UNIT - II

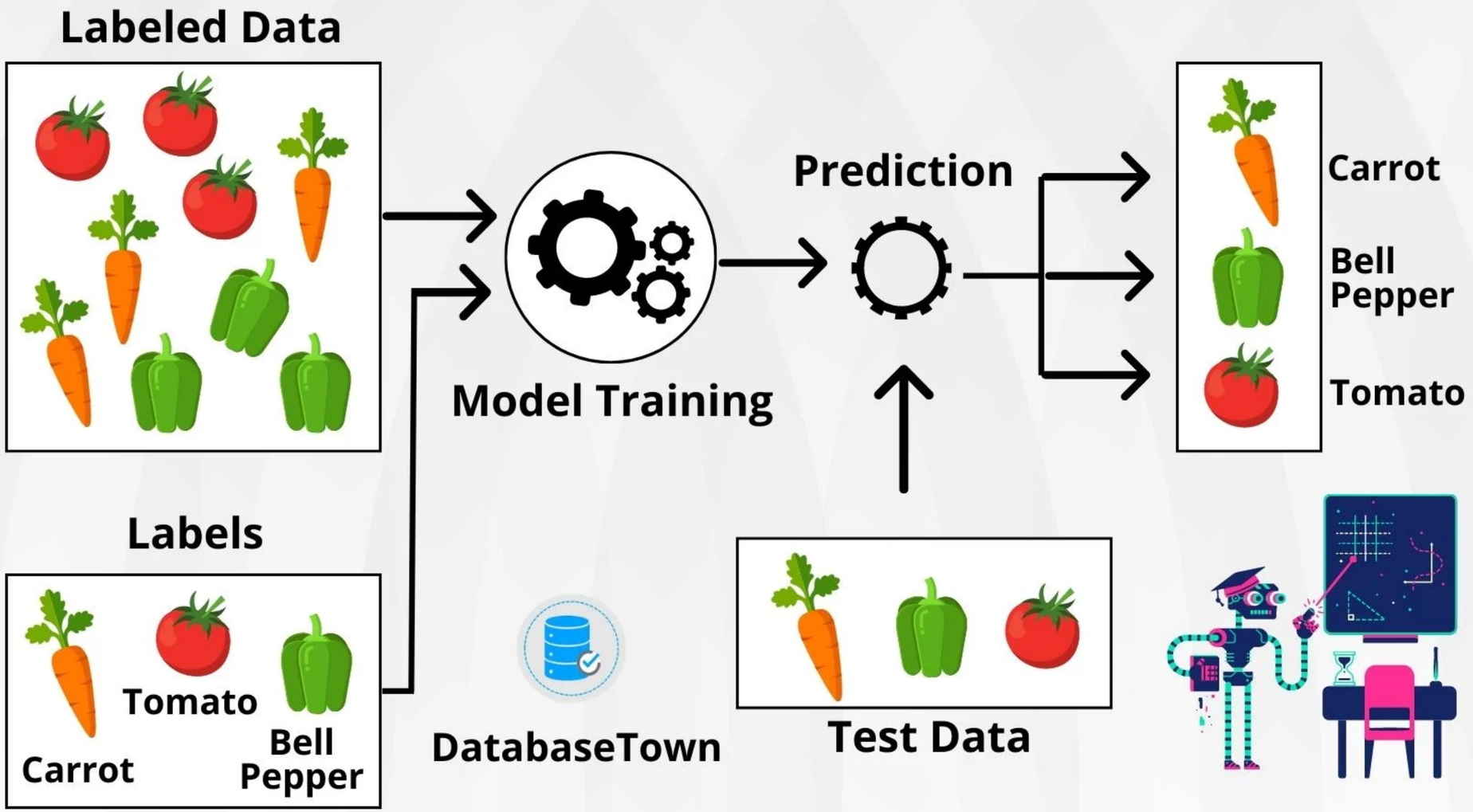
SUPERVISED LEARNING

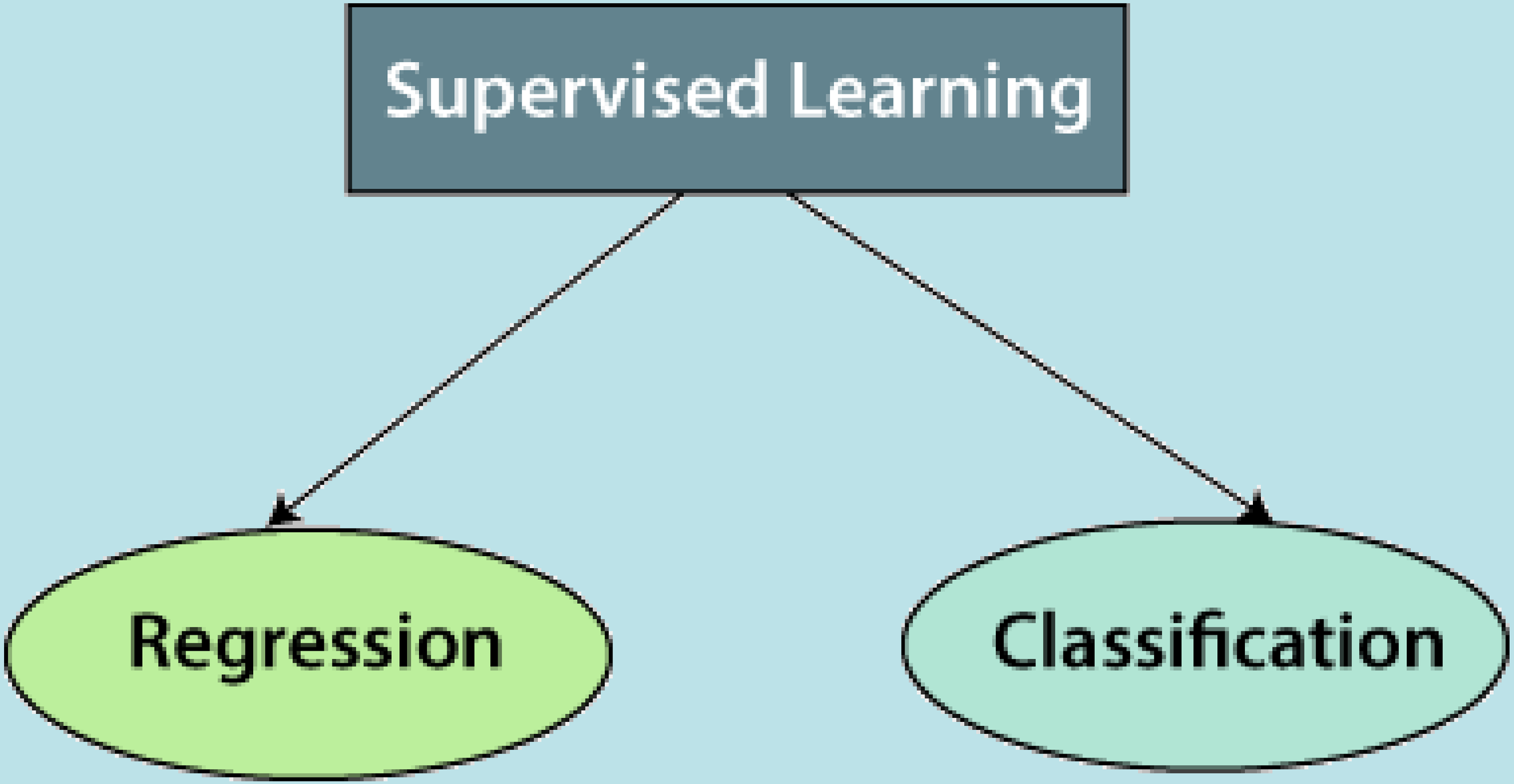
UNIT II SUPERVISED LEARNING 11

Linear Regression Models: Least squares, single & multiple variables, Bayesian linear regression, gradient descent, Linear Classification Models: Discriminant function – Perceptron algorithm, Probabilistic discriminative model - Logistic regression, Probabilistic generative model – Naive Bayes, Maximum margin classifier – Support vector machine, Decision Tree, Random Forests

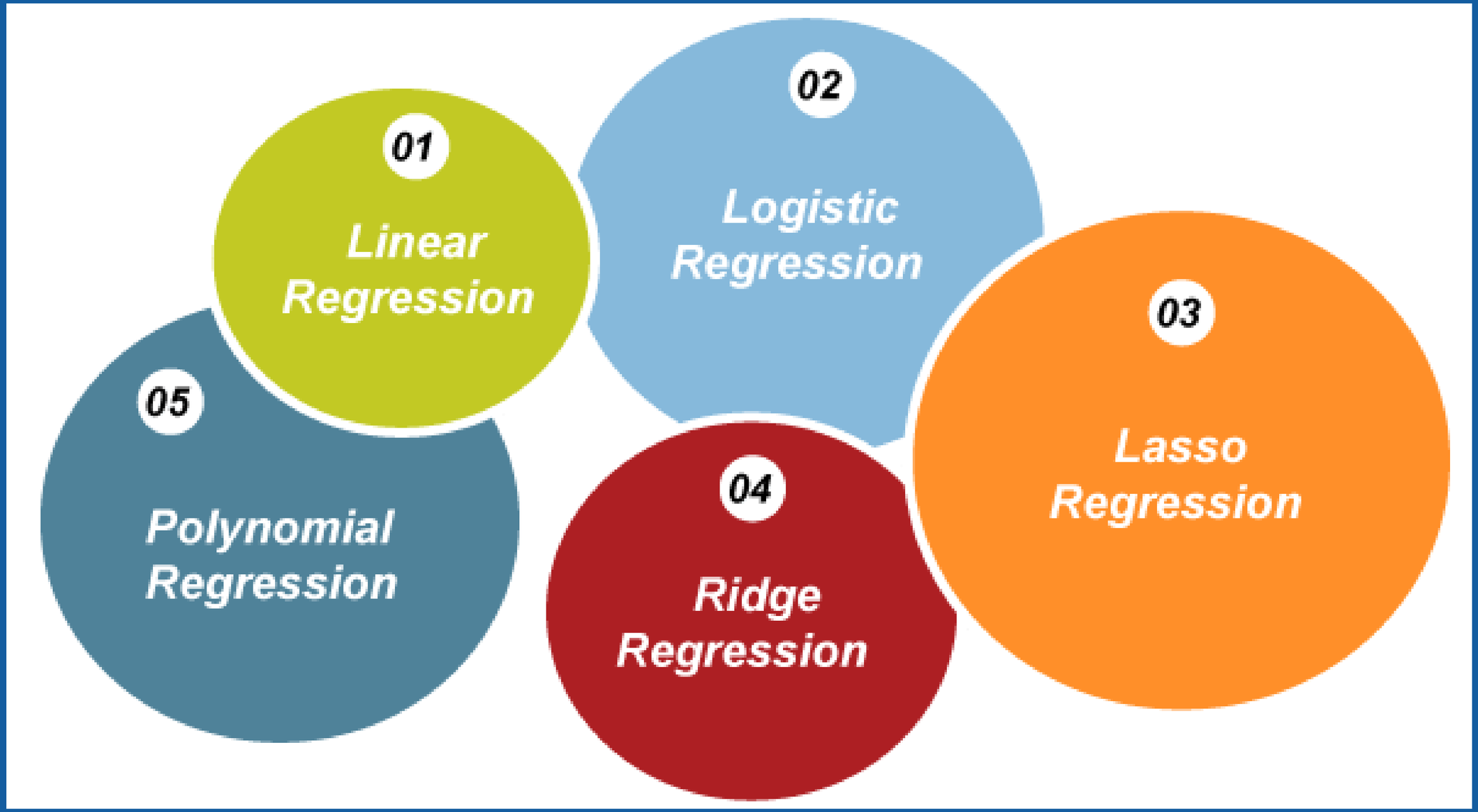
SUPERVISED LEARNING

Supervised machine learning is a branch of artificial intelligence that focuses on training models to make predictions or decisions based on labeled training data.




