# Linear Regression

## What is Linear Regression?

Regression analysis is one of the most widely used methods for prediction. Linear regression is probably the most fundamental machine learning method out there and a starting point for the advanced analytical learning path of every aspiring data scientist.

Linear Regression: A supervised learning algorithm used for regression tasks, where the goal is to predict a continuous value based on input features. It models the relationship between the independent variables and the dependent variable using a linear equation.

independent variables ↓ input feature

dependent variable ↓ output feature

Get sample data → Design a Model → Make predictions about the whole population

Supervised ML → Regression → O/P → continous

Supervised ML → Classification → O/P → category

linear + Regression → O/P → continous

linear ↓ line ↓ straight line

Independent feature ↓ Size of House

Dependent feature ↓ → O/P Price (predict)

# Simple Linear Regression

Input feature ⟶ Output feature

(Independent) (Dependent)

| Size of House | Price |
| --- | --- |
| 2 BHK | 40 lac |
| 3 BHK | 45 lac |
| 4 BHK | 60 lac |
| — | |
| — | |
| — | |

$\xrightarrow{\text{predict} \to \text{price}}$

TRAIN DATASET

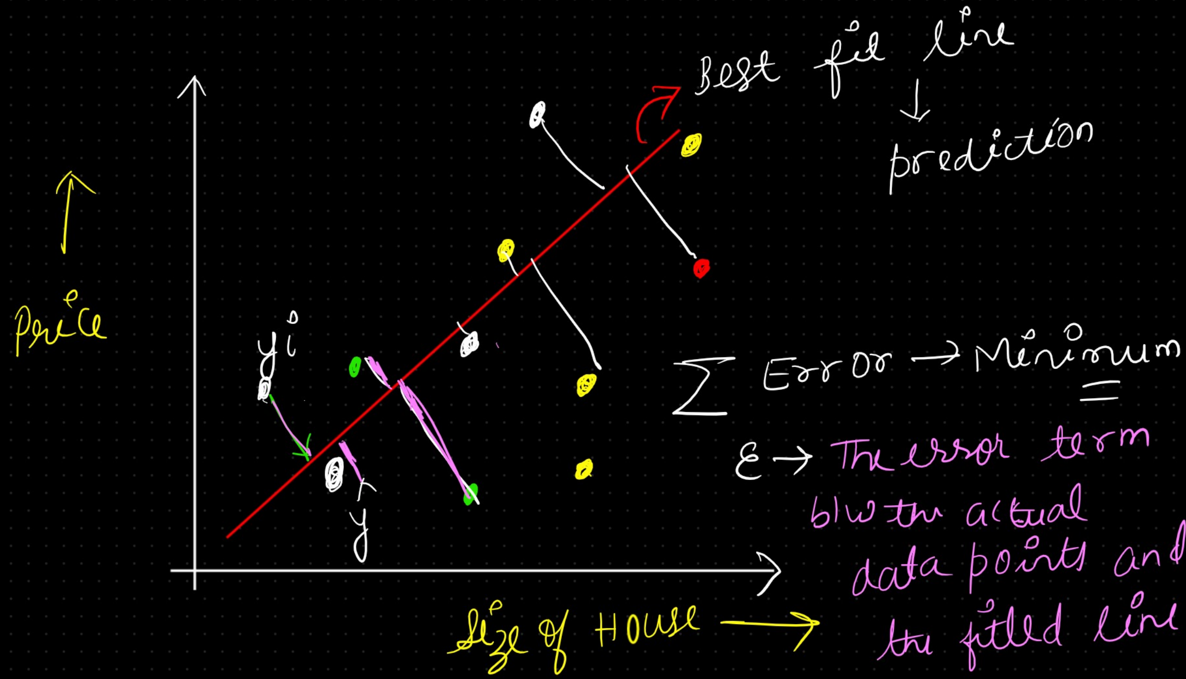
predict

↓

new size ⟶ $\boxed{\text{Model}}$ ⟶ Price

↓

ACC ↑↑




linear Equation $\boxed{\hat{y} = mx + c}$

$\hat{y}$ = predicted variable

$m$ = slope

$c$ = intercept

$x$ = Independent variable

$$\boxed{\hat{y} = \beta_0 + \beta_1 x_1}$$

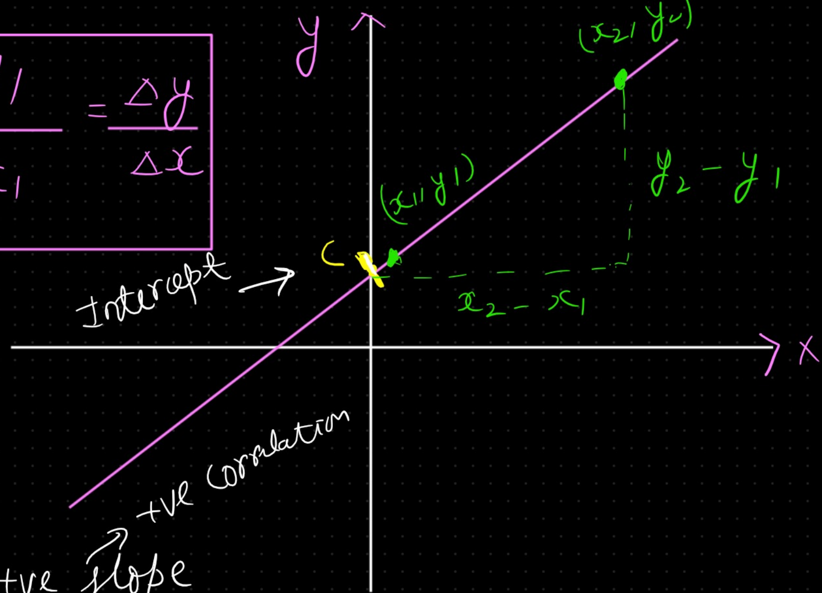
w.r.t Hypothesis Testing

$$\boxed{h_\theta(x) = \theta_0 + \theta_1 x_1}$$

