Simple Linear Regression

⇒ when we have a single input attribute (x) and we want to use linear regression, this is called Simple Linear Regression.

⇒ If we had multiple input attributes (eg. x1, x2, x3 etc.) This would be called multiple linear regression.

⇒ with Simple Linear Regression ~~model~~ we want to model our data as follows:-

$$\boxed{y = B_0 + B_1 * x}$$

y → O/P variable

B_0, B_1 → Coefficient

x → input variable

Technically B_0 is called the intercept and B_1 is called the slope.

⇒ The goal is to find the best estimates for the coefficients to minimize the errors in predicting y from x.

⇒ First we estimate the value for B1 as:

$$B1 = \frac{sum((xi - mean(x)) * (y_i - mean(y)))}{sum((xi - mean(x))^2)}.$$

mean () → average value for the variable in our dataset.

⇒ we can calculate B0 using B1 and some statistics from our dataset as follow:

$$B0 = mean(y) - B1 * mean(x).$$

(8)

Estimating the slope (B_1)

⇒ First we need to calculate the mean value of x and y.

$$\text{mean} = \frac{1}{\cancel{n \times sum(x)}} \quad \boxed{\frac{1}{n} \times sum(x)}$$

n → is the no. of values (5 in this case).
We can use the AVERAGE() function in your spreadsheet.

mean (x) = 3

mean (y) = 2.8

$\frac{1}{5} \times 15 = 3$

$\frac{1}{5} \times 14 = 2.8$

1	X	Y	
2	1	1	
3	2	3	
4	4	3	
5	3	2	
6	5	5	

⇒ Now we calculate the error of each variable from the mean

1	x	mean(x)	x − mean(x)
2	1	3	−2
3	2	3	−1
4	3	3	0
5	4	3	1
6	5	3	2

9

	y	mean(y)	y-mean(y)
1			
2	1	2.8	-1.8
3	3	2.8	0.2
4	3	2.8	0.2
5	2	2.8	-0.8
6	5	2.8	2.2

=> Now we have all parts for calculating the numerator. All we need do do in multiple the error for each x with error for each y. and calculate the sum of multiplication

	x-mean(x)	y-mean(y)	multiplication
1			
2	-2	-1.8	3.6
3	-1	0.2	-0.2
4	0	0.2	0
5	1	-0.8	-0.8
6	2	2.2	4.4

7

3.6
4.4
8.0

8.0

=> Summing the final column we have the numerator as ~~8~~ 0.7

=> now we calculate the bottom part of the eqⁿ for calculating B_1

	x-mean(x)	squared
1		
2	-2	4
3	-1	1
4	1	1
5	0	0
6	2	4

$\frac{7}{10} = 0.7$

$B_1 = 8/10 = 0.8$

now estimating the intercept (B_0)

$B_0 = \text{mean}(y) - B_1 \times \text{mean}(x)$

$B_0 = 2.8 - 0.\cancel{8}7 \times 3$

$B_0 = 0.\cancel{4}7$

⑪

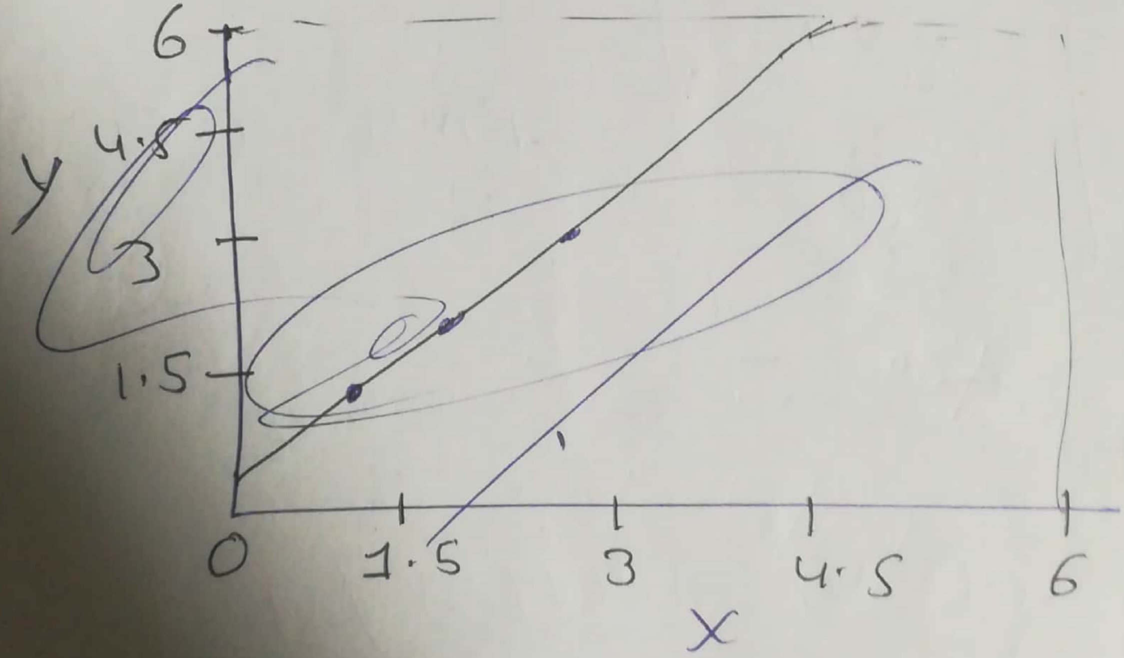
making Predictions.

$$y = B_0 + B_1 \times x$$

$$y = 0.\cancel{7} + 0.\cancel{8} \times x$$

Lets try out the mode by making Predictions for our training data.

1	x	y	Predicted y
2	1	1	1.2
3	2	3	2
4	3	3	3.6
5	4	2	2.8
6	5	5	4.4



12

Estimating the Error.

Root mean Squared Error or RMSE

$$RMSE = Sqrt\left(sum\left(\left(\underbrace{p}_{\text{Predicted value}}\, i - \underbrace{yi}_{\text{actual value}}\right)^2\right) / n\right)$$

i	Pred	y	error
2	1.2	1	0.2
3	2	3	−1
4	3.6	3	0.6
5	2.8	2	0.8
6	4.4	5	0.6

error	Squared error
0.2	0.04
~~0~~ 1	1
0.6	0.36
0.8	0.64
−0.6	0.36

RMSE = 0.692

Shortcut:-

$$B1 = \underbrace{Corr}_{\text{Pearson's correlation coefficient}}(X, Y) * Stdev(Y) / Stdev(X)$$