Date
Page

Simple Linear Regression - Math In-Depth

1) Imagine you have some data points:
x = number of hours you study
y = marks you get

The points don't fall on a perfect line (real life is messy), but we want to draw one line that best shows the trend

That line is
$\hat{y} = mx + b$ or $\hat{y} = b_0 + b_1 x$

where:
m : slope (same as b_1)
b = intercept (same as b_0)
X = input (like hours studied)
$\hat{y}$ = predicted output (like marks)

Error = $y - \hat{y}$

- Slope m : tells us how much y changes if X goes up by 1
- Intercept b : tells us the value of y when X = 0

So it's exactly like the line equation : $y = mX + c$

The formula

$$m = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma (X_i - \bar{X})^2}$$ - X and Y move together (correlation)

- is used to find the slope m of the best-fit line in simple linear regression

2) What each symbol means

- x_i = the i-th input value (like one student's study hours)
- y_i = the i-th output value (like that student's marks)

- $\bar{x}$ = the mean of all x values
  $\bar{x}$ = sum of all x's / number of data points

  $\bar{y}$ = sum of all y's / number of data points

- Simple linear regression - have only one input (x) and one output (y).
- In multiple regression we can have many inputs

3) Intuition of the formula

- The top part $\Sigma(x_i - \bar{x})(y_i - \bar{y})$ measures how x and y move together. (This is like correlation)

- The bottom part $\Sigma(x_i - \bar{x})^2$ measures how spread out x is

So slope = (how much x and y move together) / (how much x varies)

That gives us :- If x goes up by 1 unit, how much does y go up on average?

Eg

X (hours)	Y (marks)
1	2
2	4
3	5

X - (Input) Independent Variable

Y - (Output) Dependent Variable

Step 1 Find the means

$$\bar{x} = \frac{1+2+3}{3} = \frac{6}{3} = 2$$

$$\bar{y} = \frac{2+4+5}{3} = \frac{11}{3} = 3.6$$

Step 2 Apply the slope formula

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

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	-1	-1.6	1.6	1
2	4	0	0	0	0
3	5	1	1.4	1.4	1

Step 3 Add them up

$\Sigma (x_i - \bar{x})(y_i - \bar{y}) = 1.6 + 0 + 1.4 = 3.0$

$\Sigma (x_i - \bar{x})^2 = 1 + 0 + 1 = 2$

Step 4 Find slope m

$$m = \frac{3.0}{2} = 1.5$$

So, the slope is $m = 1.5$

Steps Find intercept b

$b = \bar{Y} - m\bar{X}$

$= 3.6 - (1.5 \times 2)$

$= 3.6 - 3$

$= 0.6$

for eg $x = 3$ and $x = 4$

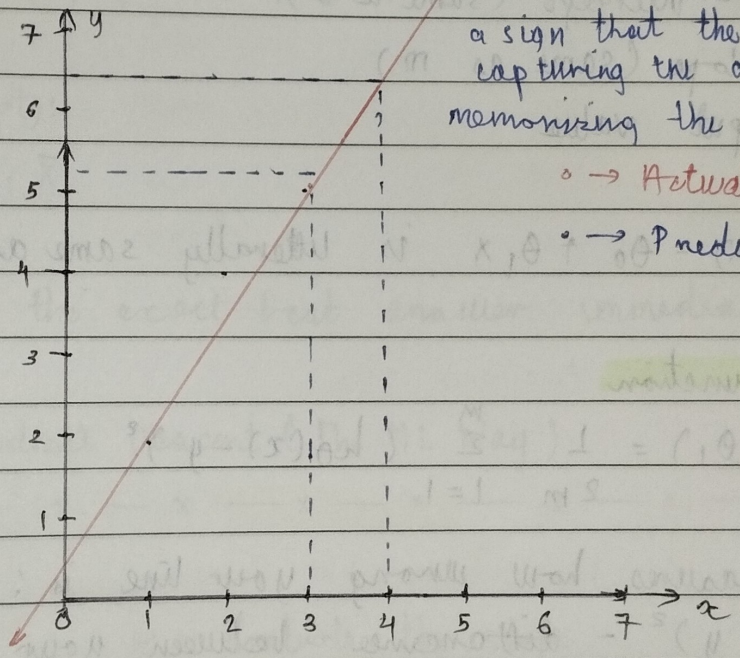
$1.5(3) + 0.6$

$4.5 + 0.6 = 5$

So intercept is $b = 0.6$

Final Line Equation

$\hat{y} = 1.5X + 0.6$



These predicted values are close to the actual values but not always exact. That's a sign that the model is capturing the overall trend and memorizing the data precisely.

∘ → Actual value

• → Predicted value