Introduction to Linear Regression

Mustafa AbdulRazek

October 1, 2024

1 What is Linear Regression?

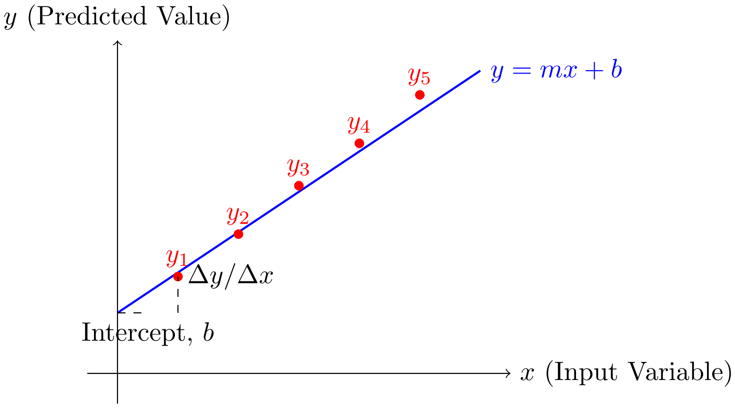
Linear Regression is a method used to predict a continuous outcome (called the dependent variable, y) based on one or more input variables (called independent variables, x).

The relationship between x and y is modeled as a straight line with the following equation:

$$y = mx + b$$

Where:

- y is the predicted value.
- x is the input variable (or feature).
- m is the slope of the line, representing how much y changes for each unit change in x.
- b is the intercept, representing the value of y when $x = 0$.



2 Minimizing the Error

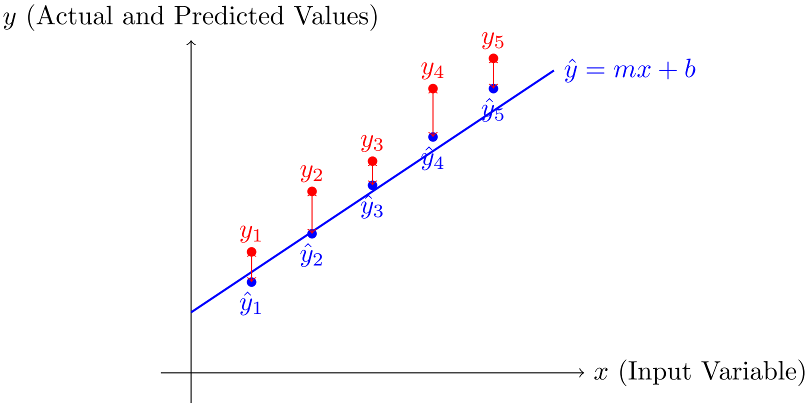
The goal of Linear Regression is to find the best line that minimizes the difference between the actual values of y and the predicted values $\hat{y}$.

The error for each point is given by:

$$\text{Error}_i = y_i - \hat{y}_i$$

We minimize the sum of squared errors (SSE) to avoid positive and negative errors from canceling each other out:

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



3 Finding the Best Line (Calculating m and b)

To find the best-fitting line, we calculate the slope m and intercept b that minimize the SSE. The formulas for m and b are:

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

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

Where:

- n is the number of data points.
- $\sum xy$ is the sum of the product of x and y.
- $\sum x^2$ is the sum of the squares of x.
- $\sum x$ and $\sum y$ are the sums of the x and y values, respectively.

4 Making Predictions

Once we have m and b, we can predict y for any new value of x using the line equation:

$$y = mx + b$$

5 Evaluating the Model

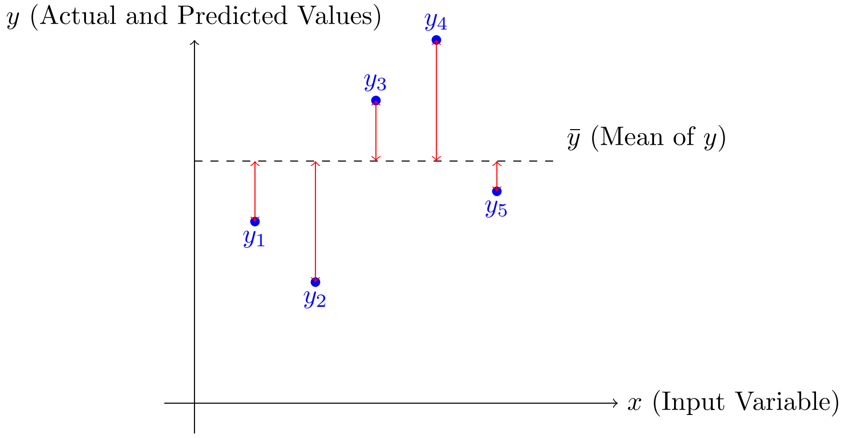
To measure how well our model fits the data, we use the R^2 (R-squared) value, which shows the proportion of the variance in y that is predictable from x:

$$R^2 = 1 - \frac{\text{SSE}}{\text{TSS}}$$

Where:

- SSE is the sum of squared errors from our model.
- TSS is the total variance of the actual data from the mean.

