Linear Regression and Line of Best Fit

Badger Code

July 22, 2023

1 Introduction

Linear regression is a statistical method used to model the relationship between a dependent variable y and one or more independent variables x. The goal is to find the "line of best fit" that represents the relationship between the variables. The line of best fit is a straight line of the form $y = mx + b$, where m is the slope and b is the y-intercept.

2 Calculating Linear Regression

To calculate linear regression, we follow these steps:

2.1 Step 1: Data Collection

Collect the data for the dependent variable y and the independent variable x.

2.2 Step 2: Data Visualization

Plot the data points on a scatter plot to visualize the relationship between x and y.

2.3 Step 3: Calculate the Mean

Find the mean (average) of x and y from the dataset. Denote the means as $\bar{x}$ and $\bar{y}$.

2.4 Step 4: Calculate the Slope (m)

The slope m can be calculated using the formula:

$$m = \frac{\sum((x_i - \bar{x})(y_i - \bar{y}))}{\sum(x_i - \bar{x})^2}$$

where $\sum$ represents the sum, x_i and y_i are individual data points, $\bar{x}$ is the mean of x, and $\bar{y}$ is the mean of y.

2.5   Step 5: Calculate the Y-Intercept (b)

The y-intercept b can be calculated using the formula:

$$b = \bar{y} - m \cdot \bar{x}$$

3   Example

Let's calculate the linear regression and the line of best fit for the following dataset:

x	y
1	2
2	4
3	6
4	8

3.1   Step 2: Data Visualization

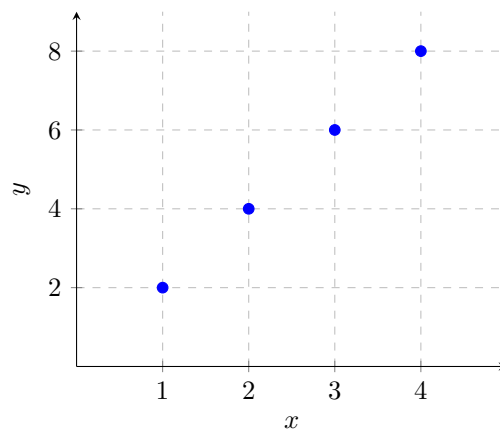| x	y
1	2
2	4
3	6
4	8

Figure 1: Scatter plot of the data points.

3.2   Step 3: Calculate the Mean

$\bar{x} = \frac{1+2+3+4}{4} = 2.5$

$\bar{y} = \frac{2+4+6+8}{4} = 5$

3.3   Step 4: Calculate the Slope (m)

$$m = \frac{((-1.5) \cdot (-3)) + ((-0.5) \cdot (-1)) + (0.5 \cdot 1) + (1.5 \cdot 3)}{(1.25)^2 + (0.75)^2 + (0.25)^2 + (1.75)^2}$$

$$m = \frac{7.5 + 0.5 + 0.5 + 4.5}{1.5625 + 0.5625 + 0.0625 + 3.0625}$$

$$m = \frac{12}{5.25} \approx 2.2857$$

3.4   Step 5: Calculate the Y-Intercept (b)

$$b = 5 - 2.2857 \cdot 2.5 \approx -0.7143$$

4   Line of Best Fit

The line of best fit for the given dataset is approximately:

$$y = 2.2857x - 0.7143$$