Experiment No. 2
Implementation of Linear Regression Algorithm
Date of Performance:
Date of Submission:
Marks:
Sign:

Aim: Implementation of Linear Regression Algorithm.

Objective: To implement Linear Regression in order to build a model that studies the relationship between an independent and dependent variable. The model will be evaluated by using least square regression method where RMSE and R-squared will be the model evaluation parameters..

Theory:

The least-squares method is a crucial statistical method that is practiced to find a regression line or a best-fit line for the given pattern. This method is described by an equation with specific parameters. The method of least squares is generously used in evaluation and regression. In regression analysis, this method is said to be a standard approach for the approximation of sets of equations having more equations than the number of unknowns. The method of least squares actually defines the solution for the minimization of the sum of squares of deviations or the errors in the result of each equation. Find the formula for sum of squares of errors, which help to find the variation in observed data. The least-squares method is often applied in data fitting.

Least Squares Regression Example
Tom who is the owner of a retail shop, found the price of different T-shirts vs the number of T-shirts sold at his shop over a period of one week.

Price of T-shirts in dollars (x)	# of T-shirts sold (y)
2	4
3	5
5	7
7	10
9	15

Let us use the concept of least squares regression to find the line of best fit for the above data.

Step 1: Calculate the slope 'm' by using the following formula:

$$m = \frac{n\sum xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

After you substitute the respective values, m = 1.518 approximately.

Step 2: Compute the y-intercept value

$$c = y - mx$$

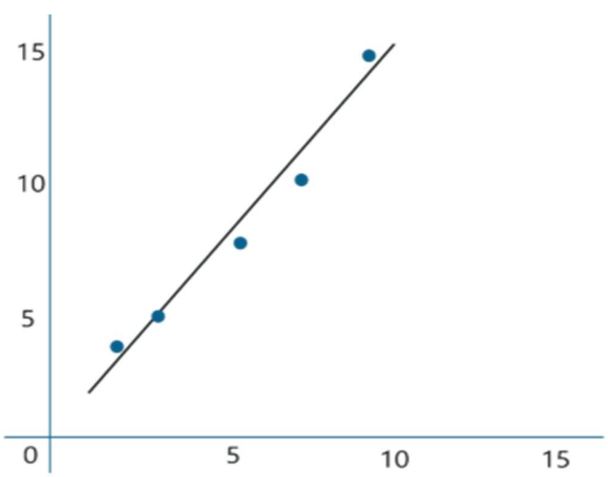
After you substitute the respective values, c = 0.305 approximately.

Step 3: Substitute the values in the final equation

$$y = mx + c$$

Price of T-shirts in dollars (x)	# of T-shirts sold (y)	Y=mx+c	error
2	4	3.3	-0.67
3	5	4.9	-0.14
5	7	7.9	0.89
7	10	10.9	0.93
9	15	13.9	-1.03

Let's construct a graph that represents the **y=mx + c** line of best fit:



Now Tom can use the above equation to estimate how many T-shirts of price $8 can he sell at the retail shop.

y = 1.518 x 8 + 0.305 = 12.45 T-shirts

This comes down to 13 T-shirts!

Dataset:

The data set contains the following variables:

- **Gender:** Male or female represented as binary variables
- **Age:** Age of an individual
- **Head size in cm^3:** An individuals head size in cm^3
- **Brain weight in grams:** The weight of an individual's brain measured in grams

These variables need to be analyzed in order to build a model that studies the relationship between the head size and brain weight of an individual.

