Self Assignment: Linear Regression and Correlation

2024

1 Introduction

Linear Regression and **Correlation**. Linear Regression is used for predicting a dependent variable based on one or more independent variables, while Correlation measures the strength and direction of a linear relationship between two variables.

2 Linear Regression

2.1 What is Linear Regression?

Linear regression is a method to model the relationship between two variables by fitting a linear equation to the observed data. The simplest form is called **Simple Linear Regression**, which involves two variables: one independent variable (x) and one dependent variable (y).

The general form of the linear regression equation is:

$$y = mx + b \tag{1}$$

Where:

- y is the dependent variable (what you're trying to predict).
- x is the independent variable (what you're using to make predictions).
- m is the slope of the line, representing how much y changes for a unit change in x.
- b is the y-intercept, representing the value of y when $x = 0$.

2.2 Fitting the Linear Model

The goal of linear regression is to find the best-fitting line through the data points, which minimizes the error between the predicted values and the actual values. This is achieved using the **Least Squares Method**.

The sum of squared errors (SSE) is calculated as:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

Where $\hat{y}_i$ represents the predicted values from the regression line.

The coefficients (m and b) are chosen such that this error is minimized.

2.3 Multiple Linear Regression

In cases where there are multiple independent variables, the model extends to:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \tag{3}$$

Where:

- k is the number of independent variables.
- $b_0, b_1, \ldots, b_k$ are the coefficients of the regression equation.

Multiple Linear Regression is widely used in machine learning models and predictive analytics, particularly in fields like finance, healthcare, and artificial intelligence.

2.4 Example: Simple Linear Regression

Consider a dataset where you have the number of hours studied (x) and the corresponding exam score (y). The goal is to predict exam scores based on study hours.

$$y = 5x + 50 \tag{4}$$

This equation implies that for every extra hour studied, the score increases by 5 points, with a base score of 50 when no study hours are logged.

3 Correlation

3.1 What is Correlation?

Correlation is a statistical measure that expresses the extent to which two variables are linearly related. It can take a value between -1 and 1:

- A correlation of +1 indicates a perfect positive linear relationship.
- A correlation of -1 indicates a perfect negative linear relationship.
- A correlation of 0 indicates no linear relationship.

3.2 Correlation Coefficient (r)

The correlation coefficient, denoted as **r**, quantifies the strength and direction of a linear relationship between two variables. It is calculated as:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \tag{5}$$

Where:

- n is the number of data points.
- $\sum xy$ is the sum of the product of the paired data.
- $\sum x$ and $\sum y$ are the sums of the individual variables.

3.3 Types of Correlation

- **Positive Correlation**: As one variable increases, the other also increases.
- **Negative Correlation**: As one variable increases, the other decreases.
- **No Correlation**: There is no apparent relationship between the two variables.

3.4 Correlation vs Causation

It's crucial to understand that **correlation does not imply causation**. Two variables may be correlated, but it does not mean that one causes the other to happen. This distinction is especially important in fields like economics, medicine, and artificial intelligence, where confounding variables might exist.

4 Scatter Plot with Linear Relationship

Consider a dataset showing the number of hours a student studies and the corresponding exam scores they receive. Below is a scatter plot that represents this relationship.

