19Z601- MACHINE LEARNING

UNIT- 2 LINEAR MODELS

LINEAR MODELS : Linear Regression Models ,Maximum Likelihood Estimation - Least Squares - Bias-Variance Decomposition - Bayesian Linear Regression - Linear Models for Classification, Probabilistic Generative Models - Probabilistic Discriminative Models - Linear Discriminant Analysis (9)

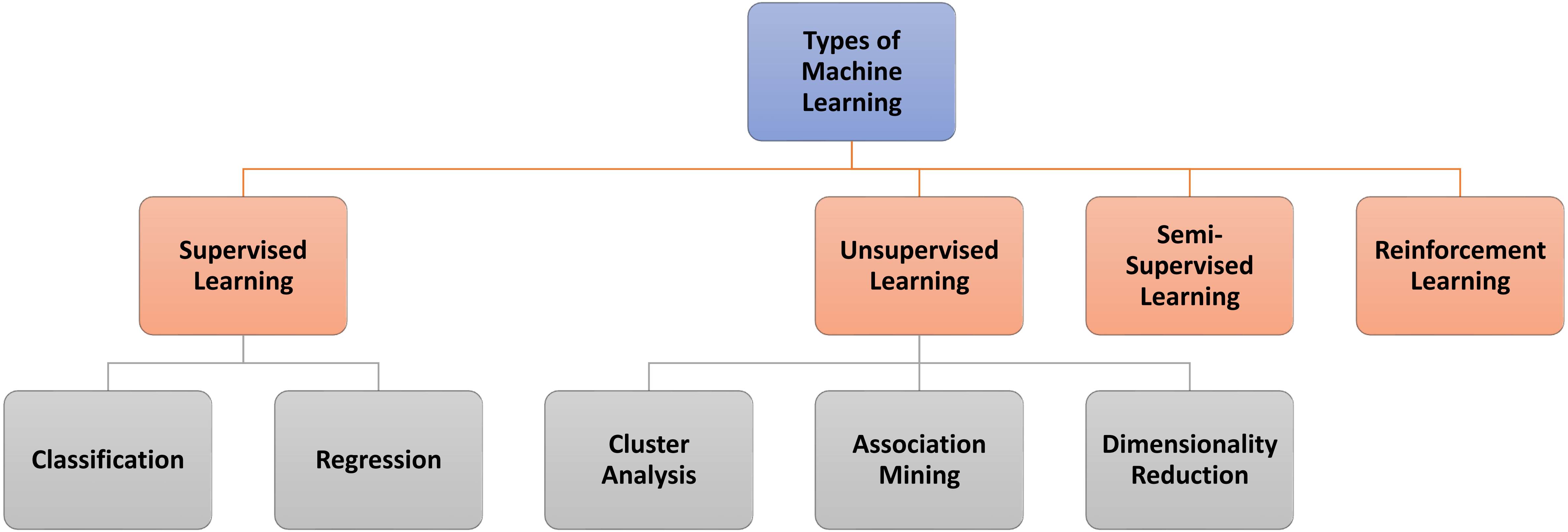
Presented by

Ms.Anisha.C.D

Assistant Professor

CSE

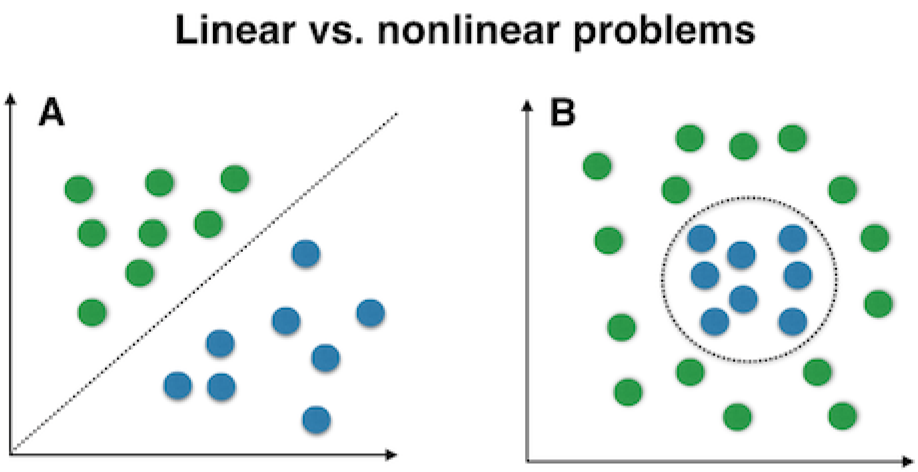
Types of Learning

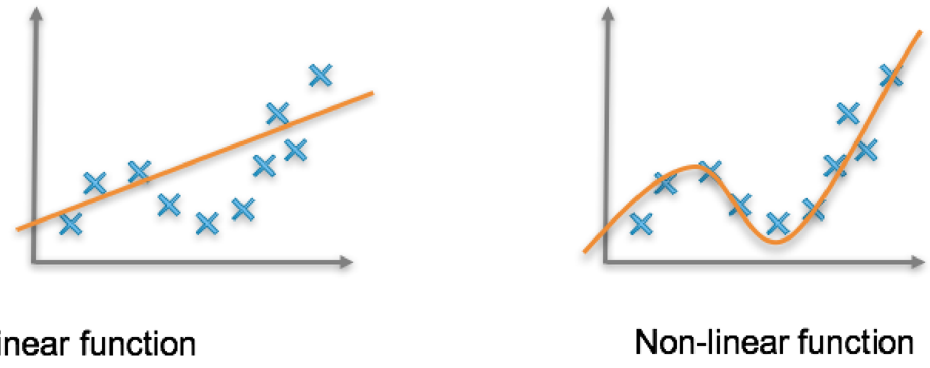


Supervised Learning

TYPE	PARAMETRIC/LINEAR MODELS	NON PARAMETRIC / NON LINEAR MODELS
REGRESSION	LINEAR REGRESSION	KNN REGREESOR
		DECISION TREE REGRESSOR
		RANDOM FOREST REGRESSOR
		BAGGING REGRESSOR
		ADA BOOST REGRESSOR
		XGBOOST REGRESSOR
CLASSIFICATION	LOGISTIC REGRESSION NAVIE BAYES	KNN CLASSIFIER
		DECISION TREE CLASSIFIER
		RANDOM FOREST CLASSIFIER
		BAGGING CLASSIFIER
		ADA BOOST CLASSIFIER
		XGBOOST CLASSIFIER
		SUPPORT VECTOR MACHINES
		ARTIFICIAL NEURAL NETWORK

What is Parametric/Linear Model?





For a linear equation, the highest order of any term is 1. (unit power)

What is Linear Model? - Example

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

- The **polynomial coefficients $w_0,\ldots,w_M$** are collectively denoted by the vector **w.**

- Note that, although the **polynomial function y(x, w)** is a **nonlinear function of x**, it is a **linear function of the coefficients w.**

- **Functions, such as the polynomial, which are linear in the unknown parameters have important properties and are called linear models.**

Linear Models - Regression

- **Regression analysis is a statistical framework** for **quantifying the relationship between a dependent variable and one or more independent variables.**
- Regression analysis comes in many forms—**linear, logistic, ridge, polynomial, and more**—many more!
- Each has an application for datasets with specific characteristics.
- Generally, these models can be categorized as **linear regression**, **multiple linear regression**, and **nonlinear regression**.

Linear Models For Regression – Simple Linear Regression Model

- Linear regression is a **powerful statistical tool** used to **quantify the relationship between variables** in ways that can be **used to predict future outcomes.**
- This method of analysis is used in **stock forecasting, portfolio management, scientific analysis, and many more applications.**
- Whenever one has at **least two variables** in their data—linear regression might be useful.

