Linear Regression

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

In this, we assume that there is a linear relationship between the input and output, which means that the output will change constantly as the input changes. Therefore, it can be represented by the equation of a straight line.

$Y = mx+b$

where:

y = the predicted value (dependent variable)

x = the input (independent variable)

m = the slope of the line (how much y changes when x changes)

b = the intercept (the value of y when x = 0)

For example, we want to predict an employee’s salary based on their years of experience. We observe that as employee gain more experience their salary goes up.

- Independent variable (input): Years of experience because it's the factor we observe.
- Dependent variable (output): Salary because it depends on years of experience.
- The model predicts Salary as a linear function of Years of experience.

Therefore, we can simply say that:

Linear regression is a statistical method used to model the relationship between a dependent variable (target, (y)) and one or more independent variables (features, (x)) by fitting a linear equation to the observed data. The goal is to predict the dependent variable as a linear function of the independent variables.

Mathematical Equations

Linear regression models the relationship between the independent variable (YearsExperience, denoted (x)) and the dependent variable (Salary, denoted (y)) using a linear equation:

$y = \beta_0 + \beta_1 x + \varepsilon$

- (**β_0**): Intercept (predicted salary when years of experience is 0).
- (**β_1**): Slope (change in salary per additional year of experience).
- (**ε**): Error term (random variation, assumed normally distributed with mean 0).

The goal is to find (β_0) and (β_1) that minimize the **Sum of Squared Errors (SSE)**:

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

Where:

- (y_i): Actual salary for the (i)-th observation.
- ($\hat{y}_i = \beta_0 + \beta_1 x_i$): Predicted salary.
- (n): Number of observations (30).

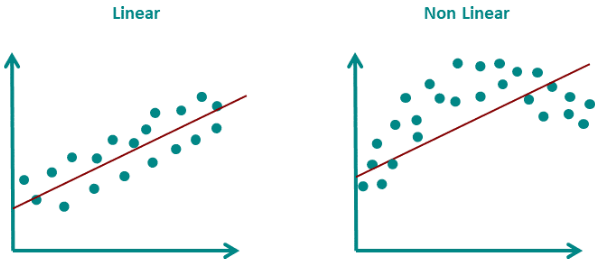
scikit-learn’s LinearRegression internally uses Ordinary Least Squares (OLS) to compute:

$$[\beta_1 = [\sum(x_i - \bar{x})(y_i - \bar{y})] / [\sum(x_i - \bar{x})^2], \beta_0 = \bar{y} - \beta_1\bar{x}]$$

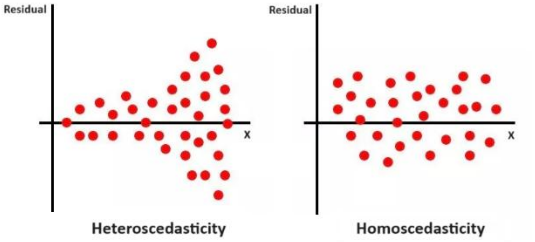
- ($\bar{x} = (1/n) \sum x_i$): Mean of YearsExperience.
- ($\bar{y} = (1/n) \sum y_i$): Mean of Salary.

Assumptions

1. **Linearity:** The relationship between inputs (X) and the output (Y) is a straight line.



2. **Independence of Errors:** The errors in predictions should not affect each other.
3. **Homoscedasticity:** Constant variance of errors. It means that the errors should have equal spread across all values of the input. If the spread changes (like fans out or shrinks), it's called heteroscedasticity and it's a problem for the model.



4. **Normality of Errors:** The errors should follow a normal (bell-shaped) distribution.

Evaluation Metrics

1. **Mean Square Error (MSE):** Average of squared differences between actual and predicted values.

   Mathematically MSE is expressed as:

   $\mathbf{MSE = (1/n) \sum(y_i - \hat{y}_i)^2}$

   Where:

   n = number of observations

   $\mathbf{y_i}$ = actual values

   $\mathbf{\hat{y}_i}$ = predicted values

   It is a way to quantify the accuracy of a model's predictions. It is **sensitive to outliers** as large errors contribute significantly to the overall score.

2. **Mean Absolute Error (MAE):** Average of absolute differences between actual and predicted values. It calculates the accuracy of a regression model.

   Mathematically MAE is expressed as:

   $\mathbf{MAE = (1/n) \sum|y_i - \hat{y}_i|}$

   Where:

   n = number of observations

   $\mathbf{y_i}$ = actual values

   $\mathbf{\hat{y}_i}$ = predicted values

   Lower MAE value indicates better model performance. It is **not sensitive** to the outliers as we consider absolute differences.

3. **Root Mean Squared Error (RMSE):** Square root of MSE, gives error in same unit as dependent variable.

   Mathematically RMSE is expressed as:

   $\mathbf{RMSE = \sqrt{[\ (1/n) \sum(y_i - \hat{y}_i)^2\ ]}}$

   Where:

n = number of observations

$\mathbf{y_i}$ = actual values

$\mathbf{\hat{y}_i}$ = predicted values

It describes how well the observed data points match the expected values or the model's absolute fit to the data.

4. **Coefficient of Determination (R-squared):** Proportion of variance in the dependent variable explained by the model. It is always in the range of 0 to 1. In general, the better the model matches the data, the greater the R-squared number.

   Mathematically R_Squared is expressed as:

   $$\mathbf{R^2 = 1 - [\ \textstyle\sum(y_i - \hat{y}_i)^2 / \sum(y_i - \bar{y})^2\ ]}$$

   Where:

   $\mathbf{y_i}$ = actual values

   $\mathbf{\hat{y}_i}$ = predicted values

Dataset Description

Source: I sourced a dataset from Kaggle. Its link is:

https://www.kaggle.com/datasets/abhishek14398/salary-dataset-simple-linear-regression/data

Columns: (30 rows, 3 columns)

- YearsExperience: Years of work experience (numeric, float, e.g., 1.1 to 10.5).
- Salary: Annual salary in dollars (numeric, float, e.g., 39343.0 to 122391.0).

Results

Linear Regression Results (sklearn):

Intercept (beta_0): 24848.20

Slope (beta_1): 9449.96

R-squared: 0.9570

Mean Squared Error (MSE): 31270951.72

Root Mean Squared Error (RMSE): 5592.04

Mean Absolute Error (MAE): 4644.20