Machine Learning

Linear Regression

Linear Regression

- Linear regression computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation with observed data.
- It predicts the continuous output variables based on the independent input variable.

Linear Regression

- **Linear regression** is also a type of supervised ML algorithm.
- It learns from the labelled datasets and maps the data points with most optimized linear functions which can be used for prediction on new datasets.

Linear Regression

- For example:
- Predict house price considering various factor such as house age, distance from the main road, location, area and number of room
- Linear regression uses all these parameter to predict house price as it consider a linear relation between all these features and price of house.

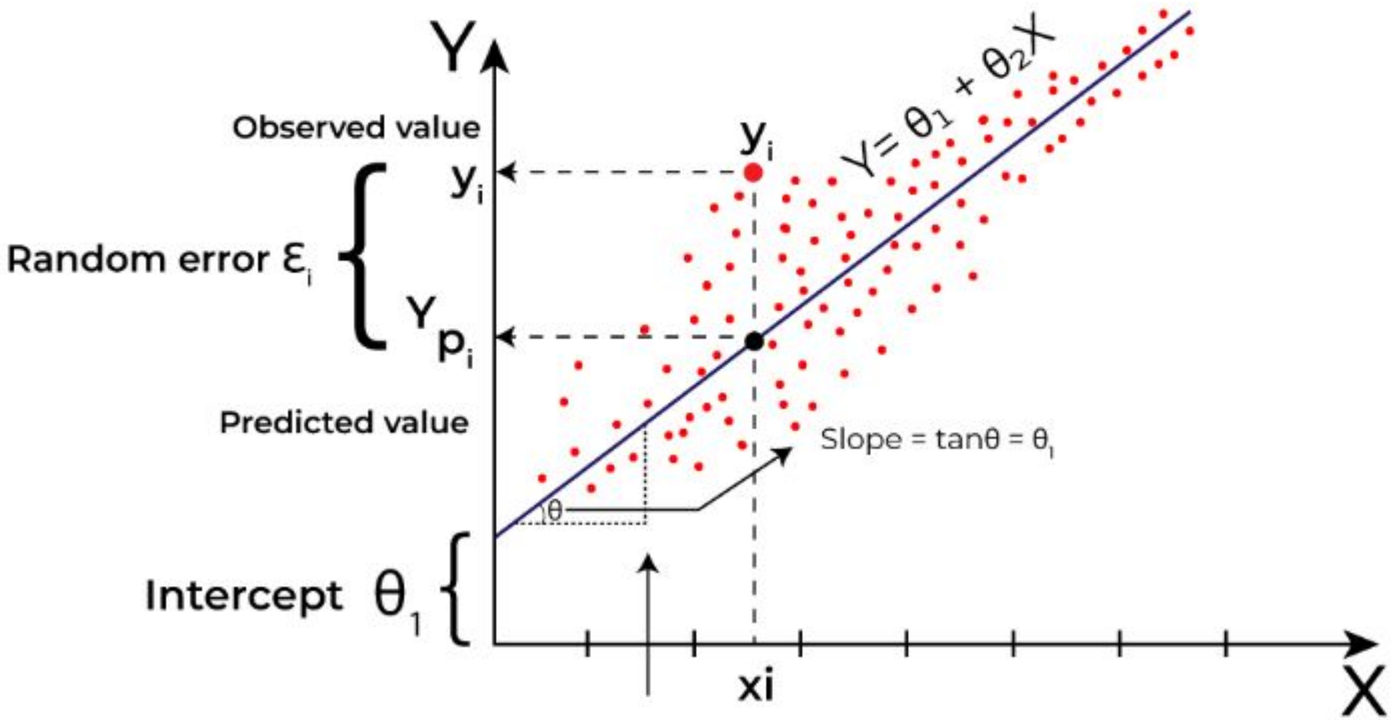
Best Fit Line

- Primary objective while using linear regression is to locate the best-fit line, which implies that the error between the predicted and actual values should be kept to a minimum.
- There will be the least error in the best-fit line.