Goal of Linear Regression Model

- The **goal of linear regression** is to **predict the value of the dependent variable based on the observed value of an independent variable.**
- In the case of **simple linear regression**, this **goal is achieved via modeling the relationship between a dependent variable and a single independent variable**.
- In the case of **multiple linear regression**, the **relationship between the dependent variable is considered with respect to two or more independent variables**.

Assumptions of Linear Regression

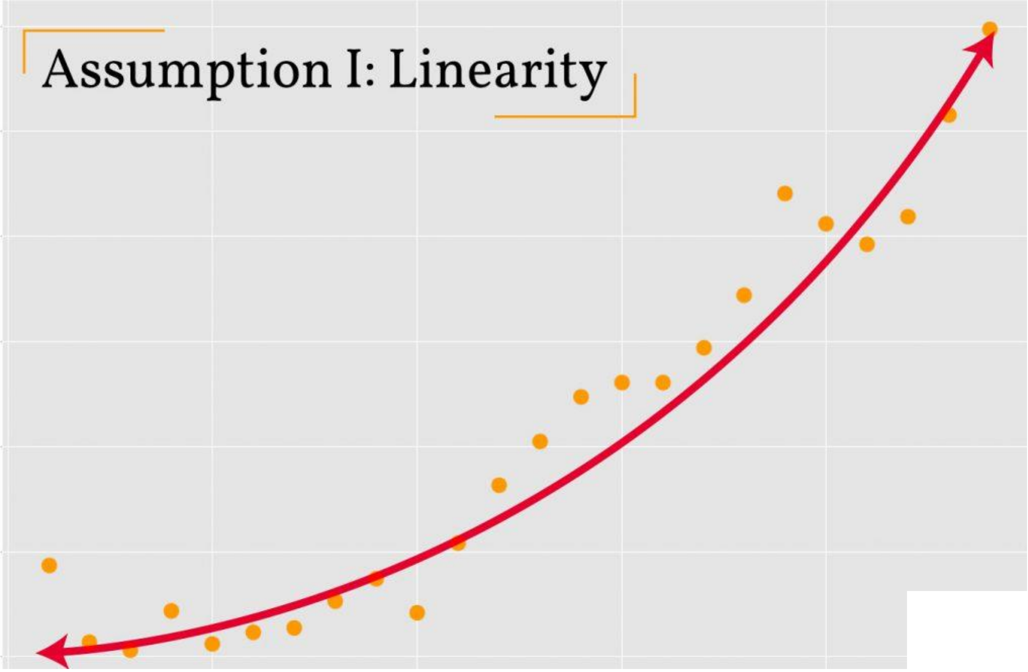


Illustration of Non-Linearity

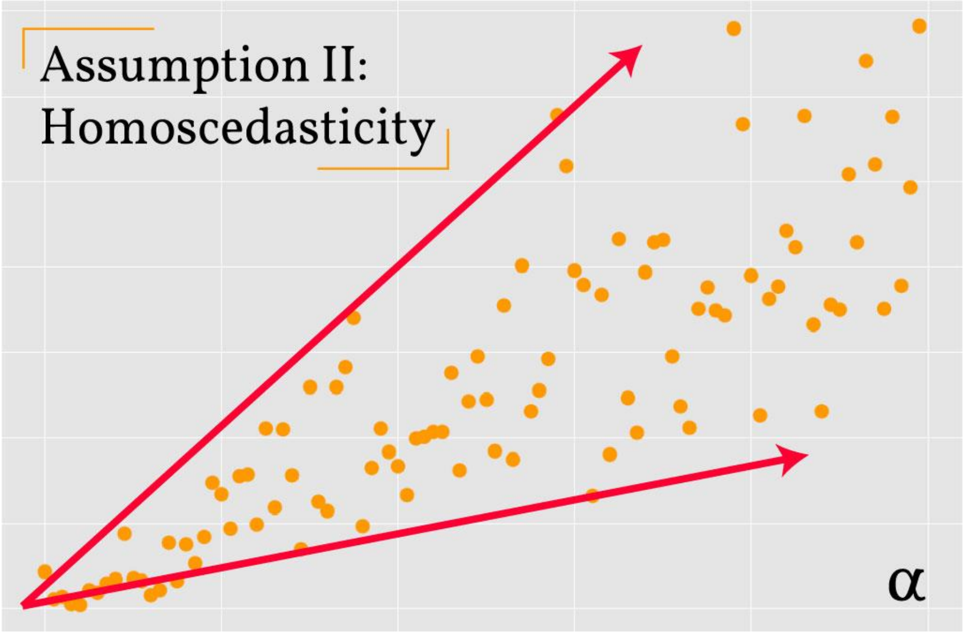
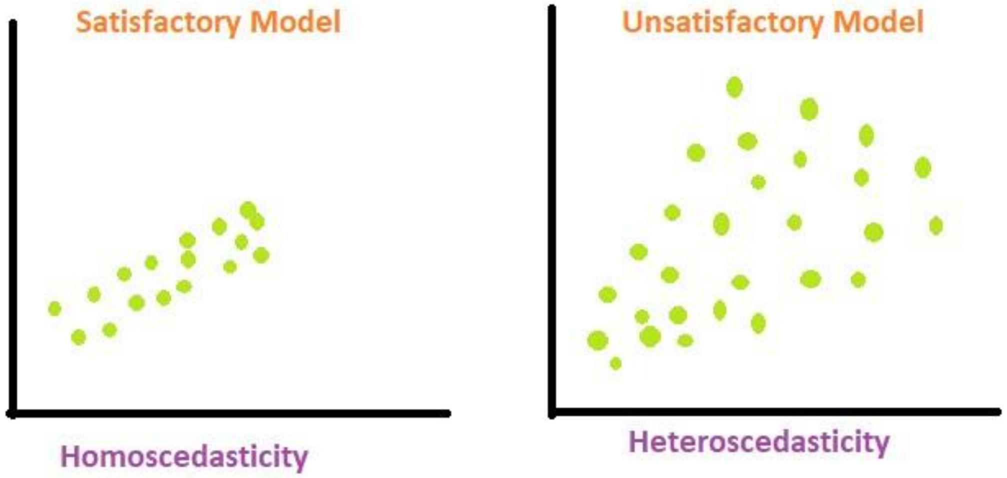
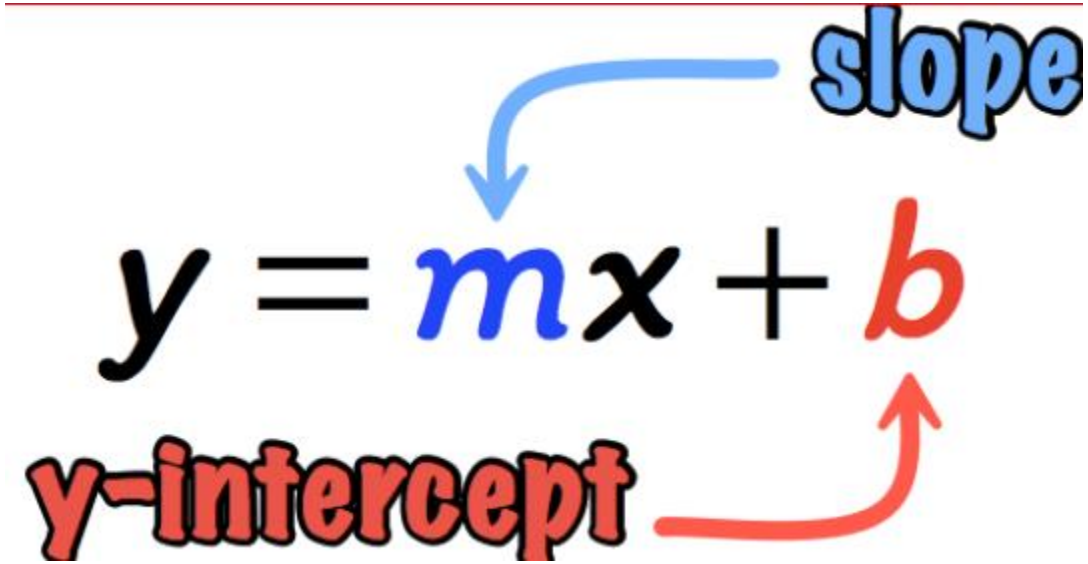