Supervised Learning
Regression
Classification

Types of Regression



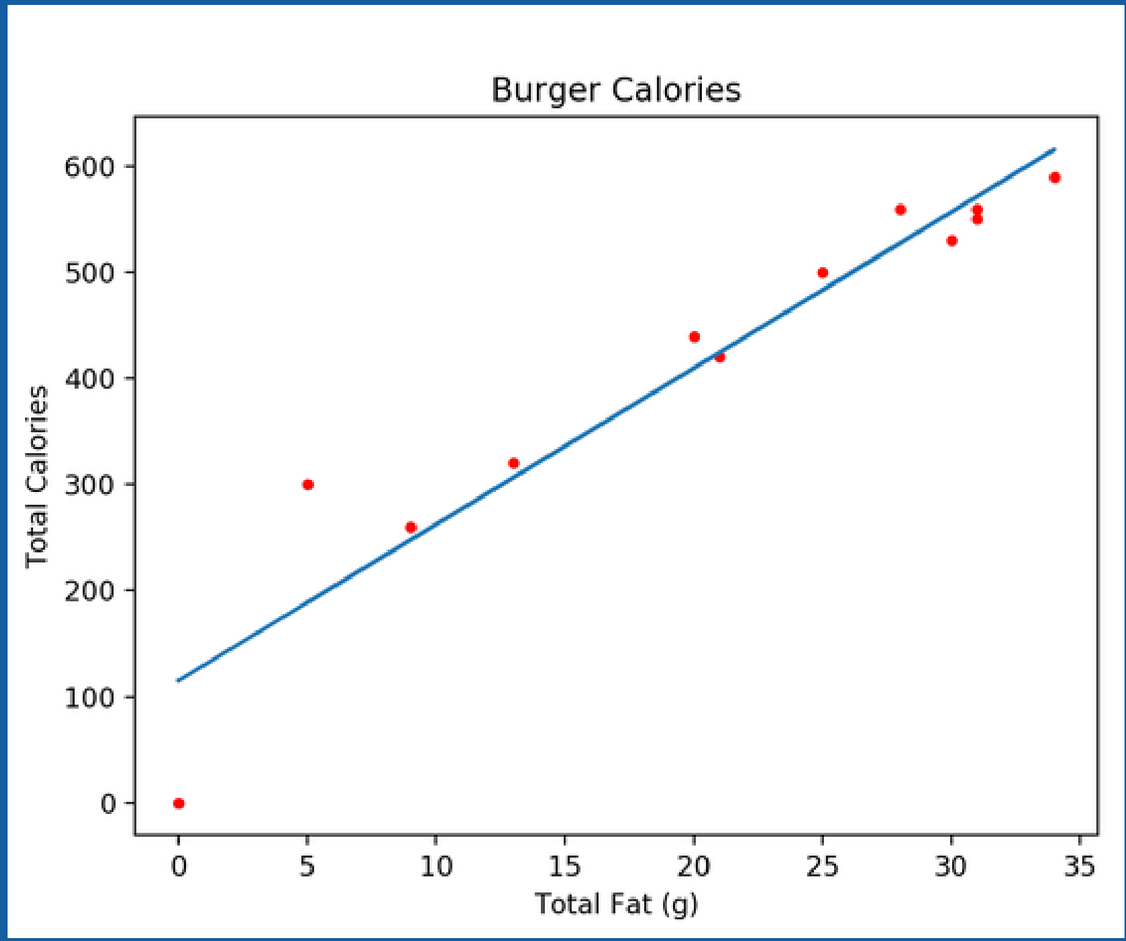
LINEAR REGRESSION MODELS

Linear Regression

- Generate predictions using an easily interpreted mathematical formula.
- Linear regression analysis is used to predict the value of a variable based on the value of another variable.
- The variable you want to predict is called the dependent variable.
- The variable you are using to predict the other variable's value is called the independent variable.
- This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable.

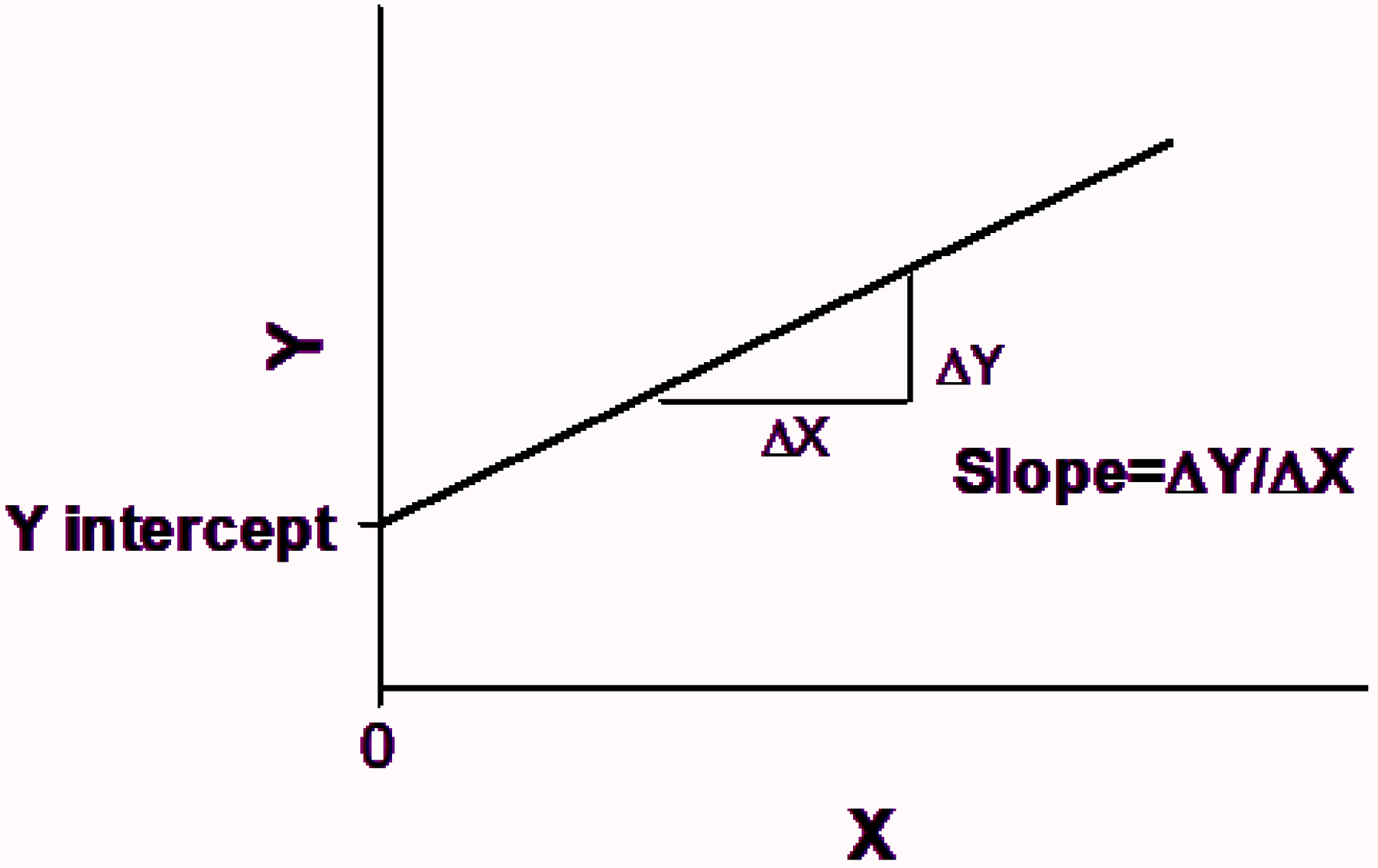
Linear Regression

- Linear regression fits a straight line or surface that minimizes the sum of squared differences between predicted and actual output values (Residual Sum of Squares).
- There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data.