Best Fit Line

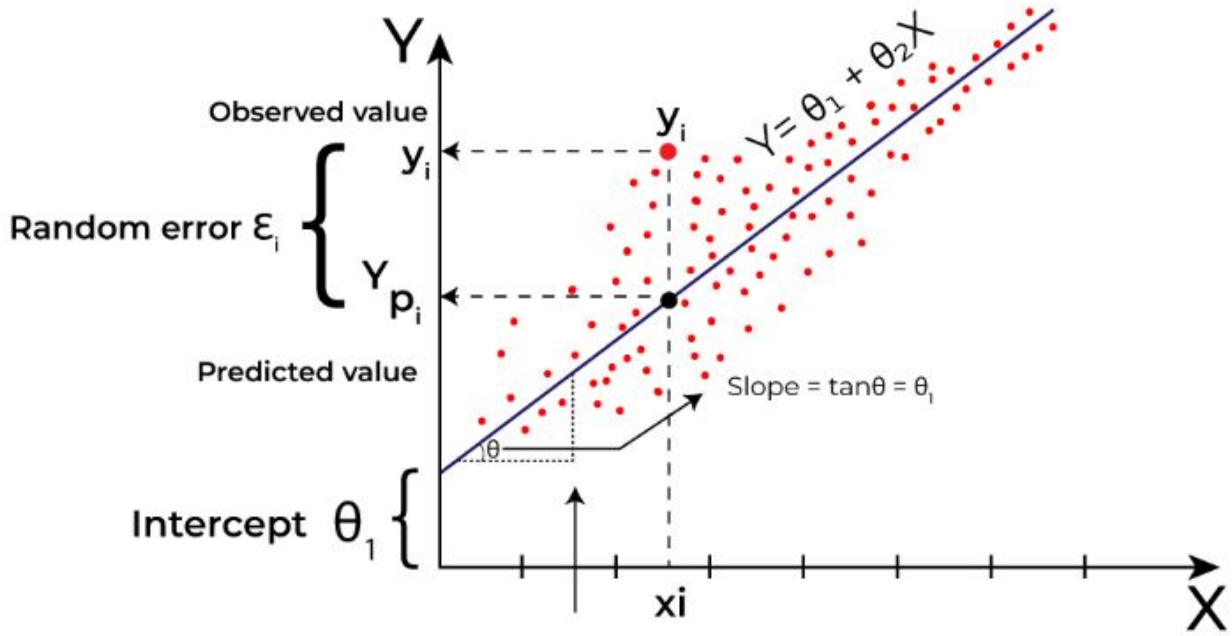
- The best Fit Line equation provides a straight line that represents the relationship between the dependent and independent variables.
- The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).

Best Fit Line



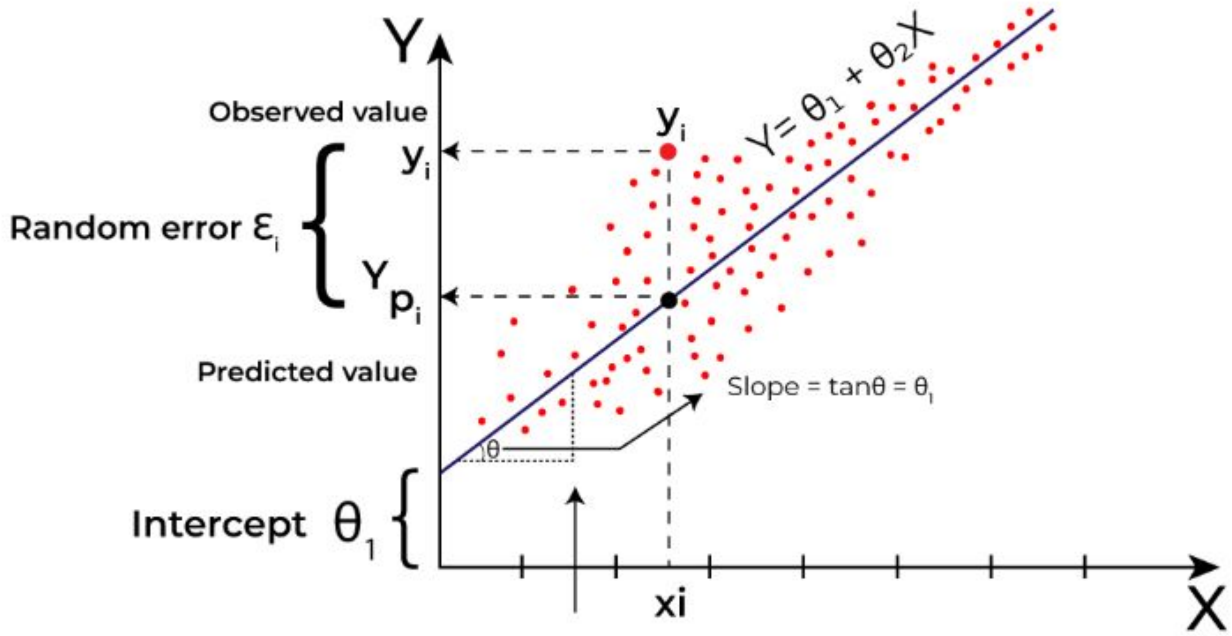
Best Fit Line

- Here Y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y.



