

Example Scenario

Imagine you are a data analyst and you want to predict the score a student will get on an exam based on the number of hours they studied.

Step 1: Collect Data

Suppose you have collected the following data:

Hours Studied (x)	Exam Score (y)
1	50
2	55
3	65
4	70
5	75

Step 2: Plot the Data

First, plot the data points on a scatter plot to visualize the relationship between hours studied and exam scores.

Step 3: Calculate the Means

Calculate the mean (average) of the independent variable (hours studied) and the dependent variable (exam score).

$$\bar{x} = \frac{1+2+3+4+5}{5} = 3$$

$$\bar{y} = \frac{50+55+65+70+75}{5} = 63$$

Step 4: Calculate the Slope (β_1)

The slope (β_1) of the regression line can be calculated using the formula:

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Calculating the numerator:

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= (1 - 3)(50 - 63) + (2 - 3)(55 - 63) + (3 - 3)(65 - 63) \\ &+ (4 - 3)(70 - 63) + (5 - 3)(75 - 63) \\ &= (-2)(-13) + (-1)(-8) + (0)(2) + (1)(7) + (2)(12) \\ &= 26 + 8 + 0 + 7 + 24 \\ &= 65\end{aligned}$$

Calculating the denominator:

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x})^2 &= (1 - 3)^2 + (2 - 3)^2 + (3 - 3)^2 + (4 - 3)^2 + (5 - 3)^2 \\ &= (-2)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2 \\ &= 4 + 1 + 0 + 1 + 4 \\ &= 10\end{aligned}$$

Thus, the slope (β_1) is:

$$\beta_1 = \frac{65}{10} = 6.5$$

Step 5: Calculate the Intercept (β_0)

The intercept (β_0) can be calculated using the formula:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Substituting the values:

$$\beta_0 = 63 - 6.5 \times 3$$

$$\beta_0 = 63 - 19.5$$

$$\beta_0 = 43.5$$

Step 6: Formulate the Regression Equation

Now that we have the slope and intercept, we can write the equation of the regression line:

$$y = \beta_0 + \beta_1 x$$

$$y = 43.5 + 6.5x$$

Step 7: Make Predictions

Using the regression equation, you can predict the exam score for a given number of hours studied. For example, to predict the score for a student who studies for 4 hours:

$$y = 43.5 + 6.5 \times 4$$

$$y = 43.5 + 26$$

$$y = 69.5$$

So, a student who studies for 4 hours is predicted to score 69.5 on the exam.

Step 8: Evaluate the Model

To evaluate the model, you can calculate metrics such as the coefficient of determination (R-squared), which measures how well the independent variable explains the variability of the dependent variable. However, this typically requires more extensive calculations or tools.