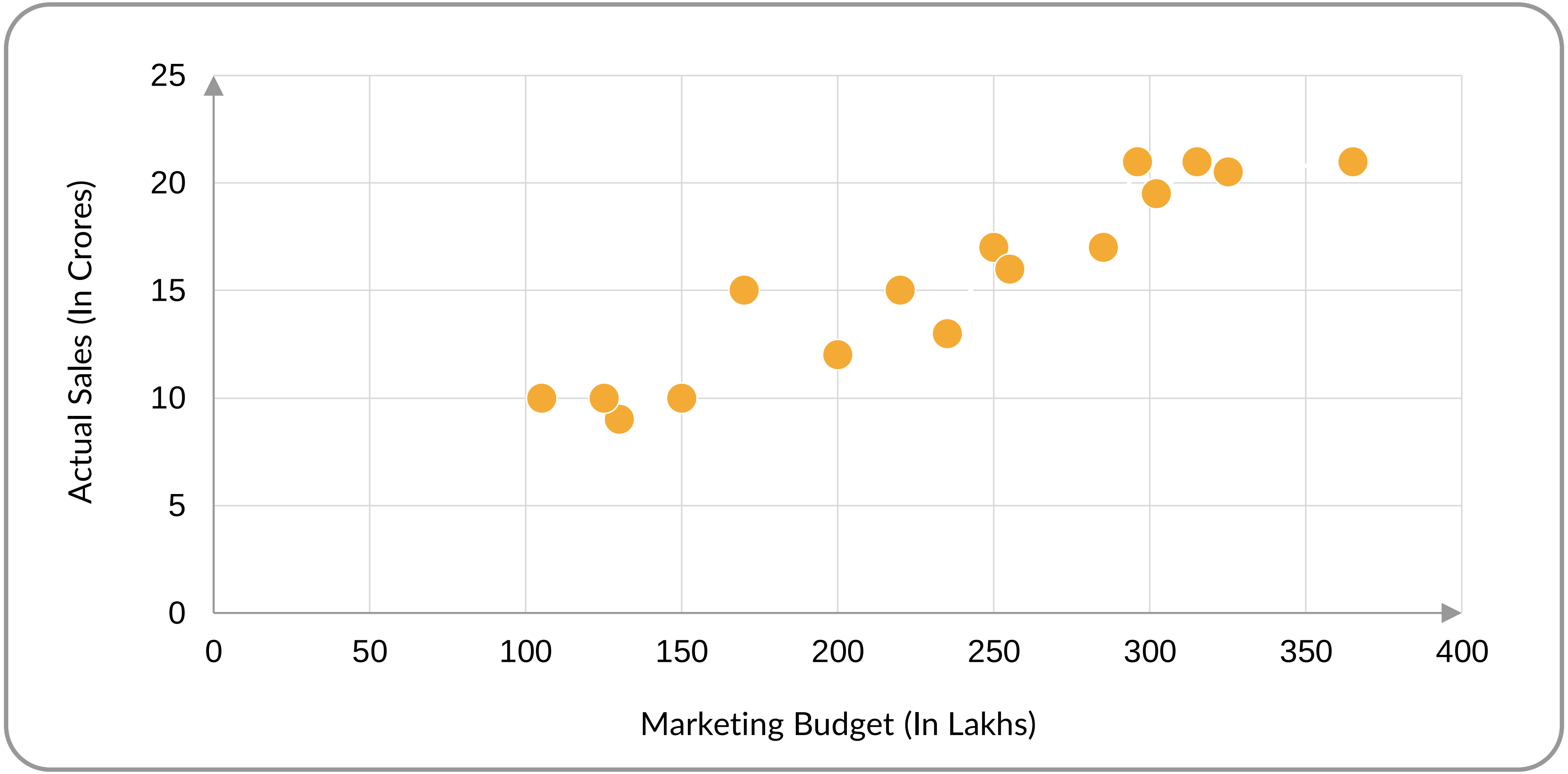
- Machine learning models can be classified into two categories on the basis of the learning algorithm:
  - Supervised learning method: Past data with labels is available to build the model.
    - Regression: The output variable is continuous in nature.
    - Classification: The output variable is categorical in nature.
  - Unsupervised learning method: Past data with labels is not available.
    - Clustering: There is no predefined notion of labels.