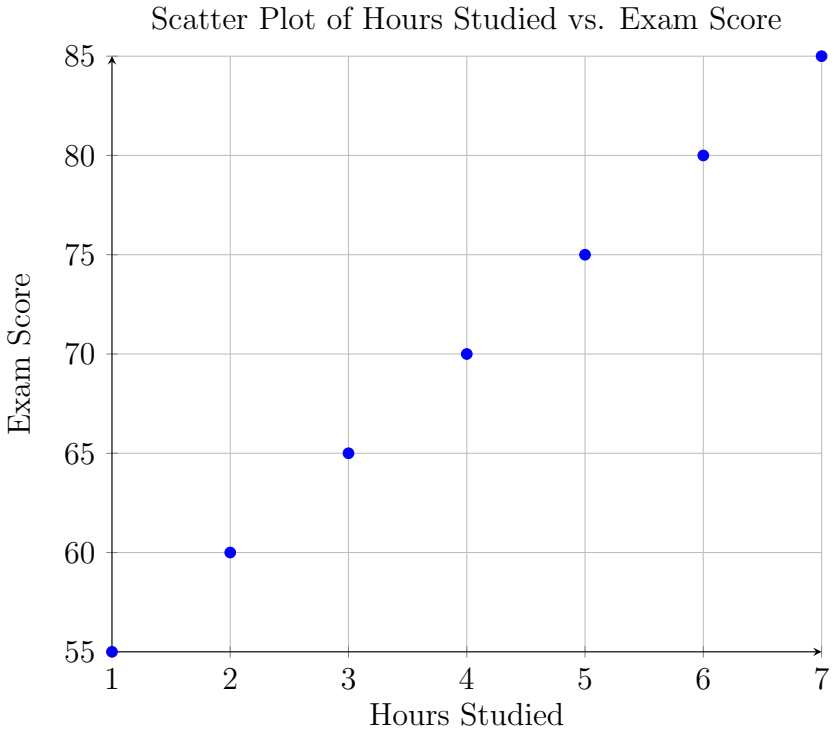

Figure 1: Scatter Plot showing the relationship between Hours Studied and Exam Score.

This scatter plot shows a positive relationship: as the number of hours studied increases, the exam score increases. The next step would be to fit a linear regression line through these points to model this relationship.

5 Fitting a Linear Regression Line

Now that we've observed a positive relationship between study hours and exam scores in the scatter plot, we can fit a linear regression line to the data. The regression line equation is given as:

$$y = 5x + 50 \tag{6}$$

Where:

- x represents the number of hours studied.
- y represents the predicted exam score.

Below is the plot showing both the scatter points and the linear regression line.

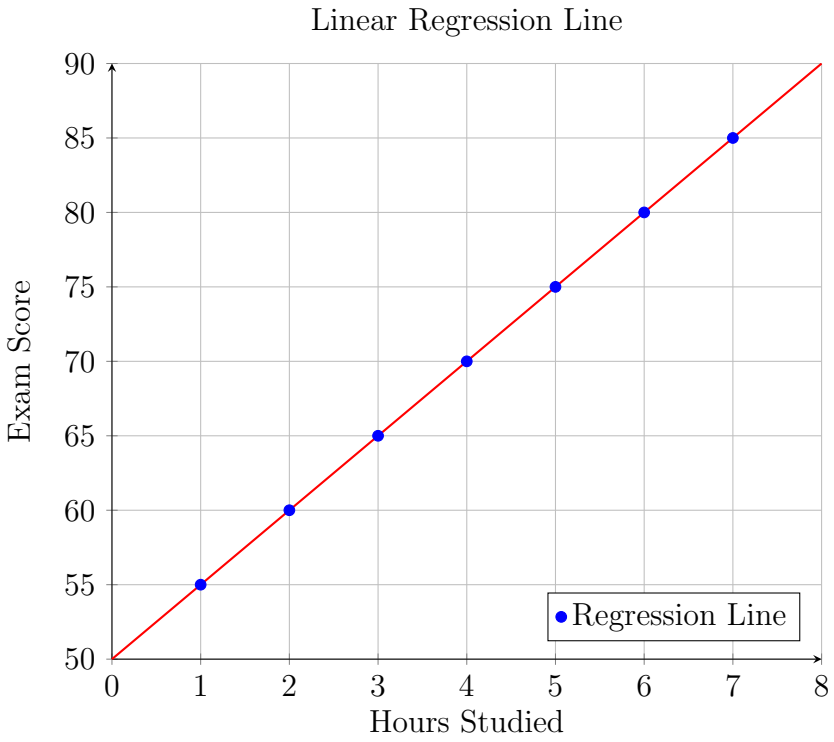

Figure 2: Linear Regression Line fitted to the data.

The red line in the graph represents the best-fitting regression line, which predicts the exam score based on the number of hours studied. For example, if a student studies for 4 hours, the model predicts an exam score of 70.