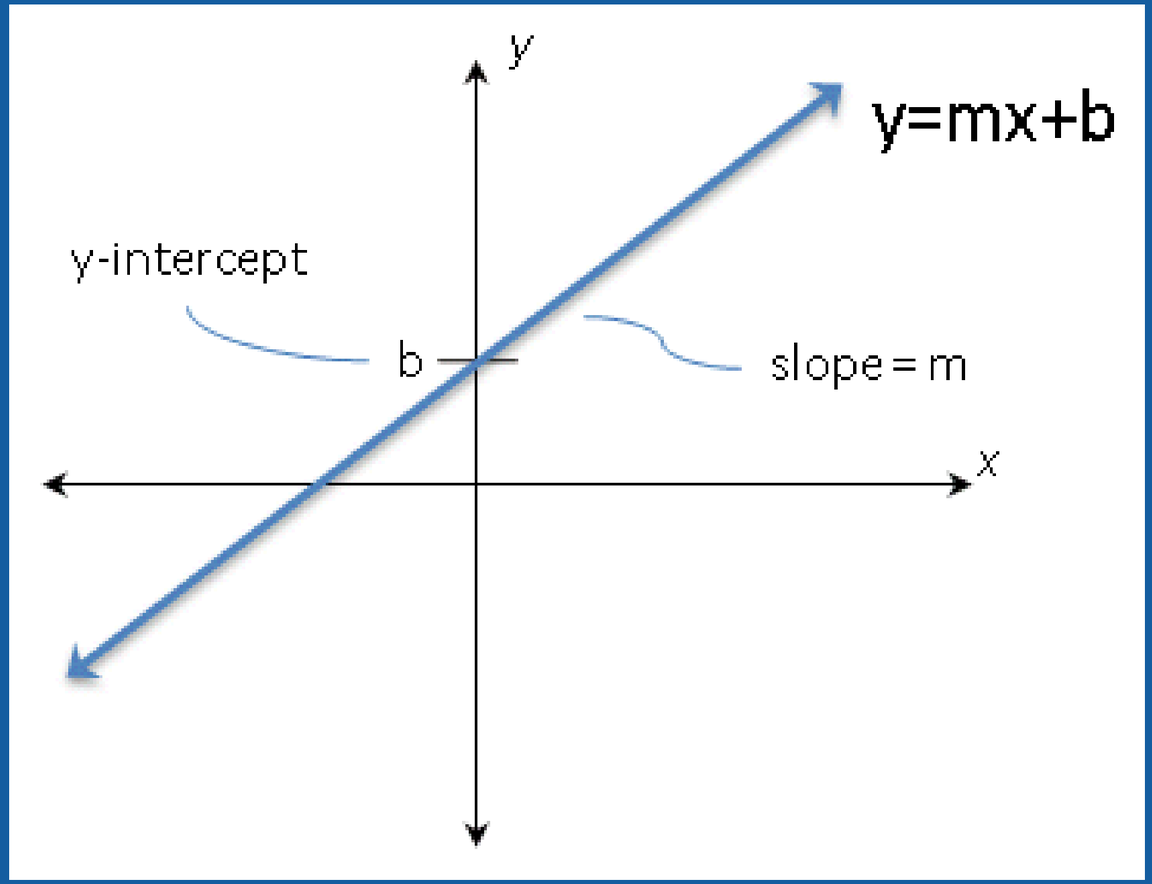


Simple Linear Regression

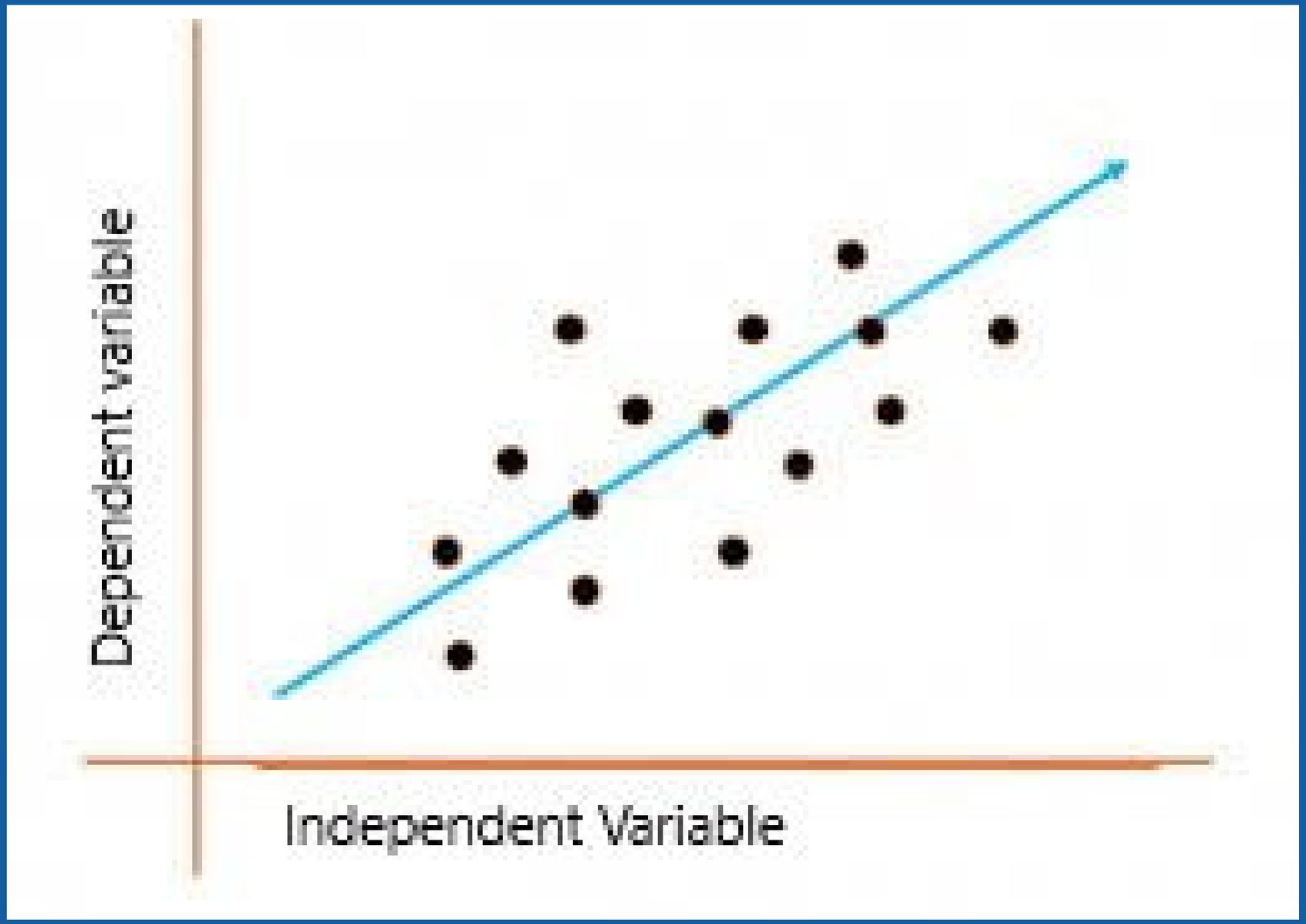
- In a simple linear regression, there is one independent variable and one dependent variable.
- The model estimates the slope and intercept of the line of best fit, which represents the relationship between the variables.
- The slope represents the change in the dependent variable for each unit change in the independent variable, while the intercept represents the predicted value of the dependent variable when the independent variable is zero.

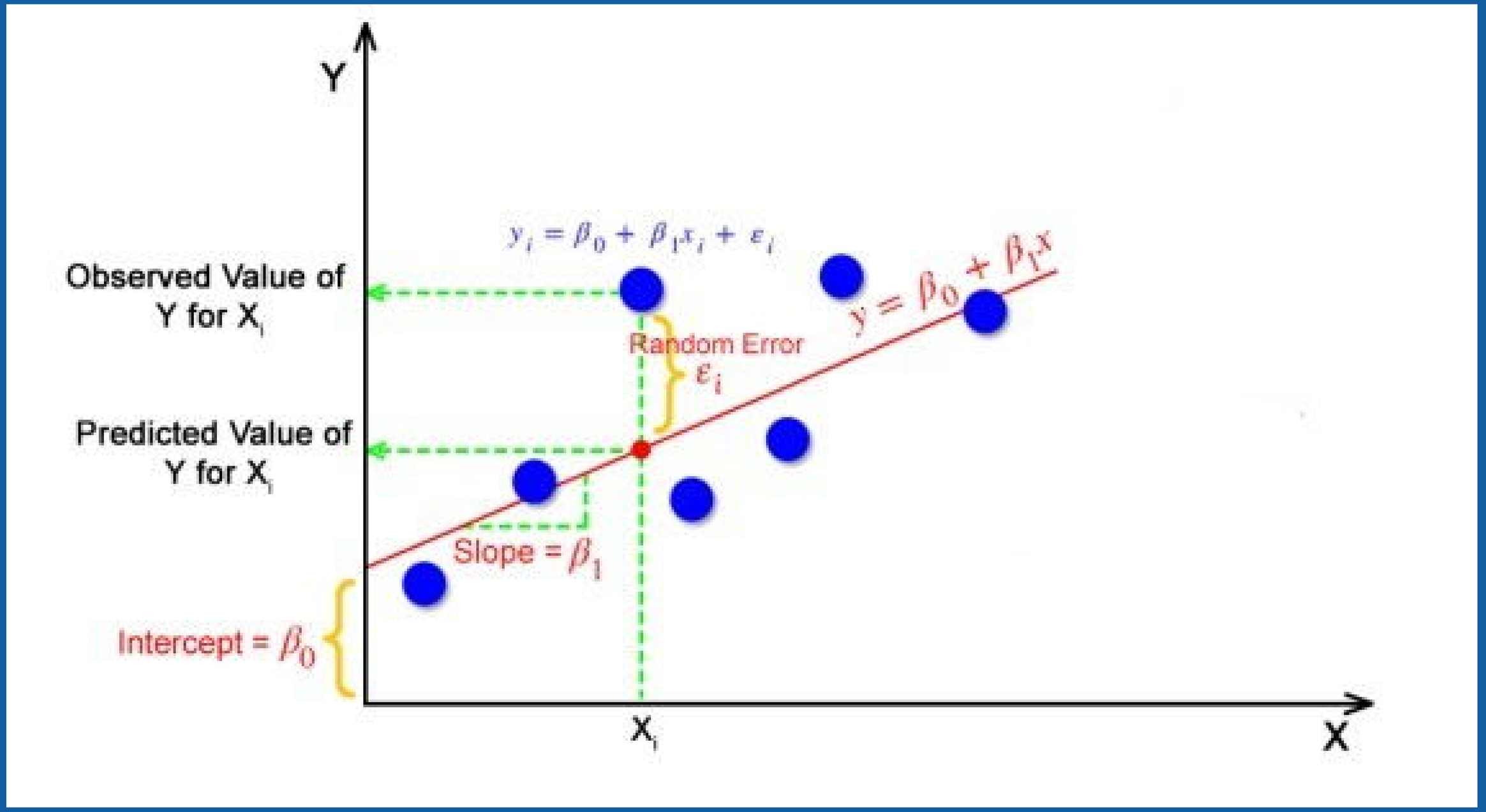

Y
ΔY
ΔX
Slope=ΔY/ΔX
Y intercept
0
X

- The slope quantifies the steepness of the line. It equals the change in Y for each unit change in X.
- The Y-intercept is the Y value of the line when X equals zero. It defines the elevation of the line.
- y = Dependent variable
- x = Independent Variable



Simple Linear Regression




Y
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
Observed Value of Y for X_i
Random Error ε_i
$y = \beta_0 + \beta_1 x$
Predicted Value of Y for X_i
Slope = β_1
Intercept = β_0
X_i
X

Random Error (Residuals)

- In regression, the difference between the observed value of the dependent variable(yi) and the predicted value(predicted) is called the residuals.

  Residual/Error (εi= Actual values – Predicted Values

  $$\varepsilon i = y_{predicted} - y_i$$

  where, $y_{predicted} = B0 + B1\ Xi$

- The vertical distance between data point and the regression line is known as error or residual.