Equation of Straight line

$$\boxed{y = mx + c}$$

$$\boxed{m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x}}$$




$+m < x\uparrow, y$

$y\uparrow\ x\uparrow \rightarrow$ +ve slope

$y\downarrow\ x\uparrow \rightarrow$ −ve slope → −ve correlation

## Linear Regression Model

$$\boxed{\hat{y} = \beta_0 + \beta_1 x + \varepsilon}$$

↓ Model

$\hat{y}$ = The predicted variable

$\beta_0$ = the intercept (Y axis value when X = 0

$\beta_1$ = The slope (how much Y changes of a unit change in x)

$\varepsilon$ = Error → $\Sigma\varepsilon$ → minimum

slope/coefficient

# Finding the Best-Fitting line

The goal is to determine the values of $\beta_0$ and $\beta_1$, that minimize the sum of squared error (SSE)

$$\boxed{SSE = \sum (y_i - \hat{y}_i)^2}$$

$$\boxed{\hat{y} = h_\theta(x) = \theta_0 + \theta_1 x_1}$$

## Cost Function

$$(\theta_0, \theta_1) = \frac{1}{n} \sum_{i=1}^{n} (\underbrace{y_i}_{\text{Actual}} - \underbrace{h_\theta(x)_i}_{\text{predicted} \; \hat{y}_i})^2$$

$$\boxed{(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \downarrow\downarrow$$

Minimum

[Mean squared Error]

$n$ = No. of data points

$y_i$ = Actual value

$\hat{y}_i = h_\theta(x) =$ predicted value

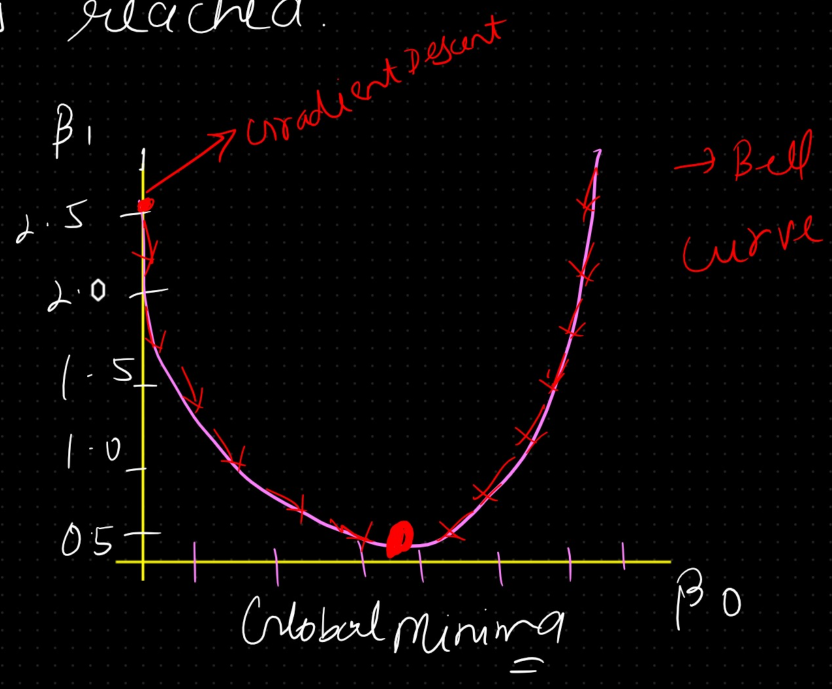
Aim: [In order to get the Best Fit line]

Simple Linear Regression is a fundamental algorithm in machine learning used for predictive analysis. It's based on the relationship between two continuous variables: one independent variable (X) and one dependent variable (Y). The algorithm aims to find the best-fitting straight line that represents the relationship between these variables.