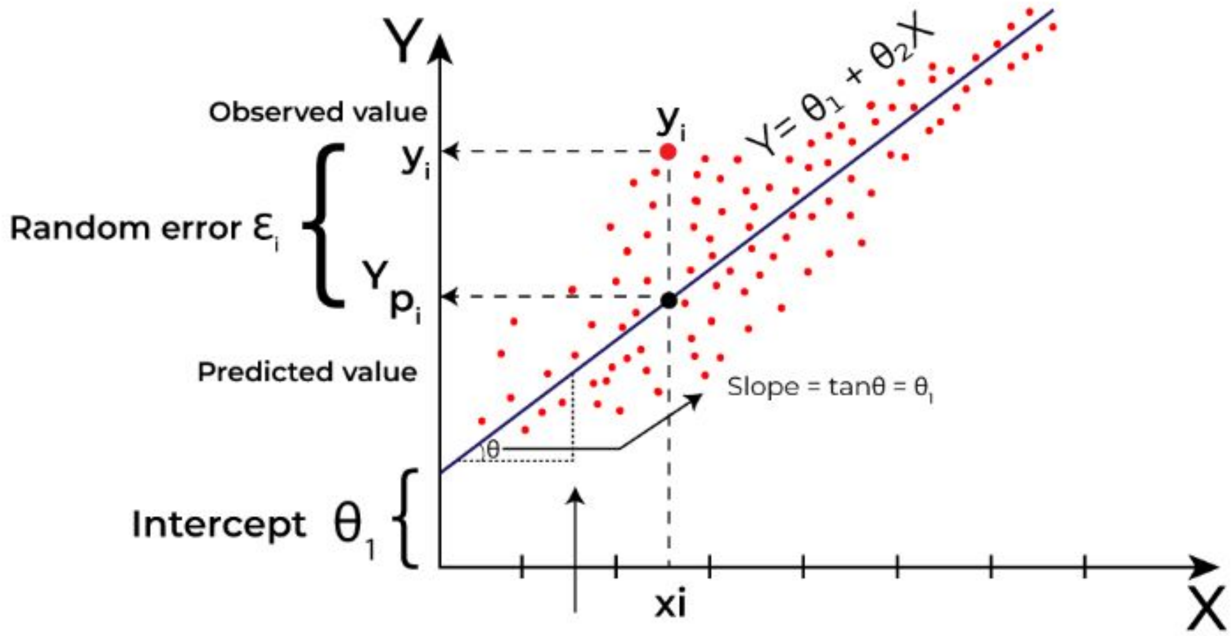
Best Fit Line

- There are many types of functions or modules that can be used for regression.
- A linear function is the simplest type of function.
- Here, X may be a single feature or multiple features representing the problem.