Cost Function for Linear Regression

- The cost function helps to find out the optimal values for slope and intercept, which provides the best-fit line for the data points.
- In Linear Regression, MSE (Mean Squared Error) cost function is used.
- MSE is the average of squared error occurred between ypredicted and yi.

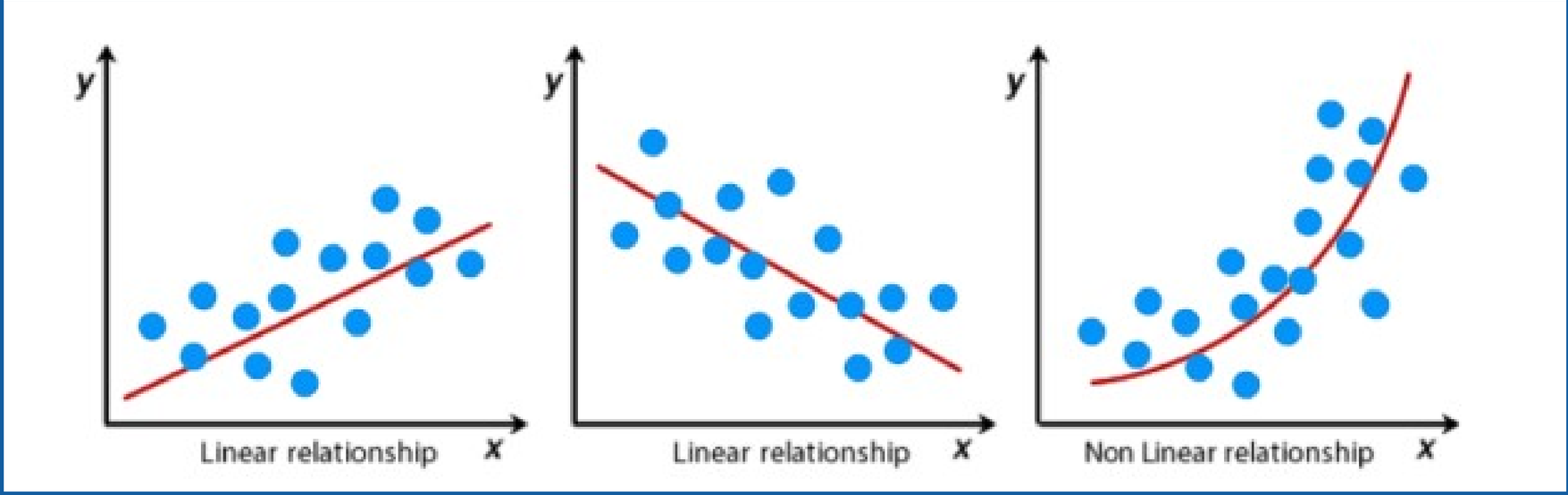
$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (B1x_i + B0))^2$$

- Using the MSE function, we'll update the values of B0 and B1 such that the MSE value settles at the minima.
- These parameters can be determined using the gradient descent method such that the value for the cost function is minimum.

Assumptions of Linear Regression

- Regression is a parametric approach it makes assumptions about the data for analysis. For successful regression analysis, it’s essential to validate the following assumptions.

1. **Linearity of residuals:** There needs to be a linear relationship between the dependent variable and independent variable(s).



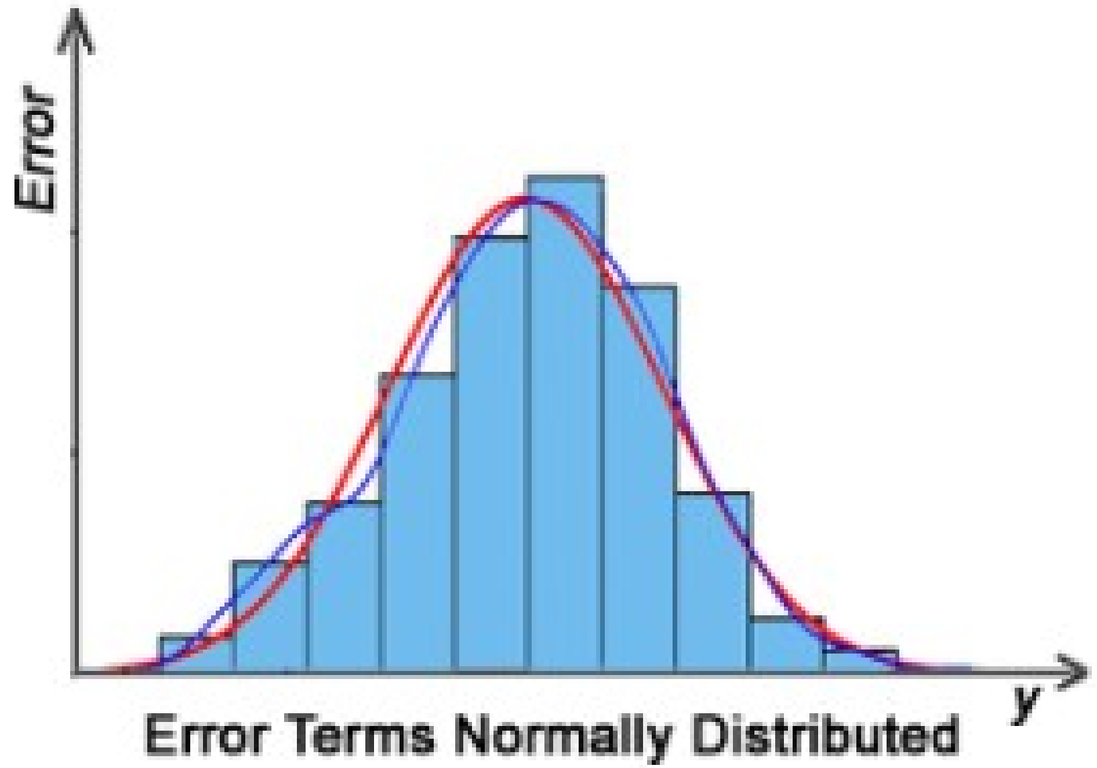
Assumptions of Linear Regression

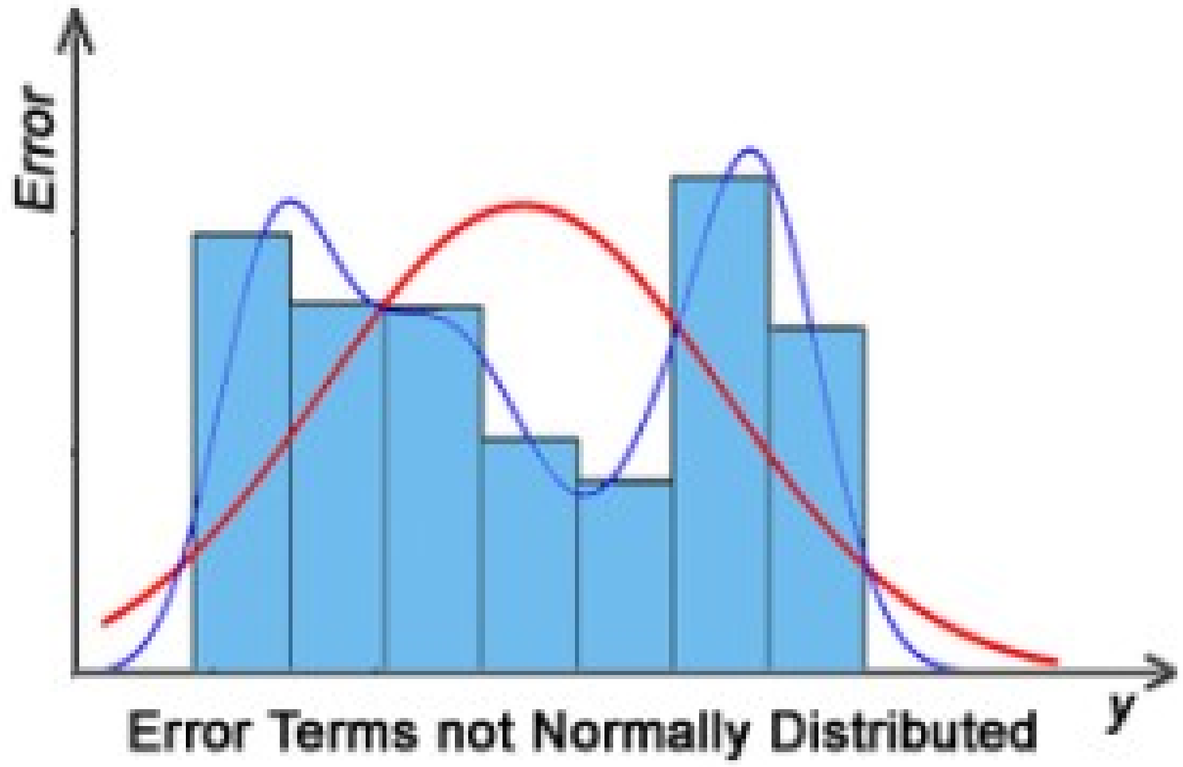
2. Independence of residuals: The error terms should not be dependent on one another.

- There should be no correlation between the residual terms.
- The absence of this phenomenon is known as Autocorrelation.

3. Normal distribution of residuals: The mean of residuals should follow a normal distribution with a mean equal to zero or close to zero.

- This is done to check whether the selected line is actually the line of best fit or not.
- If the error terms are non-normally distributed, this suggests that there are a few unusual data points that must be studied closely to make a better model.

