# Experiment/Practical 1 ****Simple Linear Regression****

**Title:** Implementation of Simple Linear Regression

**Aim**: To apply regression algorithm for prediction

**Objective:** Students will learn

* Implementation of a simple linear regression algorithm on the given dataset(s)
* To visualize and interpret the result

# Problem statement

Use given datasets to demonstrate simple linear regression.

# Explanation/Stepwise Procedure/ Algorithm:

* Give a brief description of simple linear regression.

Simple Linear Regression is a supervised learning algorithm used to predict a quantitative outcome based on a single independent variable. The goal is to find the best-fitting straight line (called the regression line) that minimizes the differences (errors) between actual data points and the predicted values.

* Give Mathematical formulation of simple linear regression:

The equation of the regression line is:

y=a+bX

Where:

y: Dependent variable (the value to predict).

X: Independent variable (the input feature).

a: Intercept (the value of yyy when X=0X = 0X=0).

b: Slope of the line, indicating the change in yyy for a one-unit change in XXX.

* Write the importance of simple linear regression in data analysis.

**Simplicity and Interpretability**: Easy to understand and visualize relationships using a straight line equation (y=a+bX).

**Identifies Relationships**: Quantifies the strength and direction of the relationship between variables.

**Predictive Capability**: Provides reliable predictions for outcomes based on input features.

**Trend Analysis**: Helps identify patterns and trends in historical data.

* Mention applications of simple linear regression in real-world scenarios.

**Marketing Analytics:** Predicting sales based on advertising budgets (as in the Sales dataset in your notebook).

**Vehicle Safety:** Predicting stopping distance based on speed (MyCar dataset).

**Human Resources:** Estimating salaries based on years of experience (Salary dataset).

**Weather prediction :** Using regression in weather forecasting, like predicting temperature changes based on humidity.

* Brief explanation of performance metrics (e.g., R², Mean Squared Error).

Performance metrics are essential for evaluating the accuracy and reliability of a regression model. They measure how well the predicted values match the actual values in the dataset. Below are the key metrics commonly used in Simple Linear Regression:

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

**Definition**: R² indicates the proportion of variance in the dependent variable (y) that is explained by the independent variable (X) in the model.

**Formula**: R² =1−Sum of Squared Residuals (SSR)/Total Sum of Squares (TSS)

**Range**: 0 ≤ R² ≤1

* + R2=0: The model explains none of the variance.
  + R2=1: The model perfectly explains the variance.

**Interpretation**: A higher R² score indicates a better fit. For example, R2=0.85 means 85% of the variation in y is explained by X.

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

**Definition**: MSE measures the average squared difference between predicted values (y) and actual values (y).

**Formula**: MSE=∑(yi−ŷi)2 / n Where n is the number of data points.

**Purpose**: Penalizes large errors more heavily than small ones, as the differences are squared.

**Interpretation**: A lower MSE indicates better model performance.

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

**Definition**: RMSE is the square root of MSE, providing error in the same unit as the dependent variable.

**Formula**: RMSE=√MSE

**Purpose**: Easier to interpret since it’s in the same scale as the output variable.

**Interpretation**: Like MSE, a lower RMSE value indicates a better model fit.

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

**Definition**: MAE calculates the average absolute difference between predicted and actual values.

**Formula**: MAE=∑ |yi− ŷi |/n

**Purpose**: Does not penalize large errors as heavily as MSE.

**Interpretation**: A lower MAE suggests better predictive accuracy.

* ***Add necessary figure(s)/Diagram(s)***

A graph with blue dots

Description automatically generated A graph with red dots

Description automatically generated

A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated

A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated

# Input & Output:

# 1. Mycar.csv

# Contains 50 entries with 3 columns:

# Unnamed: 0: An index-like column, likely redundant.

# speed: Speed of a car (integer values).

# dist: Distance traveled (integer values).

# The dataset seems to represent some relationship between speed and distance.

**2. Assignment 1 SALES.csv**

Contains **35 entries** with 1 column:

* + The column appears to have both numerical and textual data in an unclear format (e.g., "12.0 15.0").

The dataset needs cleanup for further analysis, as the structure is not well-defined.

3. Salary\_dataset

This dataset contains Years of experience and the salary of employees this giving a base to train an ml algorithm

The model fits the data well, as evidenced by high R2R^2R2 scores, low error metrics, and strong alignment between the regression line and scatter plots. Minor residuals suggest the data has limited noise, making the model effective for prediction. If the goal is improved performance, further tuning (e.g., addressing outliers or testing polynomial regression) could be explored.

Implementing a linear regression algorithm presents several challenges, including data preprocessing issues like handling missing values, outliers, and feature scaling. Multicollinearity between independent variables can distort model interpretation, and violating key assumptions like linearity, homoscedasticity, and independence of errors can lead to inaccurate predictions. Managing overfitting and underfitting, choosing the right features, and ensuring proper model evaluation are also crucial. Additionally, dealing with large datasets can introduce computational challenges, requiring techniques like regularization or optimization methods for better efficiency and generalization. Addressing these challenges ensures a more robust and accurate model..

**Conclusion:**

Insights derived from the regression line and residuals with good scores can be summarized as follows:

1. **Linear Relationship**: A well-fitting regression line indicates a strong linear relationship between the independent and dependent variables, meaning changes in the predictors are directly associated with changes in the target variable. A high R-squared value suggests that the model explains a significant portion of the variance in the data.
2. **Good Model Fit**: When the regression line aligns closely with the data points, it indicates that the model is making accurate predictions and has a low error rate. This suggests the model is effective in capturing the underlying trends in the data.
3. **Residual Distribution**: If the residuals (errors) are randomly scattered around zero with no obvious patterns, it confirms that the model assumptions are met, and the model is well-calibrated. This typically indicates that the variance of the errors is constant (homoscedasticity), and the residuals are independent.

.