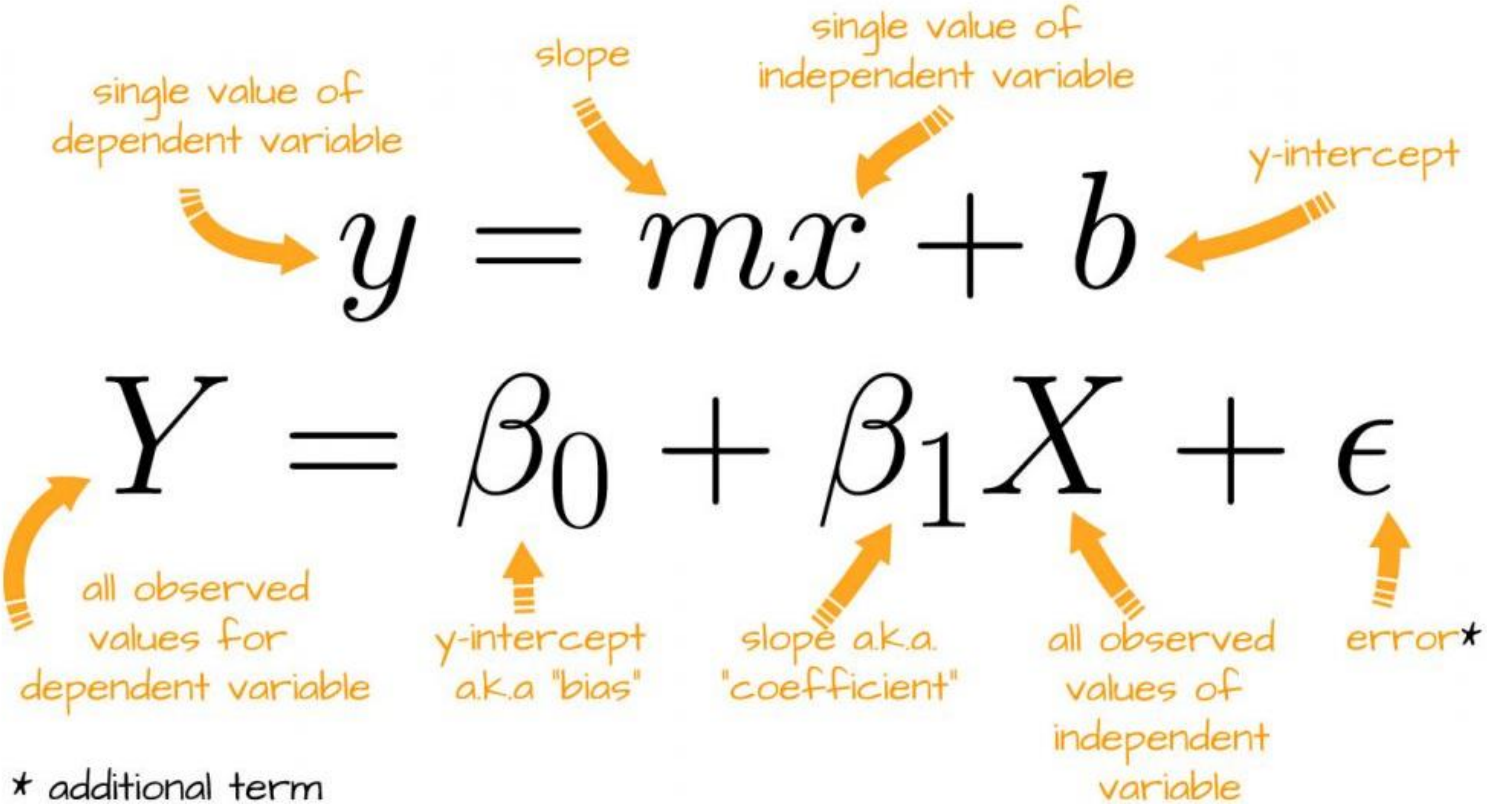
Illustration of Heteroscedasticity



Formulae of Simple Linear Regression Model

- Linear regression fits a **line of best fit** such that the **distance of predicted values from the mean of observed values is minimized.**
- The formulae for varying Linear regression models are based on the algebraic **slope-intercept form**.





Img source : https://www.alpharithms.com/simple-linear-regression-modeling-502111/

slope of X term representing the change in Y for a unit change in X

observed value for the ith term of Y