The closer R^2 is to 1, the better the model fits the data.



6 Understanding Total Variance

In the context of linear regression, total variance refers to how much the actual values of the dependent variable y vary from their mean $\bar{y}$. It provides a measure of how spread out the actual data points are.

Mathematically, total variance, also known as **Total Sum of Squares (TSS)**, is calculated as:

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

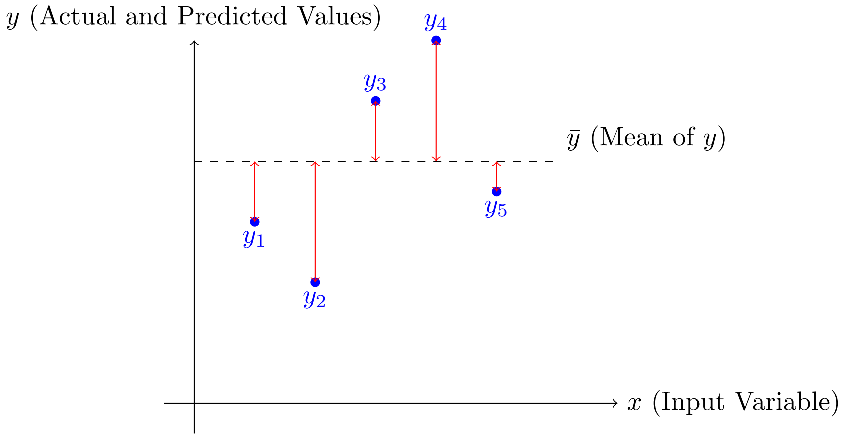
Where:

- y_i is the actual value of y for each data point.
- $\bar{y}$ is the mean of the actual values of y.
- n is the number of data points.

The TSS represents the total amount of variation in the data. Our goal in linear regression is to explain this variation using a model. Some of this variation is explained by the regression model (called the explained variance), while the rest remains unexplained (called the residual variance).

6.1 Visualization of Total Variance

In the plot below, we show how the data points deviate from their mean:



In this chart:

- The blue points represent the actual data points $y_1, y_2, \ldots, y_5$.
- The dashed horizontal line represents the mean of y, denoted by $\bar{y}$.
- The red vertical lines show the deviations of each data point from the mean $\bar{y}$.

The sum of the squared lengths of these red lines gives the Total Sum of Squares (TSS), which represents the total variance in the data. Our objective in regression is to reduce the portion of this variance that is unexplained by the model.

7 Example Problem

Let's say we have data on the number of hours students studied (x) and their test scores (y):

Hours (x)	Score (y)
1	50
2	60
3	70
4	85
5	95

7.1 Step 1: Calculate m and b

We first calculate the necessary sums:

$$\sum x = 1 + 2 + 3 + 4 + 5 = 15$$

$$\sum y = 50 + 60 + 70 + 85 + 95 = 360$$

$$\sum xy = (1 \cdot 50) + (2 \cdot 60) + (3 \cdot 70) + (4 \cdot 85) + (5 \cdot 95) = 1195$$

$$\sum x^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 55$$

Now, we can use the formulas to calculate m and b:

$$m = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} = \frac{5(1195) - (15)(360)}{5(55) - (15)^2} = \frac{5975 - 5400}{275 - 225} = \frac{575}{50} = 11.5$$

$$b = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} = \frac{(360)(55) - (15)(1195)}{5(55) - (15)^2} = \frac{19800 - 17925}{275 - 225} = \frac{1875}{50} = 37.5$$

Thus, the equation of the line is:

$$y = 11.5x + 37.5$$

7.2 Step 2: Make Predictions

Now, we can use this equation to predict the score for a student who studies for 6 hours:

$$y = 11.5(6) + 37.5 = 69_37.5 = 106.5$$

Thus, if a student studies for 6 hours, the predicted test score is 165.