Linear Regression

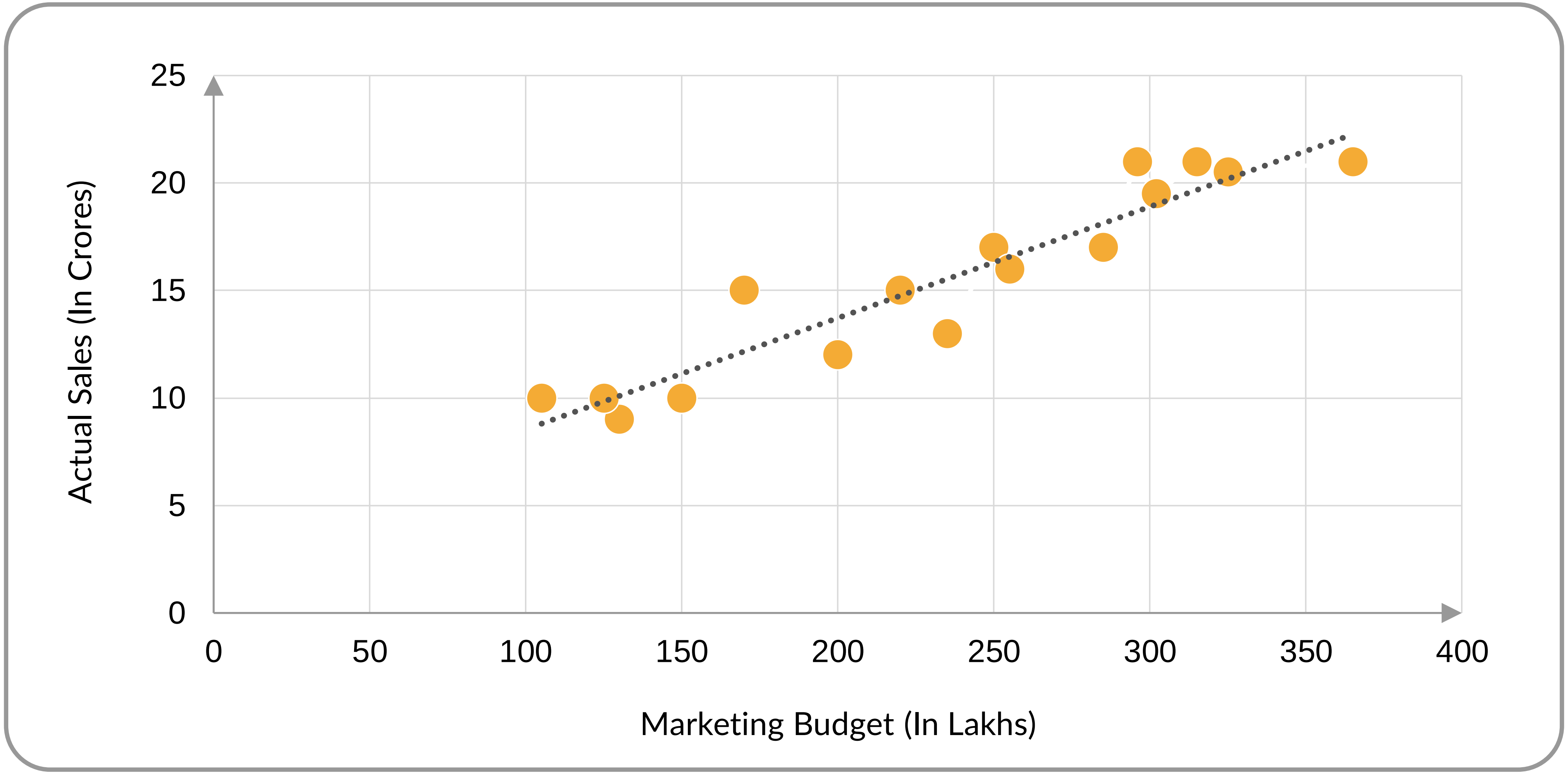
LINEAR REGRESSION

- A simple linear regression model attempts to explain the relationship between a dependent variable and an independent variable using a straight line.
- Example: Sales prediction of a company based on the marketing budget
  - Sales prediction is a dependent variable.
  - Marketing budget is an independent variable.
- Case study: Before deciding the marketing budget, the marketing head wants to know how much will be the sales number.

