**🧠 What is Linear Regression?**

**Linear Regression** is one of the simplest and most commonly used **supervised machine learning algorithms**.  
It is used to find the **relationship between one or more input variables (independent variables)** and an **output variable (dependent variable)** by fitting a **straight line** to the data.

This line is called the **Regression Line** or **Best Fit Line**.

**🎯 Goal of Linear Regression**

To find the **best-fitting straight line** through the data points such that the **difference between the predicted values and the actual values** is **minimized**.

**📈 Equation (Formula) of Linear Regression**

For **Simple Linear Regression** (one independent variable):

y=β0+β1x+εy = β\_0 + β\_1x + εy=β0​+β1​x+ε

**Where:**

* yyy → Dependent variable (the value to be predicted)
* xxx → Independent variable (input feature)
* β0β\_0β0​ → Intercept (the value of yyy when x=0x = 0x=0)
* β1β\_1β1​ → Slope (change in yyy for a unit change in xxx)
* εεε → Error term (difference between actual and predicted value)

**📊 For Multiple Linear Regression:**

If there are multiple input variables (x1,x2,...,xnx\_1, x\_2, ..., x\_nx1​,x2​,...,xn​):

y=β0+β1x1+β2x2+...+βnxn+εy = β\_0 + β\_1x\_1 + β\_2x\_2 + ... + β\_nx\_n + εy=β0​+β1​x1​+β2​x2​+...+βn​xn​+ε

**⚙️ How Linear Regression Works (Step-by-Step)**

1. **Collect Data:**  
   Gather historical data with input(s) and output(s).  
   Example: House size (sq.ft) → House price.
2. **Plot Data:**  
   Plot the data points on a graph to visualize the relationship.
3. **Find the Best Fit Line:**  
   The algorithm calculates coefficients β0β\_0β0​ and β1β\_1β1​ that minimize the total error.
4. **Use the Least Squares Method:**

The goal is to minimize the **Sum of Squared Errors (SSE):**

SSE=∑i=1n(yi−y^i)2SSE = \sum\_{i=1}^{n} (y\_i - ŷ\_i)^2SSE=i=1∑n​(yi​−y^​i​)2

where  
yiy\_iyi​ = actual value  
y^i=β0+β1xiŷ\_i = β\_0 + β\_1x\_iy^​i​=β0​+β1​xi​ = predicted value

The coefficients are found using:

β1=n(∑xy)−(∑x)(∑y)n(∑x2)−(∑x)2β\_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}β1​=n(∑x2)−(∑x)2n(∑xy)−(∑x)(∑y)​ β0=yˉ−β1xˉβ\_0 = \bar{y} - β\_1\bar{x}β0​=yˉ​−β1​xˉ

**🧮 Example: Predicting House Prices**

Suppose we have data:

| **House Size (x)** | **Price (y)** |
| --- | --- |
| 1000 | 200,000 |
| 1500 | 250,000 |
| 2000 | 300,000 |
| 2500 | 350,000 |

We find β1=100β\_1 = 100β1​=100 and β0=100,000β\_0 = 100,000β0​=100,000

Then the equation becomes:

Price=100,000+100×(House Size)\text{Price} = 100,000 + 100 × (\text{House Size})Price=100,000+100×(House Size)

✅ **If house size = 2200 sq.ft**,

Predicted Price=100,000+100×2200=320,000\text{Predicted Price} = 100,000 + 100 × 2200 = 320,000Predicted Price=100,000+100×2200=320,000

**📏 Model Evaluation Metrics**

After building the model, we evaluate how well it predicts using metrics like:

1. **Mean Absolute Error (MAE)**

MAE=1n∑∣yi−y^i∣MAE = \frac{1}{n}\sum |y\_i - ŷ\_i|MAE=n1​∑∣yi​−y^​i​∣

1. **Mean Squared Error (MSE)**

MSE=1n∑(yi−y^i)2MSE = \frac{1}{n}\sum (y\_i - ŷ\_i)^2MSE=n1​∑(yi​−y^​i​)2

1. **Root Mean Squared Error (RMSE)**

RMSE=MSERMSE = \sqrt{MSE}RMSE=MSE​

1. **R² Score (Coefficient of Determination)**

R2=1−SSresSStotR^2 = 1 - \frac{SS\_{res}}{SS\_{tot}}R2=1−SStot​SSres​​

* + SSresSS\_{res}SSres​: Residual sum of squares
  + SStotSS\_{tot}SStot​: Total sum of squares
  + R2R^2R2 close to 1 → better model fit

**✅ Applications of Linear Regression**

| **Area** | **Example** |
| --- | --- |
| **Real Estate** | Predicting house prices |
| **Finance** | Forecasting stock prices, revenue |
| **Marketing** | Predicting sales from ad spend |
| **Health** | Predicting disease progression |
| **Economics** | Predicting GDP, inflation |

**💡 Advantages**

* Easy to understand and implement
* Computationally efficient
* Works well when relationship is linear
* Good baseline model for regression tasks

**⚠️ Limitations**

* Assumes linear relationship between variables
* Sensitive to outliers
* Doesn’t perform well with non-linear data
* Multicollinearity (when predictors are correlated) reduces accuracy

**🔍 Visualization**

A straight line passes through the data points showing the trend between input and output.

|

y ↑ | \*

| \* \*

| \* \*

|\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ x

↑

Regression Line

**🧩 Summary**

| **Concept** | **Description** |
| --- | --- |
| **Type** | Supervised Learning |
| **Goal** | Predict continuous value |
| **Equation** | y=β0+β1x+εy = β\_0 + β\_1x + εy=β0​+β1​x+ε |
| **Error Metric** | MSE, RMSE, R² |
| **Use Case** | Sales forecasting, price prediction |
| **Key Assumption** | Linear relationship between variables |