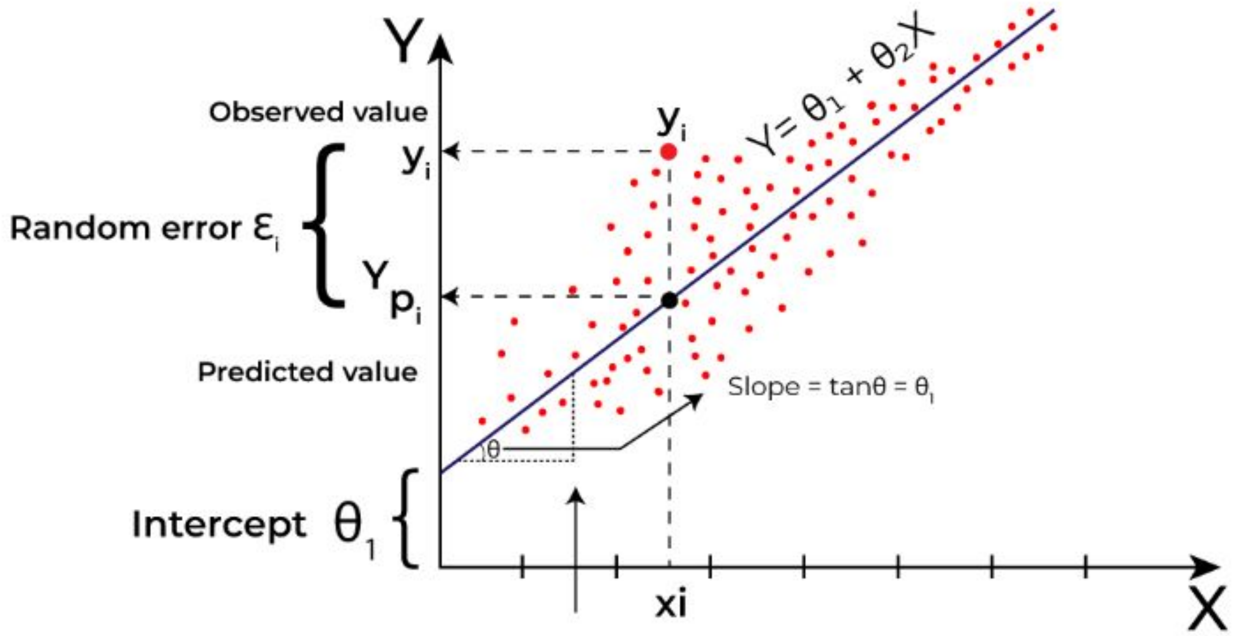
Best Fit Line

- Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)).
- Hence, the name is Linear Regression.



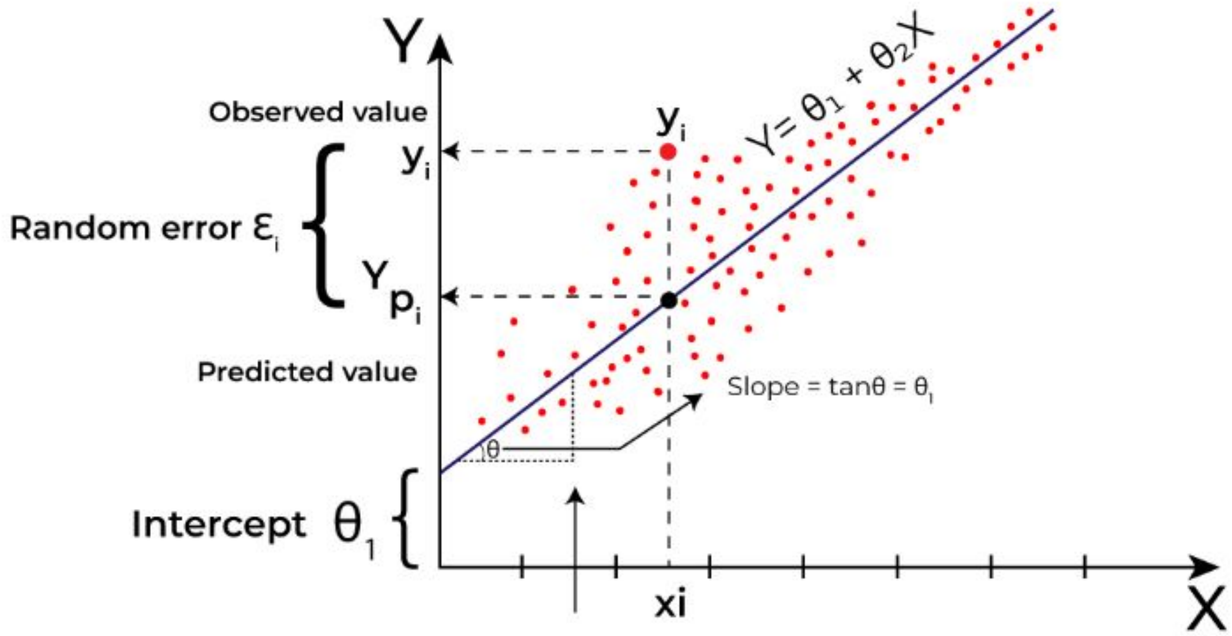
Best Fit Line

- In the figure, X (input) is the work experience and Y (output) is the salary of a person.
- The regression line is the best-fit line for the model.



