Problem Statement

Title: *Predicting Student Exam Scores Based on Study Hours Using Simple Linear Regression*

Objective

To build a simple linear regression model to predict a student's exam score based on the number of hours they study.

Background

There is a widely accepted belief that the more a student studies, the better they perform in exams. This project aims to quantify that relationship using a simple linear regression model, where we predict exam scores using the number of study hours.

Dataset

Hours Studied (X)	Exam Score (%) (Y)
1.0	52
2.0	55
3.0	60
4.0	63
5.0	66
6.0	70
7.0	73
8.0	75
9.0	78
10.0	85

Goals

- Visualize the linear relationship between study hours and exam scores
- Fit a linear regression line to the data
- Estimate model parameters (slope and intercept)
- Predict exam scores for new values of study hours
- Evaluate performance using:
  - Mean Squared Error (MSE)
  - R^2 Score

Assumptions

- There is a linear relationship between study hours and exam scores.
- The residuals are normally distributed and homoscedastic.
- No significant outliers are present.

Solution

Problem: Predict Exam Scores Based on Study Hours

We are given a dataset of 10 students with two variables:

- Independent variable (X): Hours Studied
- Dependent variable (Y): Exam Score (%)

Step 1: Dataset

Hours Studied (X)	Exam Score (Y)
1.0	52
2.0	55
3.0	60
4.0	63
5.0	66
6.0	70
7.0	73
8.0	75
9.0	78
10.0	85

Step 2: Calculate the Means

$\bar{X} = \frac{1 + 2 + \dots + 10}{10} = \frac{55}{10} = 5.5$

$\bar{Y} = \frac{52 + 55 + \dots + 85}{10} = \frac{677}{10} = 67.7$

Step 3: Compute the Slope (b_1)

$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$

(See detailed table from previous step)

$\sum (X - \bar{X})(Y - \bar{Y}) = 283.5,\quad \sum (X - \bar{X})^2 = 82.5$ $b_1 = \frac{283.5}{82.5} \approx 3.436$

Step 4: Compute the Intercept (b_0)

$b_0 = \bar{Y} - b_1 \cdot \bar{X} = 67.7 - (3.436 \cdot 5.5) = 67.7 - 18.898 = 48.802$

Step 5: Regression Equation

$\hat{Y} = b_0 + b_1 X = 48.802 + 3.436 X$

Step 6: Example Prediction

Predict the exam score for 7.5 hours of study:

$\hat{Y} = 48.802 + 3.436 \cdot 7.5 = 74.57$

Step 7: Model Evaluation

We evaluate the model using:

1. Mean Squared Error (MSE)

$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

Using predicted (\hat{Y} = 48.802 + 3.436X), compute errors for each row:

X	Y (actual)	$(\hat{Y})$	Error $(Y - \hat{Y})$	Squared Error
1	52	52.238	-0.238	0.057
2	55	55.674	-0.674	0.454
3	60	59.110	0.890	0.792
4	63	62.546	0.454	0.206
5	66	65.982	0.018	0.0003
6	70	69.418	0.582	0.339
7	73	72.854	0.146	0.021
8	75	76.290	-1.290	1.664
9	78	79.726	-1.726	2.979
10	85	83.162	1.838	3.378

$\text{MSE} = \frac{1}{10}(0.057 + 0.454 + 0.792 + 0.206 + 0.0003 + 0.339 + 0.021 + 1.664 + 2.979 + 3.378) = \frac{9.89}{10} = 0.989$

---

2. Root Mean Squared Error (RMSE)

$\text{RMSE} = \sqrt{MSE} = \sqrt{0.989} \approx 0.994$

---

3. R^2 Score

$R^2 = 1 - \frac{\sum (Y - \hat{Y})^2}{\sum (Y - \bar{Y})^2}$

- Numerator: $(\sum (Y - \hat{Y})^2 = 9.89)$
- Denominator: $(\sum (Y - \bar{Y})^2)$

From earlier step:

$\sum (Y - \bar{Y})^2 = \sum (X - \bar{X})(Y - \bar{Y}) \cdot \frac{1}{b_1} = \frac{283.5}{3.436} \approx 82.5$

$R^2 = 1 - \frac{9.89}{82.5} = 1 - 0.12 = 0.88$

Final Answers

- **Regression Equation:**
  $\hat{Y} = \boxed{48.802 + 3.436X}$

- **Prediction for 7.5 hours:**
  $\boxed{74.57%}$

- **Mean Squared Error (MSE):**
  $\boxed{0.989}$

- **Root MSE (RMSE):**
  $\boxed{0.994}$

- **R^2 Score:**
  $\boxed{0.88}$