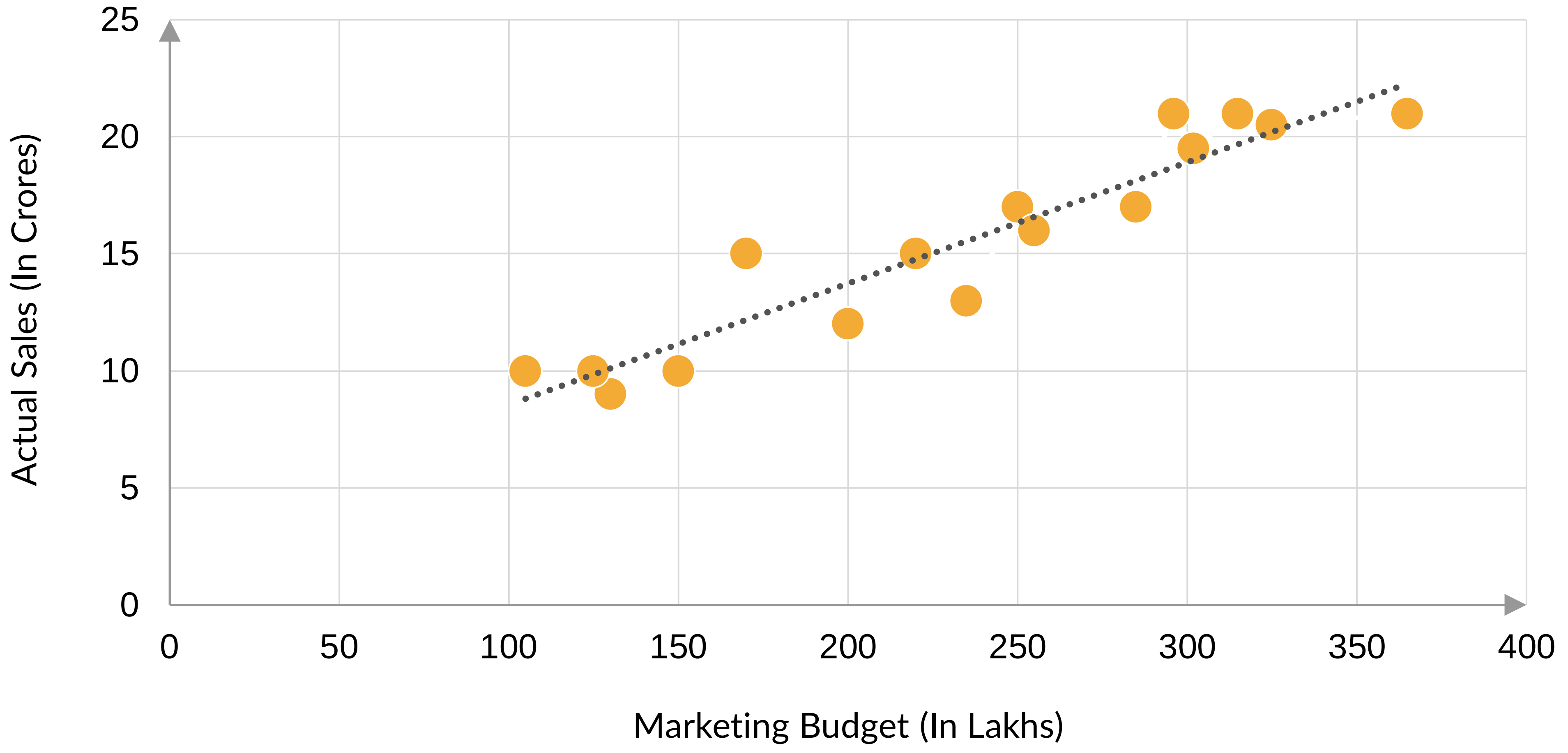
SCATTER PLOT



SIMPLE LINEAR REGRESSION



SIMPLE LINEAR REGRESSION

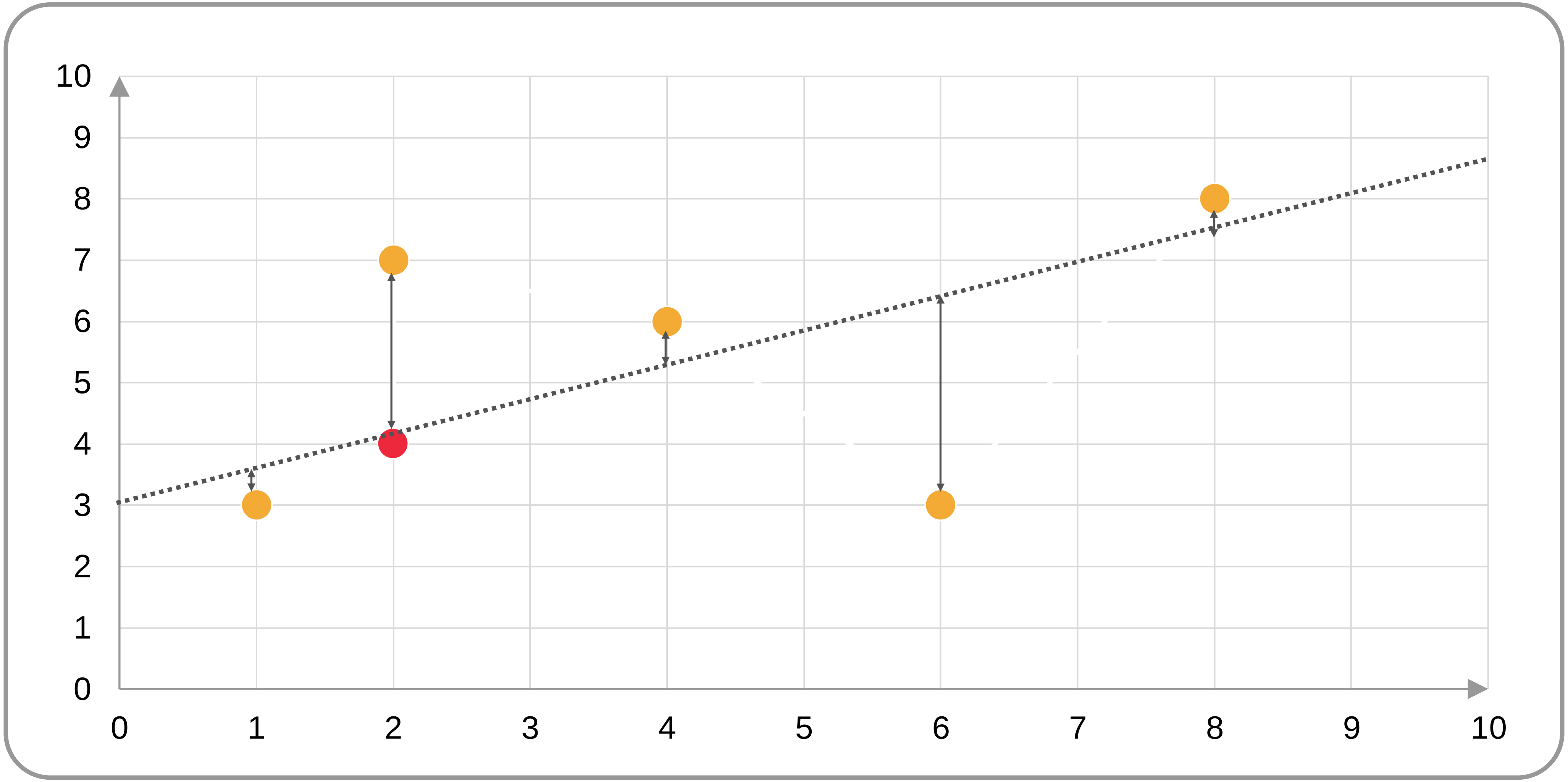


$Y = \beta_0 + \beta_1 X$

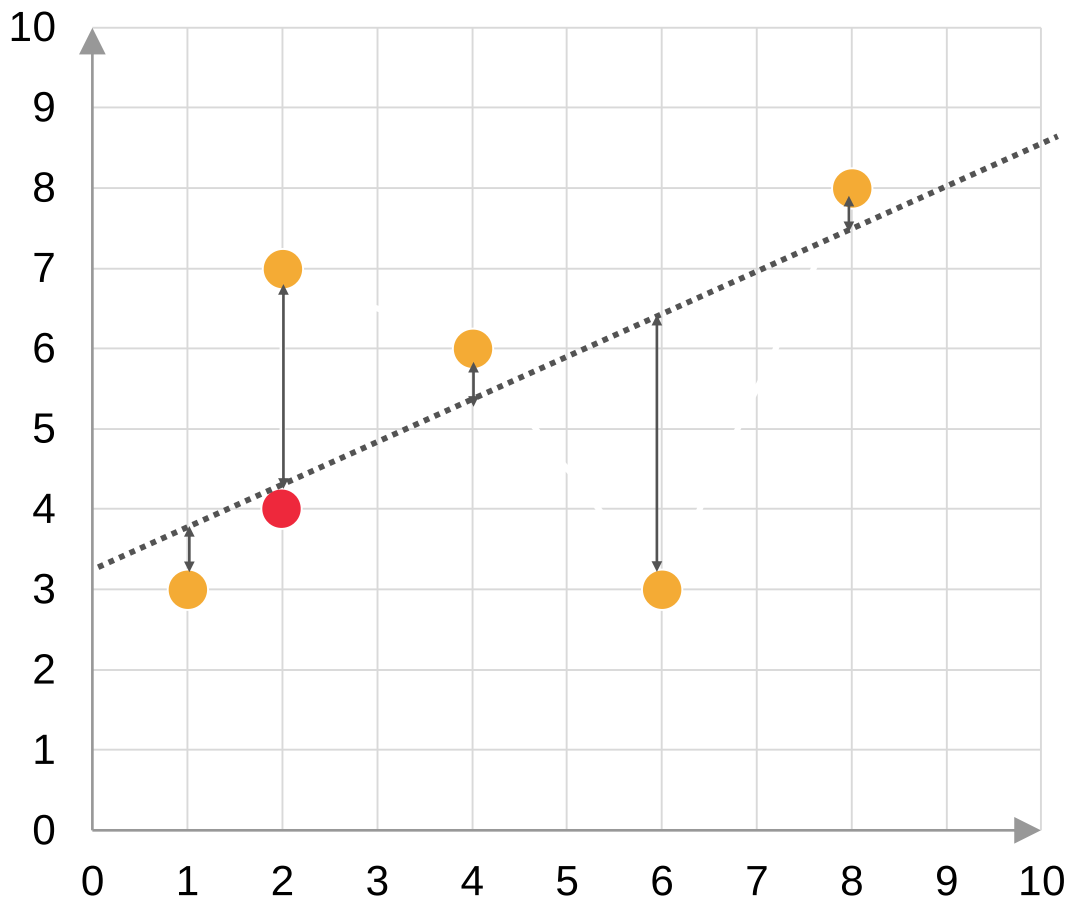
Slope

Intercept

RESIDUALS



RESIDUALS



$Y=\beta_0+\beta_1 X$

Slope

Intercept

$e_i=y_i-y_{pred}$

Ordinary Least Squares Method:

$e_1^2+e_2^2+\_\_\_+e_n^2$ (Residual Sum of Squares)

$RSS=(Y_1-\beta_0-\beta_1 X_1)^2 + (Y_2-\beta_0-\beta_1 X_2)^2\_\_\_\_+(Y_n-\beta_0-\beta_1 X_n)^2$

$$RSS= \sum_{i=1}^{n} (Y_i-\beta_0-\beta_1 X_i)^2$$

LINEAR REGRESSION STEPS

- Start with a scatter plot to check the relationship between Sales and the Marketing Budget.
- Find residuals and the residual sum of squares (RSS) for any given line passing through the scatter plot
- Find the equation of the best-fit line by minimising the RSS and find the optimal values of β_0 and β_1

LINEAR REGRESSION

- You can find the equation of the best-fit regression line ($Y = \beta_0 + \beta_1 X$) by minimising the cost function.
- (RSS in this case, using the ordinary least squares method), which is done using the following two methods:
  - Differentiation
  - Gradient descent
- The drawback in RSS is that it looks at the absolute number.
- If we change the sales number from rupees to dollar, then the RSS value will also change accordingly.
  - Example: (100–95) rupees vs (2–1.5) dollars
  - Both will give vastly different RSS values.

LINEAR REGRESSION

- The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS/TSS)$
  - RSS: Residual sum of squares
  - TSS: Total sum of squares
- TSS: The sum of squares is a measure of how a data set varies around a central number (for example, mean).