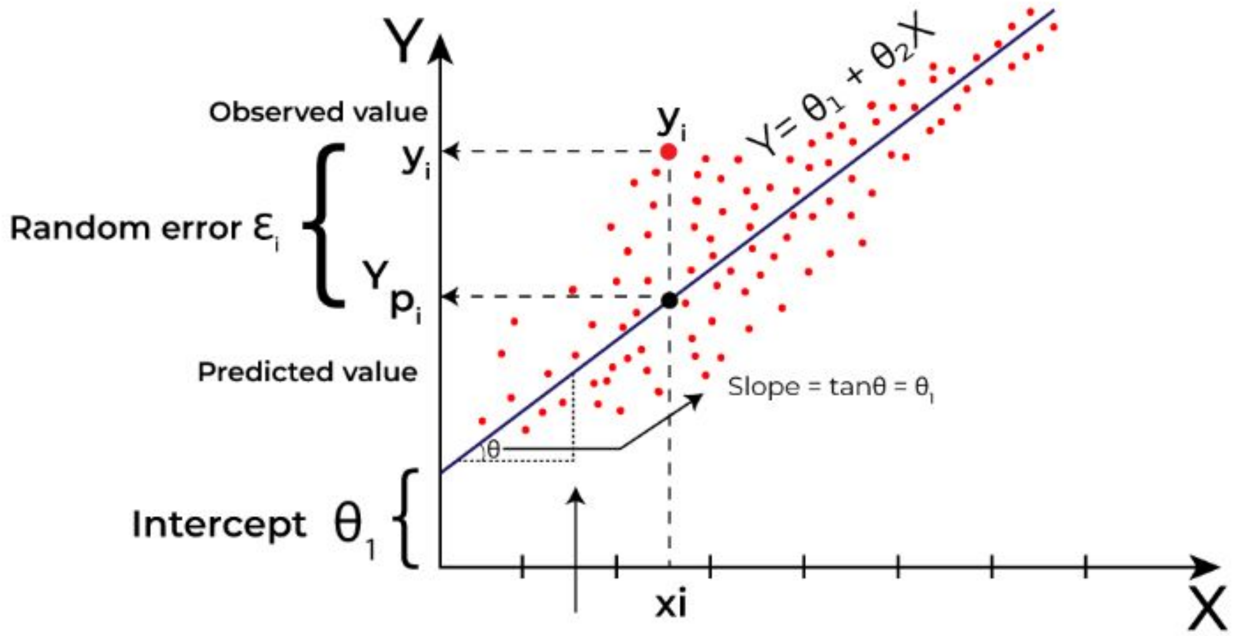
Best Fit Line

- Let's assume there is a linear relationship between X and Y, then the salary can be predicted using:
- $\hat{Y} = \theta_1 + \theta_2 X$
- OR
- $\hat{y}_i = \theta_1 + \theta_2 x_i$
- Here,
- $y_i \in Y (i = 1, 2, \ldots, n)$ are labels to data (Supervised learning)
- $x_i \in X (i = 1, 2, \ldots, n)$ are the input independent training data (univariate – one input variable(parameter))
- $\hat{y}_i \in \hat{Y} (i = 1, 2, \ldots, n)$ are the predicted values.