Step 1: Import the required libraries

Step 2: Import the data set

Step 3: Assigning 'X' as independent variable and 'Y' as dependent variable

Step 4: Calculate the values of the slope and y-intercept

Step 5: Plotting the line of best fit

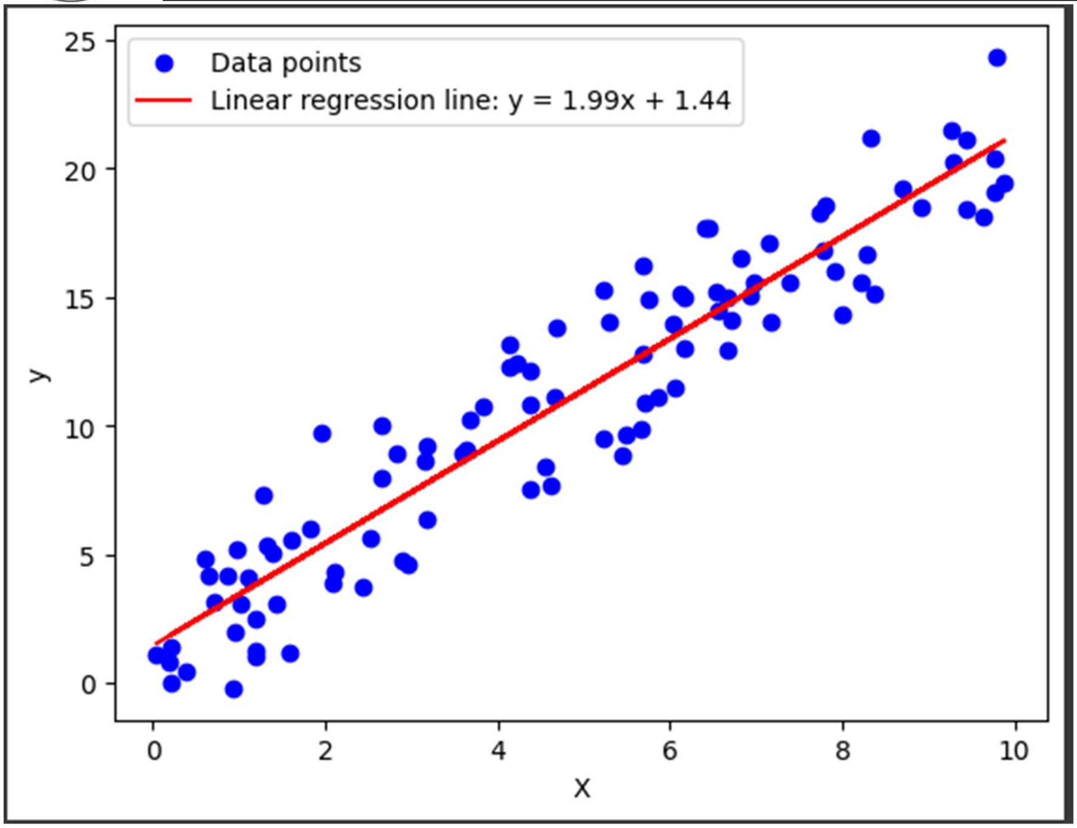
Step 6: Model Evaluation

Implementation:

Code:

```
import numpy as np
import matplotlib.pyplot as plt

# Generate sample data (for example, let's say we're predicting y from x)
np.random.seed(0)  # For reproducibility
X = np.random.rand(100, 1) * 10  # 100 random data points for X
y = 2 * X + 1 + np.random.randn(100, 1) * 2  # Linear relation with noise

# Add a column of ones to X to account for the intercept term
X_b = np.c_[np.ones((X.shape[0], 1)), X]  # Adding a bias column (X0 = 1)

# Calculate the optimal parameters (theta) using the Normal Equation
```

```
theta = np.linalg.inv(X_b.T.dot(X_b)).dot(X_b.T).dot(y)

# Extract the intercept and slope from theta
intercept, slope = theta[0], theta[1]

# Print out the coefficients
print(f"Intercept: {intercept[0]}")
print(f"Slope: {slope[0]}")

# Plotting the data points
plt.scatter(X, y, color='blue', label='Data points')

# Plotting the regression line
plt.plot(X, X_b.dot(theta), color='red', label=f'Linear regression line: y = {slope[0]:.2f}x + {intercept[0]:.2f}')

plt.xlabel("X")
plt.ylabel("y")
plt.legend()

# Save the plot as an image
plt.savefig("linear_regression_plot.png")

# Show the plot
plt.show()
```



Conclusion:

Comment on the Least Square Method used for regression.

The Least Squares Method is a foundational and widely used technique for regression analysis, offering simplicity, interpretability, and optimal performance under linearity and normality assumptions. However, it is sensitive to outliers, struggles with multicollinearity, and relies on strict assumptions like homoscedasticity. Despite its limitations, it remains a powerful tool for linear modeling, with applications in diverse fields. Extensions like Ridge Regression, LASSO, and robust regression methods address some of its weaknesses, making it adaptable to more complex scenarios. Overall, it is a cornerstone of statistical modeling but should be applied with caution and awareness of its assumptions.