error term representing $y_i - \bar{y}$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

constant term where y = 0

observed value for the ith term of X

$$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2$$
$$y_3 = \beta_0 + \beta_1 x_2 + \varepsilon_3$$
$$\cdots \qquad \cdots \qquad \cdots \qquad \cdots$$
$$y_n = \beta_0 + \beta_1 x_n + \varepsilon_n$$

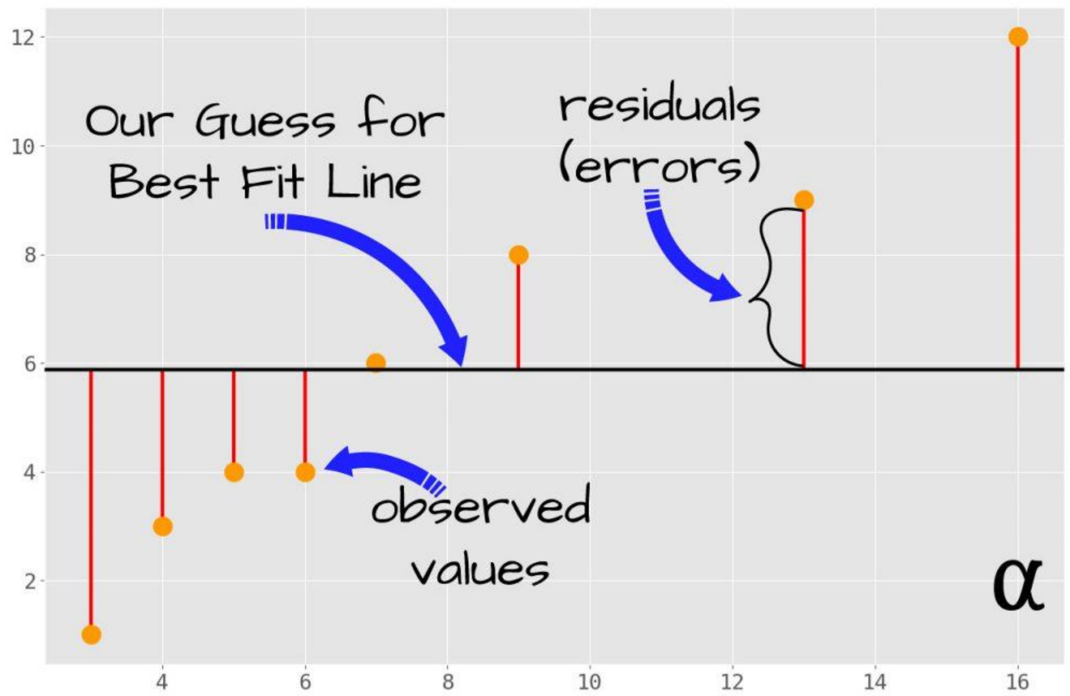
Simple Linear Regression – Sample Data

x	y
3	1
4	3
5	4
6	4
7	5
9	8
13	9
16	12

Img source : https://www.alpharithms.com/simple-linear-regression-modeling-502111/