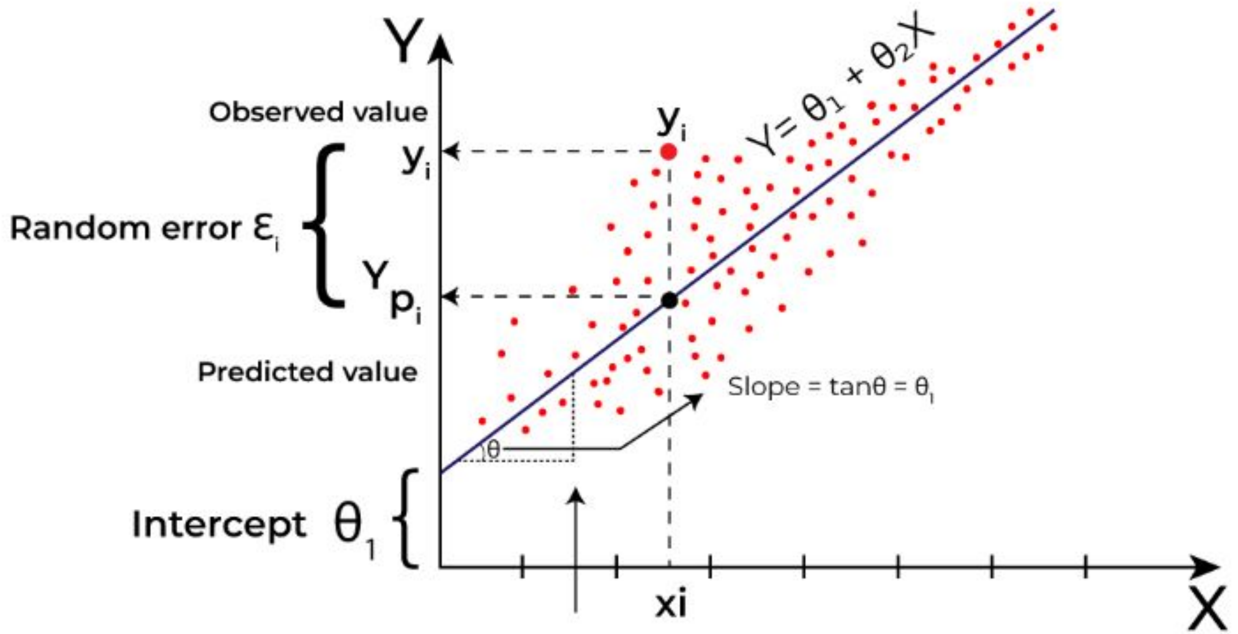
Best Fit Line

- The model gets the best regression fit line by finding the best θ_1 and θ_2 values.
  - θ_1: intercept
  - θ_2: coefficient of x
- Once we find the best θ_1 and θ_2 values, we get the best-fit line.
- So when we are finally using our model for prediction, it will predict the value of y for the input value of x.



Best Fit Line

- To achieve the best-fit regression line, the model aims to predict the target value $\hat{Y}$ such that the error difference between the predicted value $\hat{Y}$ and the true value Y is minimum.
- So, it is very important to update the θ_1 and θ_2 values, to reach the best value that minimizes the error between the predicted $\hat{y}$ value (pred) and the true y value.
- $minimize \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$



Types of Linear Regression

- **Simple Linear Regression** or Univariate Linear Regression
- **Multiple Linear Regression** or Multivariate Regression

Simple Linear Regression

- Simple linear regression is the simplest form of linear regression
- It involves only one independent variable and one dependent variable.
- The equation for simple linear regression is:

$$\mathrm{y} = \theta_1 + \theta_2 x$$

where:

y is the dependent variable

x is the independent variable

θ_1 is the intercept

θ_2 is the slope

Simple Linear Regression

- Use cases:
- To examine the relationship between study hours and exam scores.
- To forecast sales based on historical data particularly examining how factors like advertising expenditure influence revenue.

Multiple Linear Regression

- Multiple linear regression involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

  $$\mathrm{y} = \theta_1 + \theta_2 x_1 + \theta_3 x_2 + \cdots + \theta_n x_3$$

  where:

  y is the dependent variable

  $x_1, x_2, \ldots, x_n$ are the independent variables

  θ_1 is the intercept

  θ_2 are the slopes

Multiple Linear Regression

- **Real Estate Pricing:** In real estate MLR is used to predict property prices based on multiple factors such as location, size, number of bedrooms, etc. This helps buyers and sellers understand market trends and set competitive prices.
- **Financial Forecasting:** Financial analysts use MLR to predict stock prices or economic indicators based on multiple influencing factors such as interest rates, inflation rates and market trends. This enables better investment strategies and risk management.

Multiple Linear Regression

- **Agricultural Yield Prediction:** Farmers can use MLR to estimate crop yields based on several variables like rainfall, temperature, soil quality and fertilizer usage. This information helps in planning agricultural practices for optimal productivity
- **E-commerce Sales Analysis:** An e-commerce company can utilize MLR to assess how various factors such as product price, marketing promotions and seasonal trends impact sales.

Cost function for Linear Regression

- The difference between the predicted value $\hat{Y}$ and the true value Y and it is called cost function or the loss function.