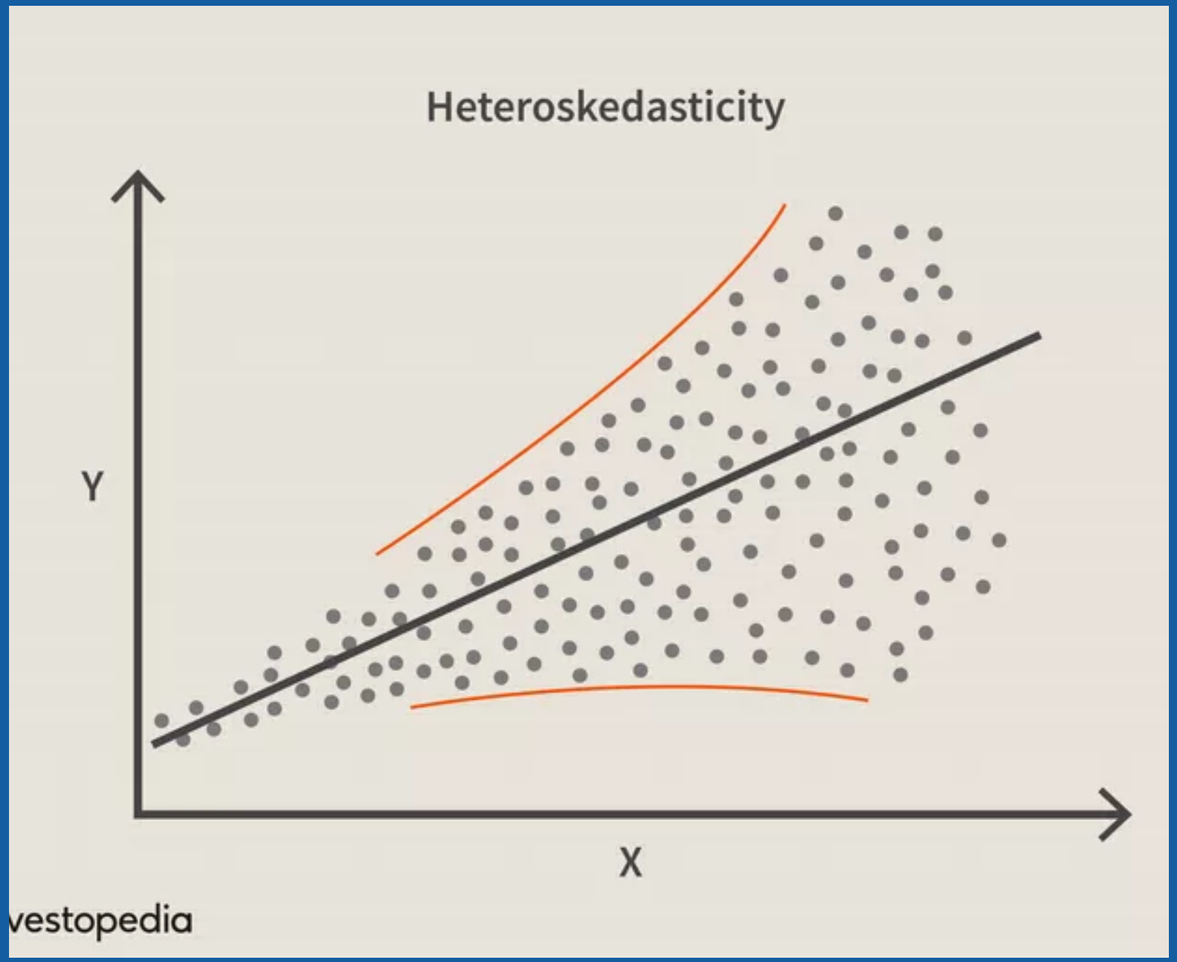

Error
y
Error Terms Normally Distributed

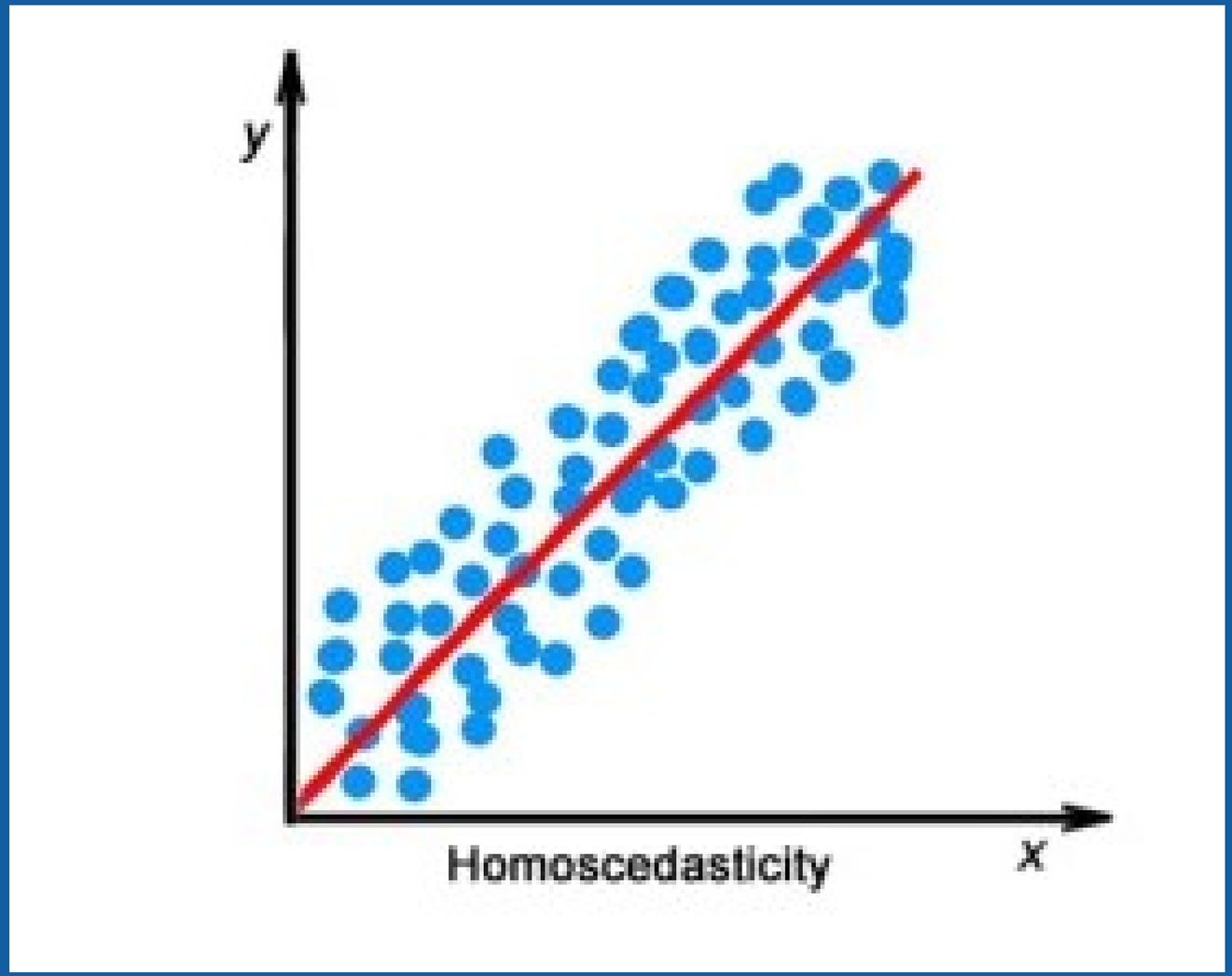

Error
y
Error Terms not Normally Distributed

Assumptions of Linear Regression

4. The equal variance of residuals:

- The error terms must have constant variance and this phenomenon is known as Homoscedasticity.
- The presence of non-constant variance in the error terms is referred to as Heteroscedasticity.
- Generally, non-constant variance arises in the presence of outliers or extreme leverage values.
- In statistics, heteroscedasticity happens when the standard errors of a variable, monitored over a specific amount of time, are non-constant.


Heteroskedasticity
Y
X
vestopedia


y
Homoscedasticity
x

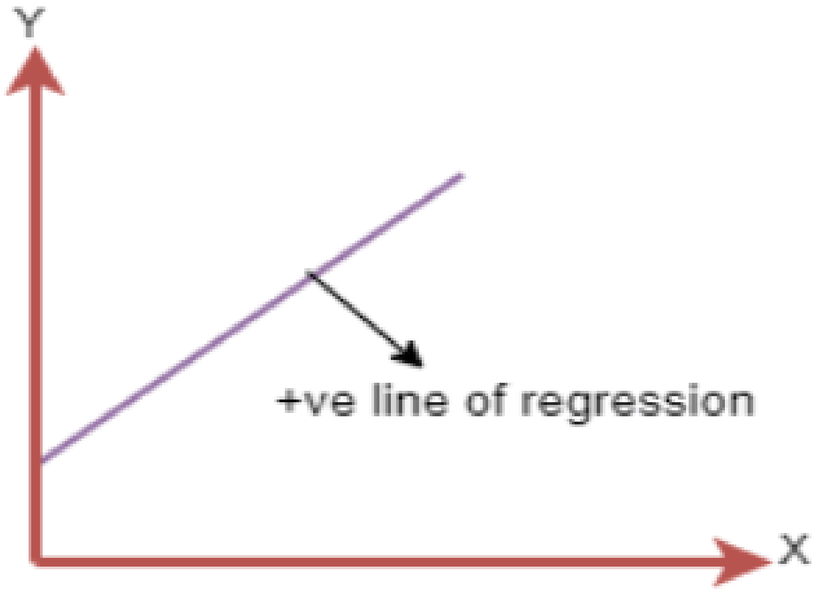
Multiple Linear Regression

- Multiple linear regression is a technique to understand the relationship between a single dependent variable and multiple independent variables.
- The formulation for multiple linear regression is

$$Y = B_0 + B_1X_1 + B_2X_2 + \ldots + B_pX_p + \varepsilon$$

Positive Linear Relationship

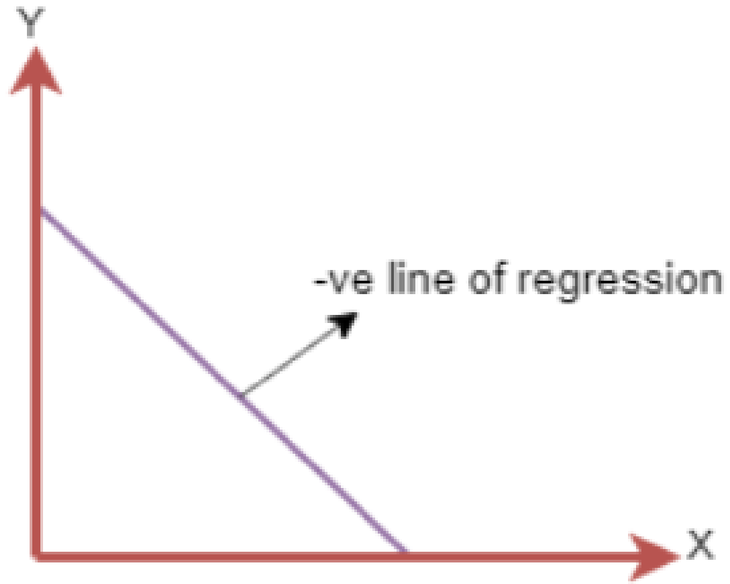
- If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: $\mathbf{Y= a_0+a_1x}$