Create a **scatterplot of the observed values** and also an initial "best guess" line—being "fit" using the **mean of the dependent variable (y)**

values y = mean(Y).



x	y
3	1
4	3
5	4
6	4
7	5
9	8
13	9
16	12

Line of Best Fit: the black horizontal line which is currently just our "best guess" which is simply y = mean(y-values).
Observed Values: the yellow dots representing the (x, y) pairs of our data where the X is our independent (predictor) variable and the y is our dependent (response) variable.
Residuals: the red lines illustrating the between our current y-values and our line of best fit.
Coefficient of Determination (r2): A sum of the *standardized* residual values that provides a non-zero estimate of the total error in our model. Simply the sum of the squared values of all the red lines.

Img source : https://www.alpharithms.com/simple-linear-regression-modeling-502111/

Calculating the Error

$$\varepsilon_i = y_i - \hat{y}_i$$

error term for a single predicted value

single observed value for independent variable

estimated (predicted) value of a single observed value

The goal of regression is to find the equation of the line that will <u>minimize</u> the sum of the squared values of our residuals (Coefficient of Determination.)

x	y	$y - \hat{y}$	$(y - \hat{y})^2$
3	1	1 – 5.75 = -4.75	22.5625
4	3	3 – 5.75 = -2.75	7.5625
5	4	4 – 5.75 = -1.75	3.0625
6	4	4 – 5.75 = -1.75	3.0625
7	5	5 – 5.75 = -0.75	0.5625
9	8	8 – 5.75 = 2.25	5.0625
13	9	9 – 5.75 = 3.25	10.5625
16	12	12 – 5.75 = 6.25	39.0625
Totals	5.75	0	91.5

Img source : https://www.alpharithms.com/simple-linear-regression-modeling-502111/

Simple Linear Regression Model Building

line of least squares

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

single predicted value of dependent variable

estimator of β_0

estimator of β_1

single value for the independent variable

Img source : https://www.alpharithms.com/simple-linear-regression-modeling-502111/

Estimation of Parameters

estimate of β_1 (slope)

error for single value of y

error for single value of x

$$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

sum of terms for all observed values

standardized error for single value of x

Img source : https://www.alpharithms.com/simple-linear-regression-modeling-502111/

estimator for intercept

estimator for slope

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

sample mean of observed values for dependent variable

sample mean of observed values for independent variable

Img source : https://www.alpharithms.com/simple-linear-regression-modeling-502111/