Cost function for Linear Regression

- In Linear Regression, the **Mean Squared Error (MSE)** cost function is employed
- MSE calculates the average of the squared errors between the predicted values $\hat{y}_i$ and the actual values y_i.
- The purpose is to determine the optimal values for the intercept θ_1 and the coefficient of the input feature θ_2 providing the best-fit line for the given data points. The linear equation expressing this relationship is $\hat{y}_i = \theta_1 + \theta_2 x_i$

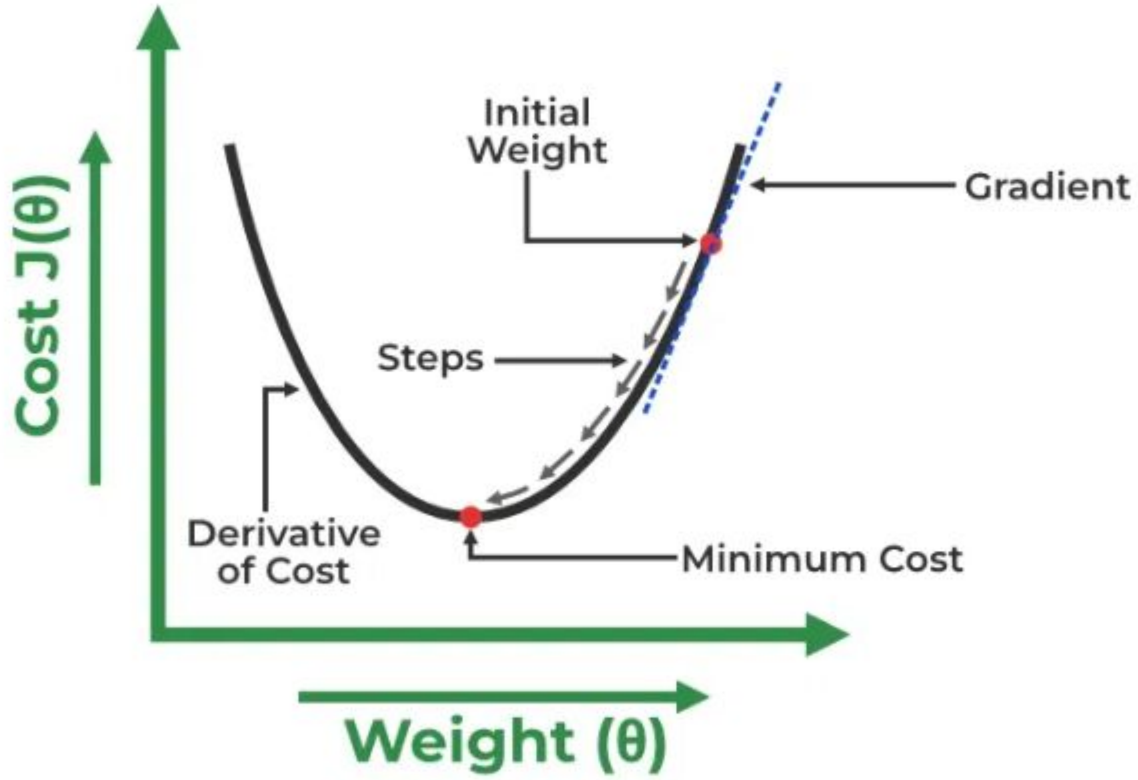
Cost function for Linear Regression

- MSE function can be calculated as:
  - Cost function $(J) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$
- Utilizing the MSE function, the iterative process of gradient descent is applied to update the values of θ_1& θ_2.
- This ensures that the MSE value converges to the global minima, signifying the most accurate fit of the linear regression line to the dataset.

Cost function for Linear Regression

- A linear regression model can be trained using the optimization algorithm **gradient descent** by iteratively modifying the model's parameters to reduce the mean squared error (MSE) of the model on a training dataset.
- To update θ_1 and θ_2 values in order to reduce the Cost function (minimizing RMSE value) and achieve the best-fit line the model uses Gradient Descent.
- The idea is to start with random θ_1 and θ_2 values and then iteratively update the values, reaching minimum cost.

Cost function for Linear Regression



Regularization Techniques for Linear Models

- Lasso Regression (L1 Regularization)
- Ridge Regression (L2 Regularization)
- Elastic Net Regression

Lasso Regression (L1 Regularization)

- L1 regularization is a technique used for regularizing a linear regression model,
- It adds a penalty term to the linear regression objective function to prevent overfitting.
- The objective function after applying lasso regression is:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 + \lambda\sum_{j=1}^{n}\left|\theta_j\right|$$

  - the first term is the least squares loss, representing the squared difference between predicted and actual values.
  - the second term is the L1 regularization term, it penalizes the sum of absolute values of the regression coefficient θ_j.