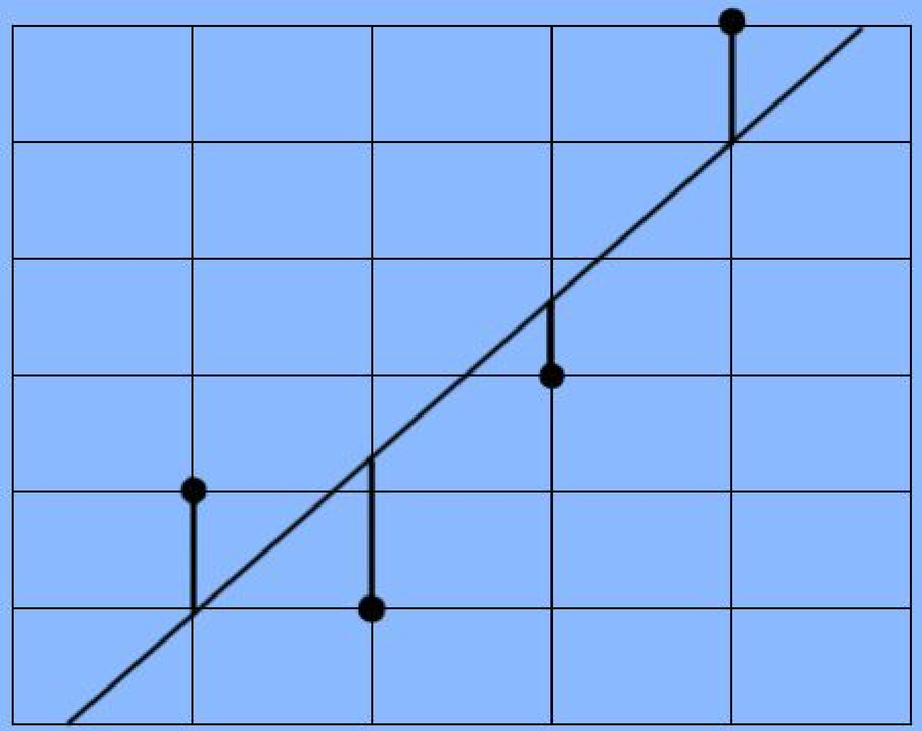
Negative Linear Relationship

- If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



The line of equation will be: $\mathbf{Y= -a_0+a_1x}$

LEAST SQUARES REGRESSION METHOD



Least Squares Method

['lēst 'skwers 'me-thəd]

A form of mathematical regression analysis used to determine the line of best fit for a set of data, providing a visual demonstration of the relationship between the data points.

- The regression line under the least squares method can be calculated using the following formula:

$$\hat{y} = a + bx$$

- Where,
- $\hat{y}$ = dependent variable
- x = independent variable
- a = y-intercept
- b = slope of the line

The slope of line is calculated using the formula,

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Y-intercept, 'a' is calculated using the following formula:

$$a = \frac{\sum y - (b \sum x)}{n}$$

Find the best fit line for the below problem using Least square regression.

X	Y
Years of Expeience (Years)	Salary (in 1000$)
18	90
12	64
2	15
8	47
16	75
11	61
1	8
9	49
5	25

X	Y (Actual)	Y (Predicted)	Error (Residual)
Years of Expeience (Years)	Salary (in 1000$)		
18	90	91.27	-1.27
12	64	63.07	0.93
2	15	16.07	-1.07
8	47	44.27	2.73
16	75	81.87	-6.87
11	61	58.37	2.63
1	8	11.37	-3.37
9	49	48.97	0.03
5	25	30.17	-5.17

Exercise

- Find the salary of a person having years of experience as 20, 25 and 27

Problem 2

Experience of Technician (in Years)	Performance Rating
16	87
12	88
18	89
4	68
3	78
10	80
5	75
12	83

KEY TAKEAWAYS

- The least squares method is a statistical procedure to find the best fit for a set of data points.
- The method works by minimizing the sum of the offsets or residuals of points from the plotted curve.
- Least squares regression is used to predict the behavior of dependent variables.
- The least squares method provides the overall rationale for the placement of the line of best fit among the data points being studied.
- Traders and analysts can use the least squares method to identify trading opportunities and economic or financial trends.

BAYESIAN LINEAR REGRESSION

Bayesian Linear Regression

- Bayesian Linear Regression is the Bayesian interpretation of linear regression.
- in linear regression we chose values for the bias that minimized our mean squared error cost function.

Bayesian Approach:

- In Bayesian approach we don’t work with exact values but with probabilities.
- In nearly all real-world situations, our data and knowledge about the world is incomplete, indirect and noisy.
- Hence, uncertainty must be a fundamental part of our decision-making process.

Bayesian Linear Regression

- Linear regression is based on the assumption that the underlying data is normally distributed and that all relevant predictor variables have a linear relationship with the outcome.
- But In the real world, this is not always possible, it will follows these assumptions, Bayesian regression could be the better choice.
- Bayesian regression employs prior belief or knowledge about the data to “learn” more about it and takes into account the data’s uncertainty and create more accurate predictions.

Bayesian Linear Regression

- Bayesian regression uses a Bayes algorithm to estimate the parameters of a linear regression model from data, including prior knowledge about the parameters.
- Because of its probabilistic character, it can produce more accurate estimates for regression parameters than ordinary least squares (OLS) linear regression
- Bayesian regression is a type of linear regression that uses Bayesian statistics to estimate the unknown parameters of a model.
- Bayesian Regression can be very useful when we have insufficient data in the dataset or the data is poorly distributed.

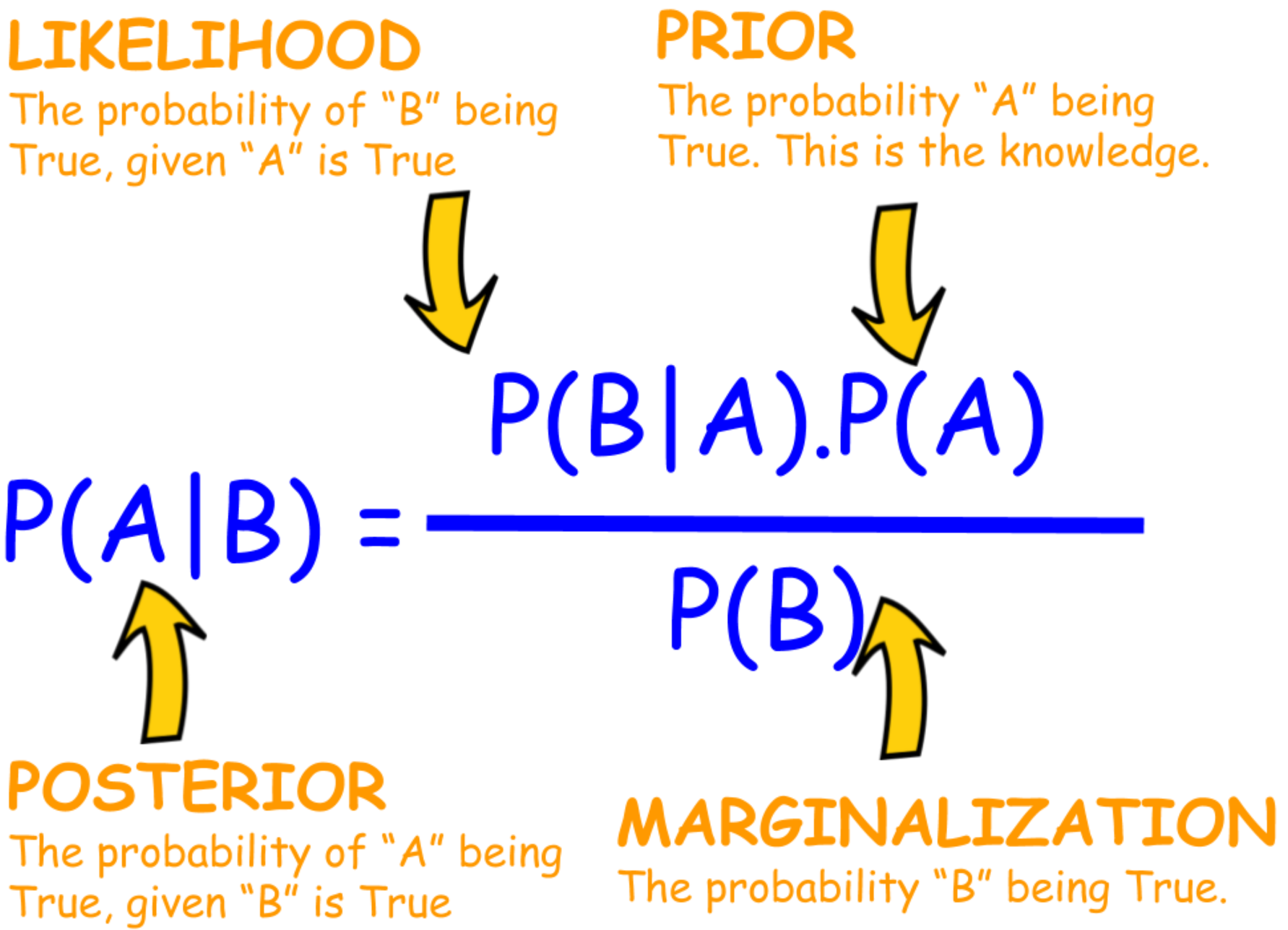
Bayesian Linear Regression

- It uses Bayes' theorem to estimate the likelihood of a set of parameters given observed data.
- The goal of Bayesian regression is to find the best estimate of the parameters of a linear model that describes the relationship between the independent and the dependent variables.
- The main difference between traditional linear regression and Bayesian regression is the underlying assumption regarding the data-generating process.
- Traditional linear regression assumes that data follows a Gaussian or normal distribution, while Bayesian regression has stronger assumptions about the nature of the data and puts a prior probability distribution on the parameters.

Bayes Theorem

- Bayes Theorem gives the relationship between an event's prior probability and its posterior probability after evidence is taken into account.
- It states that the conditional probability of an event is equal to the probability of the event given certain conditions multiplied by the prior probability of the event, divided by the probability of the conditions.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$


LIKELIHOOD
The probability of "B" being True, given "A" is True
PRIOR
The probability "A" being True. This is the knowledge.
P(A|B) = P(B|A).P(A) / P(B)
POSTERIOR
The probability of "A" being True, given "B" is True
MARGINALIZATION
The probability "B" being True.

Priori Probability

- Priori Probability is the initial probability of an event occurring before any new data is taken into account.
- P(Ei) is the priori probability of hypothesis Ei.

Posterior Probability

- Posterior Probability is the updated probability of an event after considering new information.
- Probability P(Ei|A) is considered as the posterior probability of hypothesis Ei

Maximum Likelihood Estimation

- **Maximum Likelihood Estimation** is a method used to estimate the parameters of a statistical model by **maximizing the likelihood function**.
- It seeks to find the parameter values that make the observed data most probable under the assumed model.
- Maximum Likelihood Estimation does not incorporate any **prior information** or assumptions about the parameters, and it **provides point estimates of the parameters.**

Maximum A Posteriori (MAP) Estimation

- MAP estimation is a Bayesian approach that combines prior information with the likelihood function to estimate the parameters.
- It involves finding the parameter values that maximize the posterior distribution, which is obtained by applying Bayes' theorem.
- In MAP estimation, a prior distribution is specified for the parameters, representing prior beliefs or knowledge about their values.