$\beta_0$, $\beta_1$

② Gradient Descent

→ This is an iterative optimization technique that repeatedly adjusts $\beta_0$ and $\beta_1$ to decrease the SSE until a minimum is reached.



| Size of House ($x$) | Price ($y$) |
| --- | --- |
| 2BHK | 40 lac |
| 3BHK | 50 lac |
| 4BHK | 60 lac |

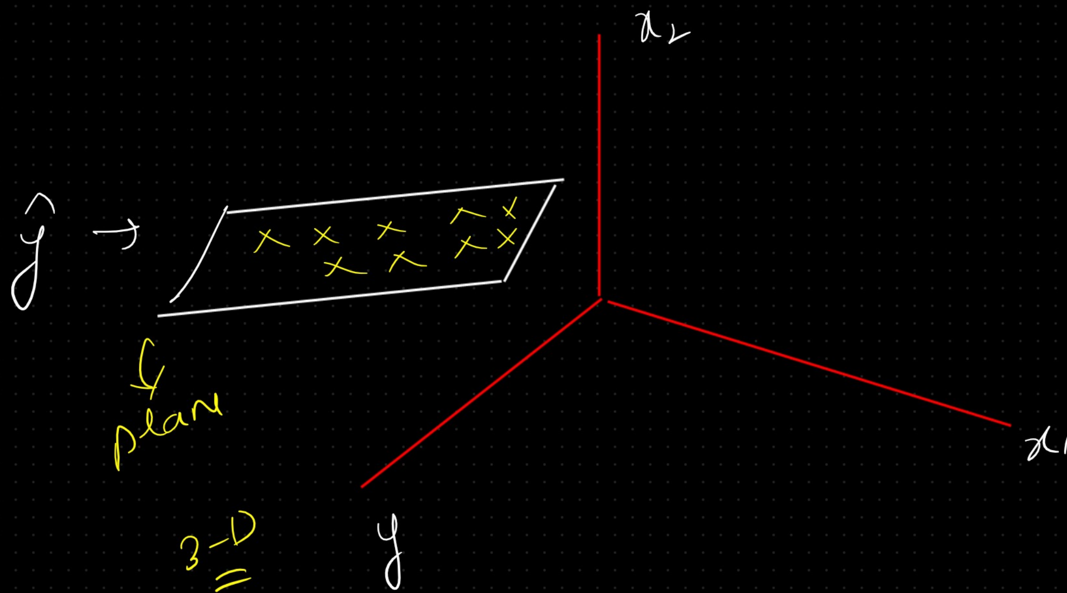
Conclusion:
In simple terms, simple linear regression finds the line that best fits the data by minimizing the vertical distances between the line and the data points. It allows us to make predictions about the dependent variable based on the values of the independent variable.

## II) Multiple Linear Regression

| No of Rooms ($x_1$) | Size ($x_2$) | Price ($y$) |
| --- | --- | --- |

Two Independent variable

Dependent variable




$$\boxed{\hat{y} = h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2}$$

$\theta_0$ = Intercept

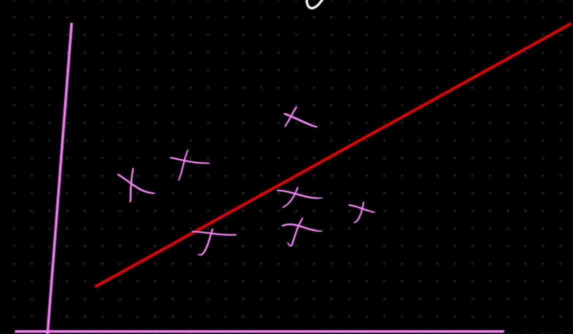
$\theta_1, \theta_2$ = slope

III)

## Polynomial Regression

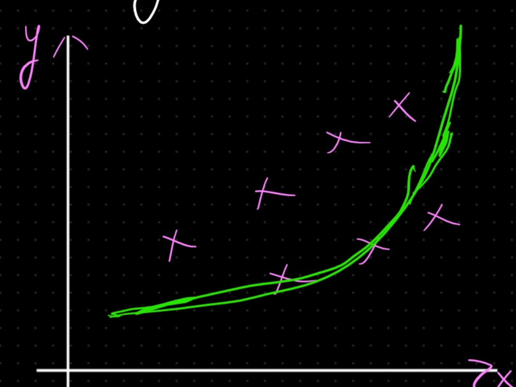
(Non-linear)

Linear Regression



$\hat{y} = b_0 + b_1 x$

Polynomial




$\hat{y} = b_0 + b_1 x_1^1 + b_2 x_1^2$

Simple linear Regression

$$\hat{y} = \beta_0 + \beta_1 x_1$$

Multiple linear Regression

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \\ + \cdots + \beta_n x_n$$

polynomial degree (1)

$$= \beta_0 + \beta_1 x_1$$

simple linear regression $= \beta_0 + \beta_1 x_1^1 =$

degree of polynomial = 2

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

degree = 3

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3$$

Polynomial regression is a form of regression analysis used when the relationship between the independent variable $x$ and the dependent variable $y$ is believed to follow a polynomial rather than a linear function. In polynomial regression, the relationship is modeled as an nth degree polynomial. The general form of a polynomial regression model of degree $n$ is given by:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_n x^n + \varepsilon$$

Overall, polynomial regression is a flexible approach that can capture nonlinear relationships between variables, making it useful in various fields such as economics, engineering, and the natural sciences. However, careful consideration must be given to model selection and validation to ensure robust and reliable results.