$x = \{3, 4, 5, 6, 7, 9, 13, 16\}$ $y = \{1, 3, 4, 4, 6, 8, 9, 12\}$

estimate for first term

$$\hat{\beta}_1 = \frac{(1-5.88)(3-7.88)}{(3-7.88)^2} = 1.0$$

estimate for all terms

$$\hat{\beta}_1 = \frac{111.875}{144.875} = \sim 0.772$$

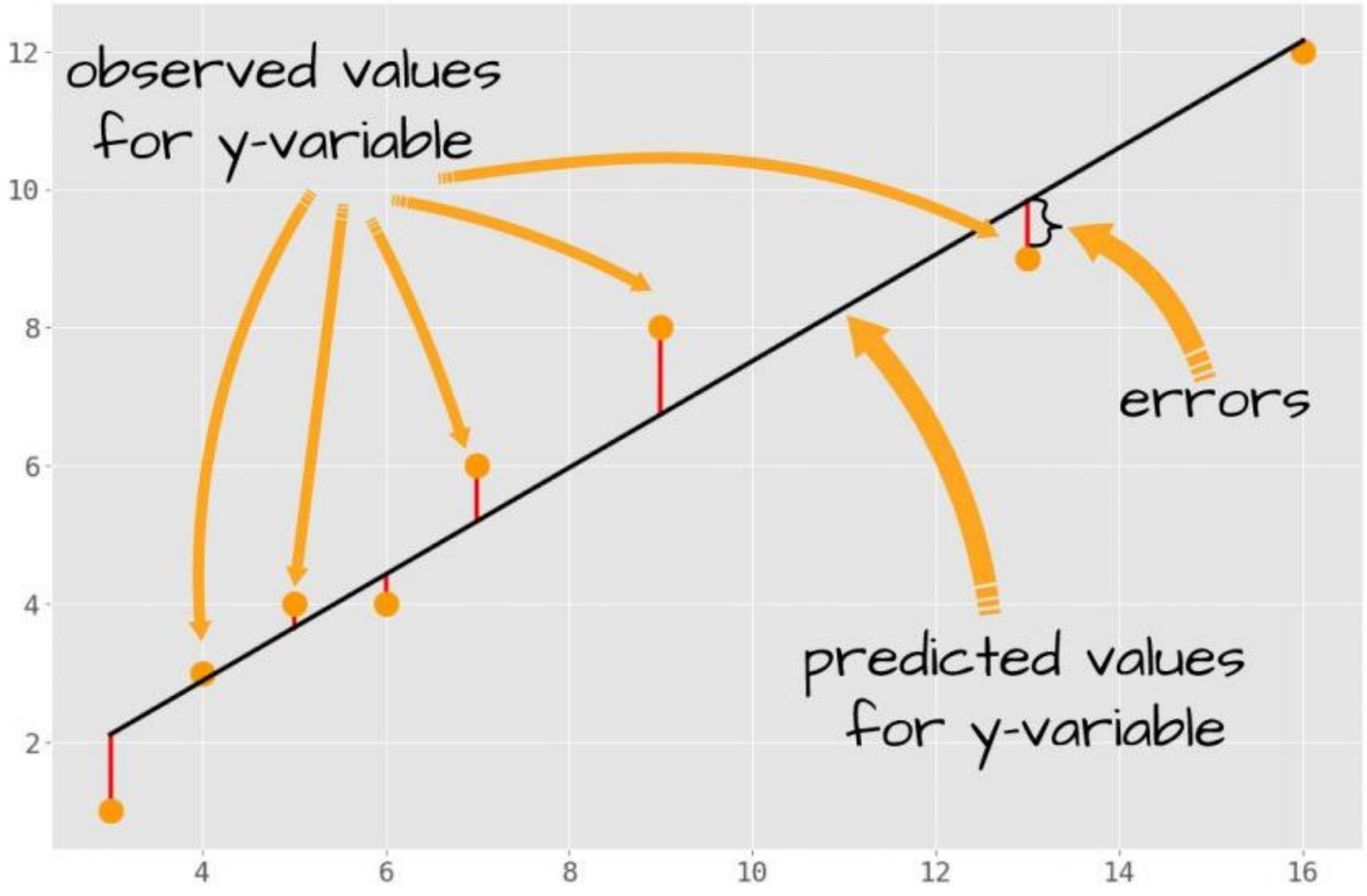
Img source : https://www.alpharithms.com/simple-linear-regression-modeling-502111/

$$\hat{\beta}_0 = \overset{\text{sample mean of y values}}{\bar{y}} - \overset{\text{estimated slope}}{\beta_1}\,\overset{\text{sample mean of x values}}{\bar{x}}$$

$$\hat{\beta}_0 = 5.88 - .772 * 7.88$$

$$\hat{\beta}_0 \cong \underbrace{-0.21}_{\text{estimated y-intercept value}}$$

Img source : https://www.alpharithms.com/simple-linear-regression-modeling-502111/



This line represents the **least-squares regression line** such that the sum of the square errors between our observed values and predicted values is minimized.

The equation for this line is

y = -.21 + .772 * x and can be used to predict future values of x.

Img source : https://www.alpharithms.com/simple-linear-regression-modeling-502111/

Try this (Homework)

You are given a dataset containing information about the number of hours students spend studying and their corresponding scores on a test. Your task is to perform simple linear regression to predict test scores based on the number of hours studied using the following dataset.

No.of Hours Studies	Test Scores
2	75
3	82
4	93
5	89
6	98