Maximum A Posteriori (MAP) Estimation

- The likelihood function is then multiplied by the prior distribution to obtain the joint distribution, and the parameter values that maximize this joint distribution are selected as the MAP estimates.
- MAP estimation provides point estimates of the parameters, similar to MLE, but incorporates prior information.

GRADIENT DESCENT

Gradient Descent

- Gradient descent (GD) is an iterative first-order optimisation algorithm, used to find a local minimum/maximum of a given function.
- This method is commonly used in machine learning (ML) and deep learning (DL) to minimise a cost/loss function (e.g. in a linear regression).
- Gradient descent algorithm does not work for all functions.

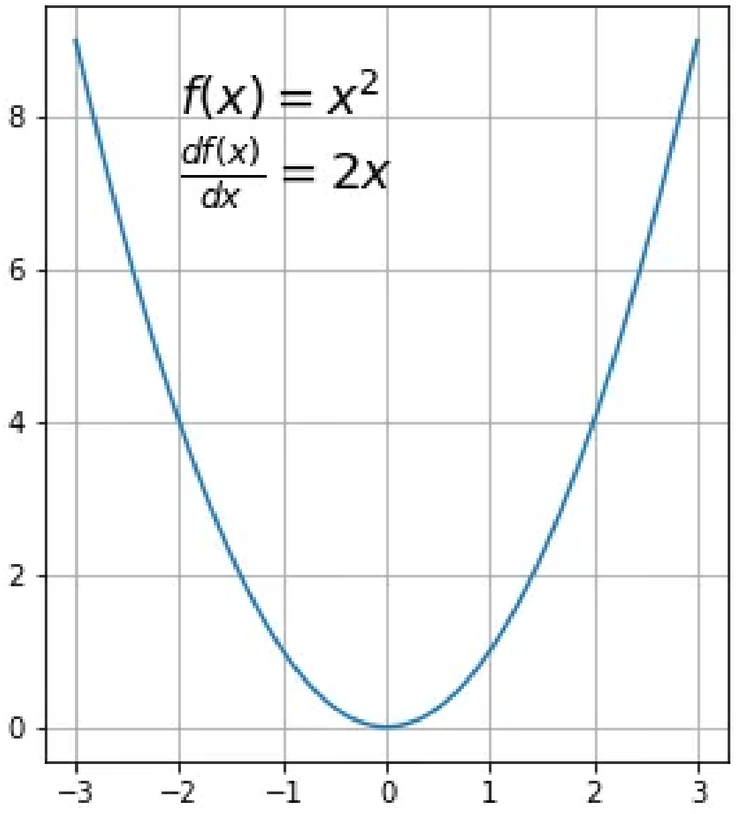
Gradient Descent

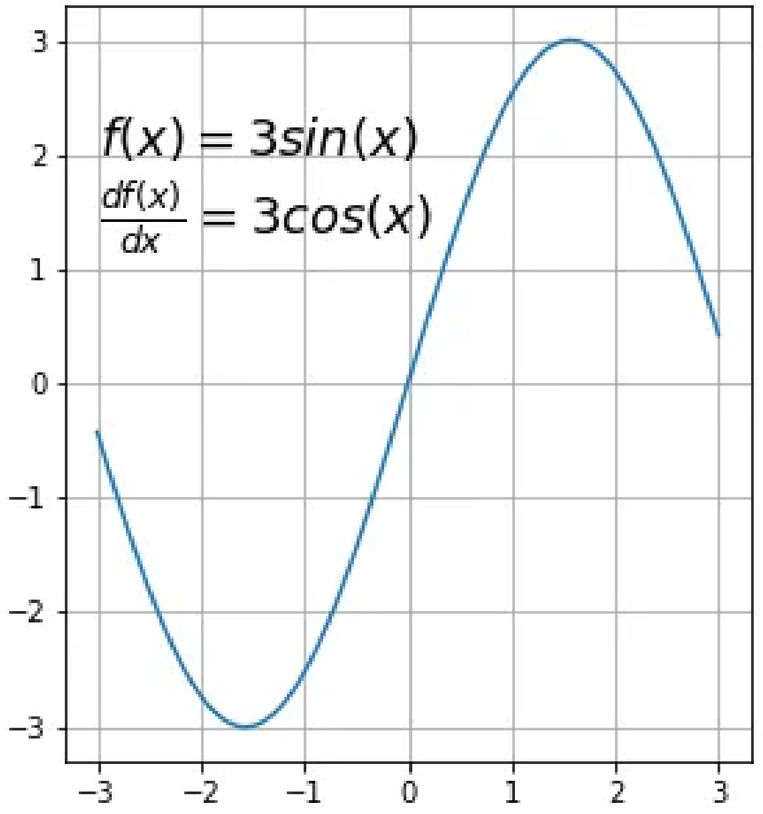
- There are two specific requirements.
- A function has to be:
    - Differentiable
    - Convex

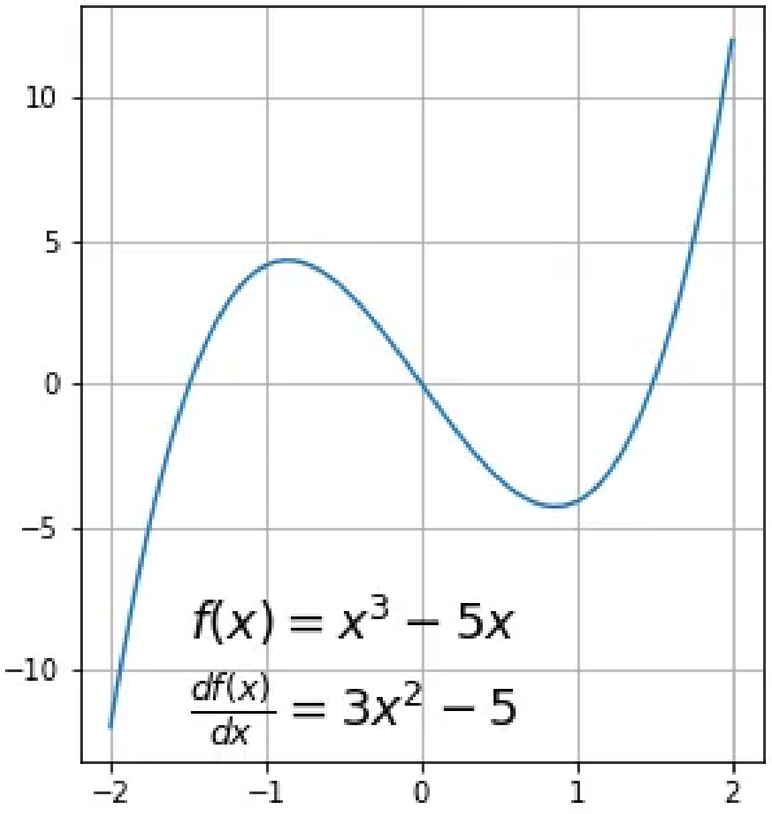
1) If a function is differentiable it has a derivative for each point in its domain