Ridge Regression (L2 Regularization)

- Ridge regression is a linear regression technique that adds a regularization term to the standard linear objective.
- Again, the goal is to prevent overfitting by penalizing large coefficient in linear regression equation.
- It is useful when the dataset has multicollinearity where predictor variables are highly correlated.
- The objective function after applying ridge regression is:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 + \lambda\sum_{j=1}^{n}\theta_j^2$$

  - the first term is the least squares loss, representing the squared difference between predicted and actual values.
  - the second term is the L1 regularization term, it penalizes the sum of square of values of the regression coefficient θ_j.

Elastic Net Regression

- L1 and L2 regularization is a hybrid regularization technique that combines the power of both L1 and L2 regularization in linear regression objective.

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 + \alpha\lambda\sum_{j=1}^{n}\left|\theta_j\right| + \frac{1}{2}(1-\alpha)\lambda\sum_{j=1}^{n}\theta_j^2$$

  - the first term is least square loss.
  - the second term is L1 regularization and third is ridge regression.
  - λ is the overall regularization strength.
  - α controls the mix between L1 and L2 regularization.

Logistic Regression

Logistic Regression

- **Logistic regression** is a **supervised machine learning algorithm** used for **classification tasks**.
- The goal is to predict the probability that an instance belongs to a given class or not.

Logistic Regression

- Logistic regression is used for binary classification where we use *sigmoid function*, that takes input as independent variables and produces a probability value between 0 and 1.
- For example, we have two classes Class 0 and Class 1 if the value of the logistic function for an input is greater than 0.5 (threshold value) then it belongs to Class 1 otherwise it belongs to Class 0.
- It’s referred to as regression because it is the extension of linear regression but is mainly used for classification problems.

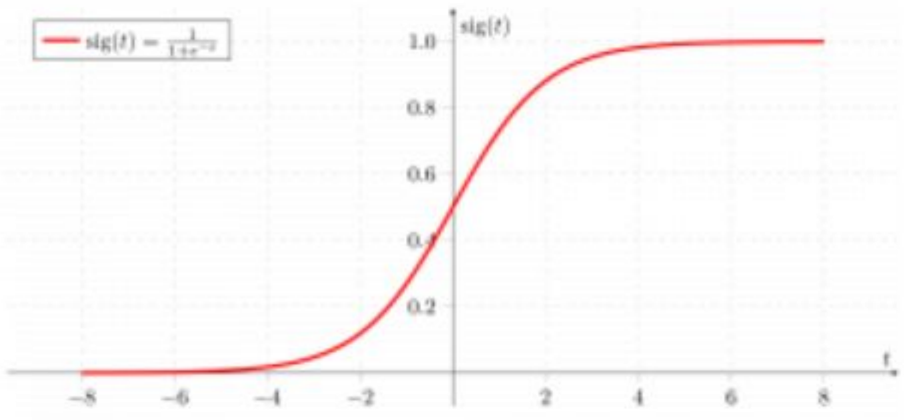
Logistic Regression

- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value.
- It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

Logistic Function – Sigmoid Function

- The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

$$\sigma(z) = \frac{1}{1+e^{-z}}$$



Sigmoid function

Linear regression equation

- Using Analytical Method
  - Used for simple linear regression

Linear regression equation

- Y on X= Variation on Y when the changes in X is given

$$(\mathrm{y} - \bar{y}) = \frac{\sum xy}{\sum x^2}(x - \bar{x})$$

Linear regression equation

- For example:
- Data Points: (1, 2), (2, 5), (3, 11), (4, 8), (5, 14)

Linear regression equation

1	-2	4	2	-6	36	12
2	-1	1	5	-3	9	3
3	0	0	11	3	9	0
4	1	1	8	0	0	0
5	2	4	14	6	36	12

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

$$\Rightarrow (y - 8) = \frac{27}{10}(x - 3) \Rightarrow (y - 8) = 2.7(x - 3) \Rightarrow (y - 8) = 2.7x - 8.1 \Rightarrow y = 2.7x - 8.1 + 8$$

$$\Rightarrow y = 2.7x - 0.1$$