- Not all functions meet these criteria

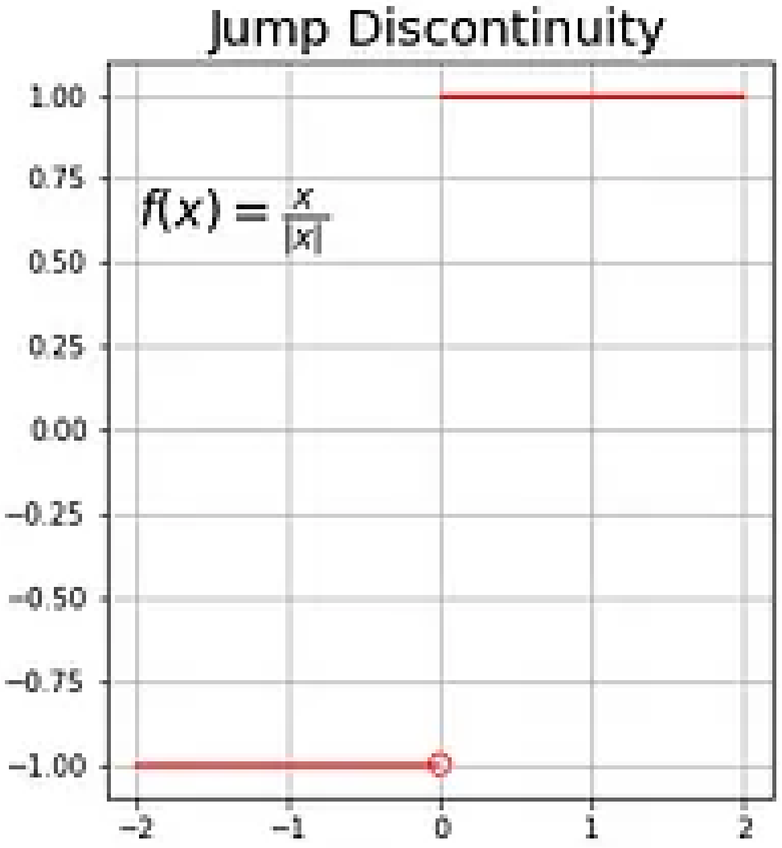
Differentiable functions

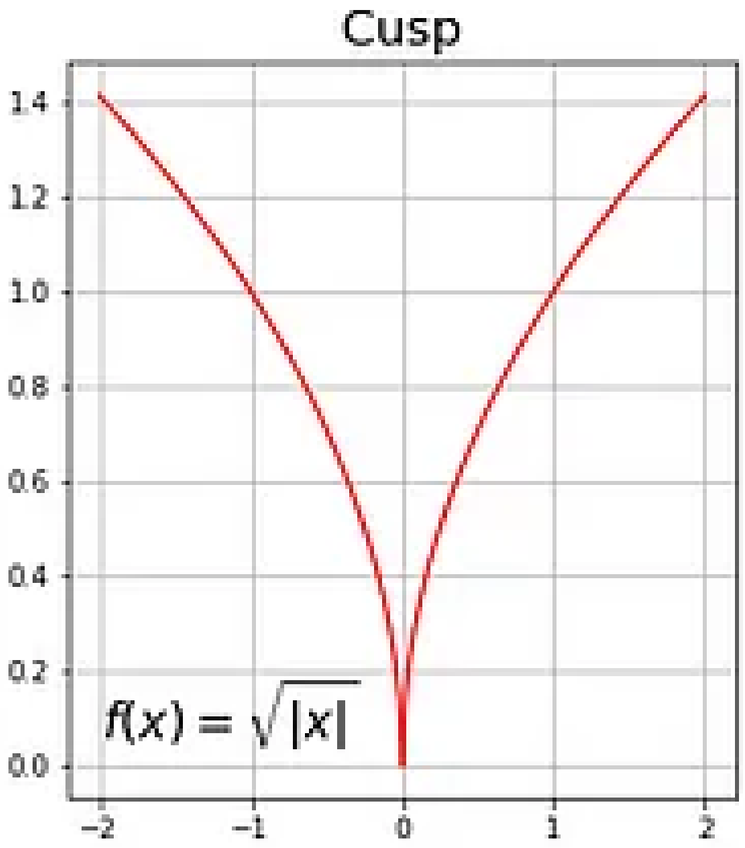


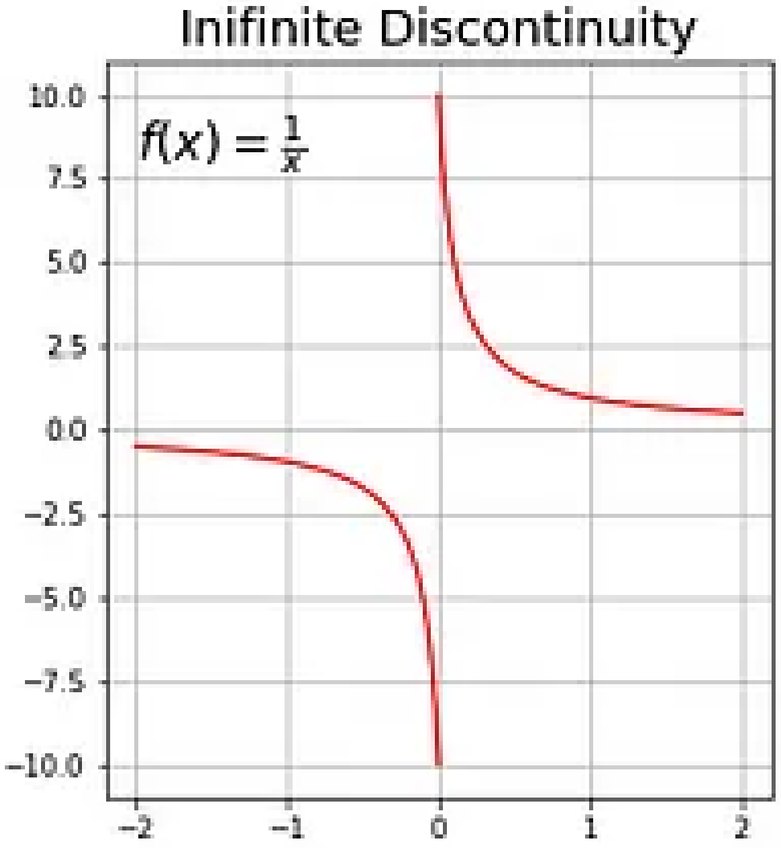




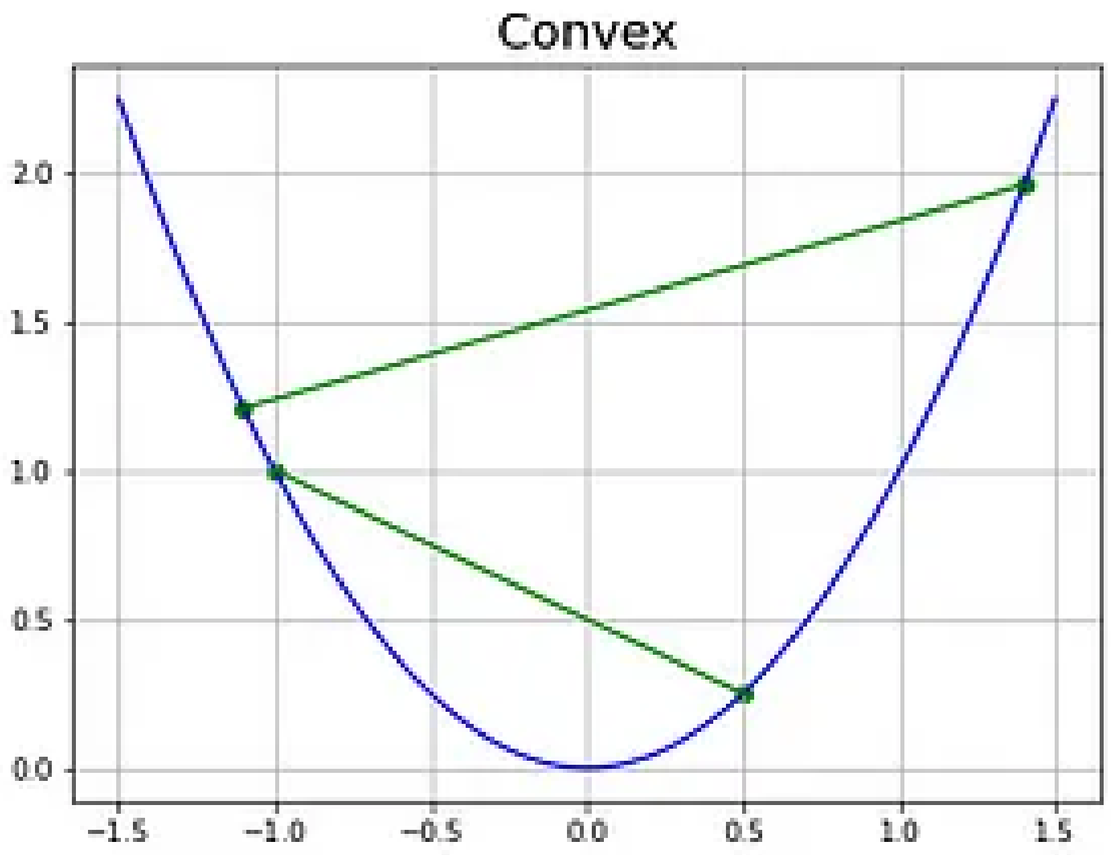
Non-Differentiable functions

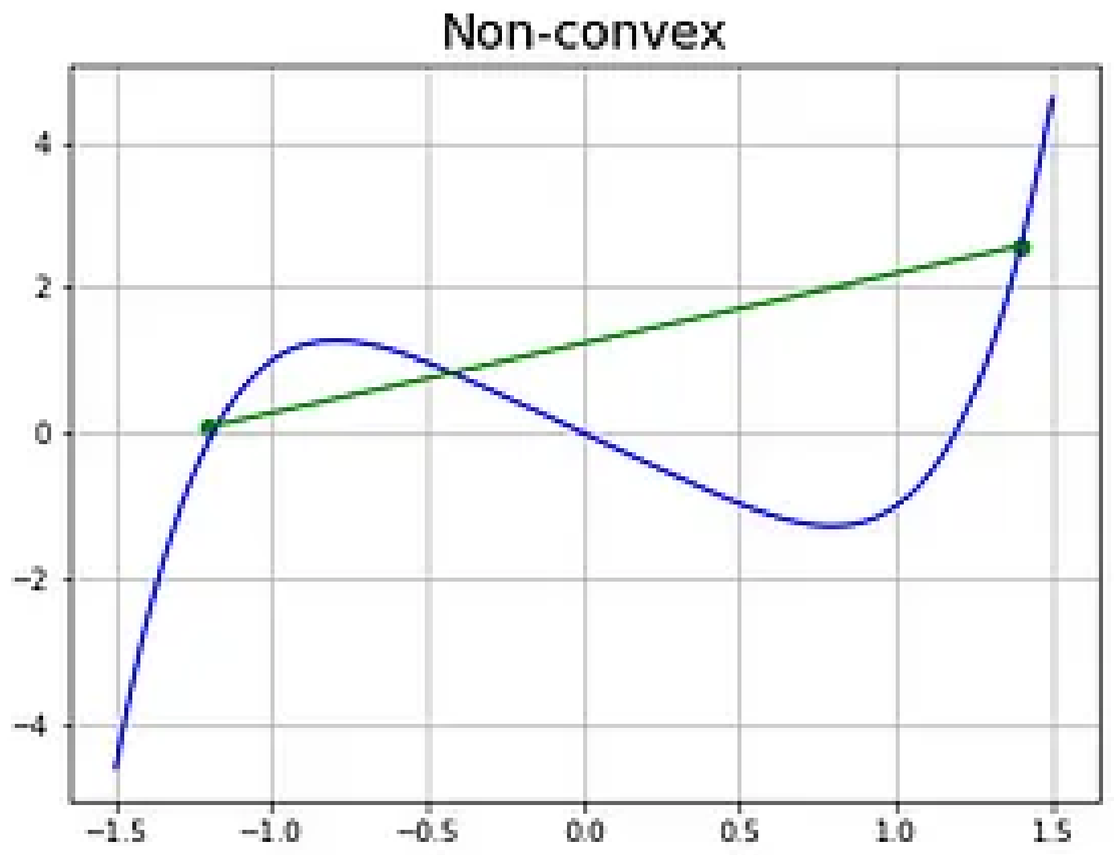






Convex & Non-Convex functions





Gradient

- **Consider a two dimensional function:**

$$f(x) = 0.5x^2 + y^2$$

- **Let's assume we are interested in a gradient at point p(10, 10)**

$$\frac{\partial f(x,y)}{\partial x} = x, \quad \frac{\partial f(x,y)}{\partial y} = 2y$$

$$\nabla f(x,y) = \begin{bmatrix} x \\ 2y \end{bmatrix} \qquad \nabla f(10,10) = \begin{bmatrix} 10 \\ 20 \end{bmatrix}$$

Gradient Descent Algorithm

- Gradient Descent Algorithm iteratively calculates the next point using gradient at the current position, scales it (by a learning rate) and subtracts obtained value from the current position (makes a step).
- It subtracts the value because we want to minimize the function.

$$p_{n+1} = p_n - \eta \nabla f(p_n)$$

Gradient Descent Algorithm

- ‘η’ (Learning rate) is the important parameter which scales the gradient and thus controls the step size.
- Learning rate have a strong influence on performance.
- The smaller learning rate the longer GD converges, or may reach maximum iteration before reaching the optimum point
- If learning rate is too big the algorithm may not converge to the optimal point (jump around) or even to diverge completely.

Gradient Descent Algorithm

1. choose a starting point (initialisation)
2. calculate gradient at this point
3. make a scaled step in the opposite direction to the gradient (objective: minimise)
4. repeat points 2 and 3 until one of the criteria is met:
    - maximum number of iterations reached
    - step size is smaller than the tolerance (due to scaling or a small gradient).

- Tolerance - to conditionally stop the algorithm