Regression Math Explained

Assume dataset below:

X	*Y*
0	*4*
1	*12*
2	*28*
3	*52*
4	*80*

To find the line of the best fit, we can choose two points and calculate $\boldsymbol{y = mx + c}$ *easily:*

X1, Y1 = 0, 4

X2, Y2 = 4, 80

Firstly, calculate the slope using two values:

$$m = \frac{y1 - y2}{x1 - x2}$$

$$m = \frac{4 - 80}{0 - 4}$$

$$m = -\frac{76}{-4} = \mathbf{19}$$

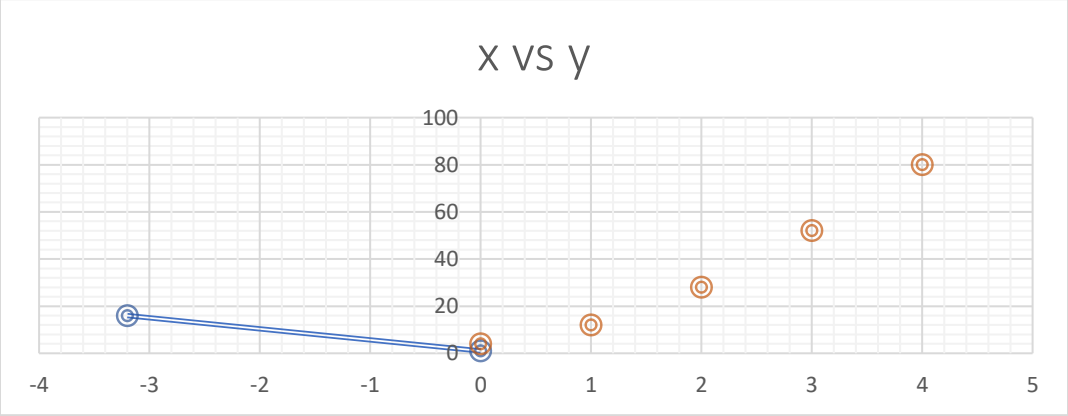
Y-intercept or coefficient bias is what the value of y is when x is 0. Given $m = 19$ *and x, y = 1, 4:*

$$y - y1 = m * (x - x1)$$

$$y - 4 = 19x - 19$$

$$\boldsymbol{y = 19x - 15}$$

This formula above is fairly accurate however, we need to minimize the errors. Notice graph below. It perfectly describes points used, however ones in between are not described as well.



To minimize the error, we use Least Square Regression method which minimizes the error in such a way that all square errors are minimized. First, the formula for calculating slope is:

$$m = \frac{sum\big((x - x_{mean}) * (y - y_{mean})\big)}{sum(x - x_{mean})^2}$$

Firstly, calculate x and y means:

$$x_{mean} = \frac{0 + 1 + 2 + 3 + 4}{5} = \mathbf{2}$$

Additionally, calculate x-x mean for each value of x:

$$x1 - x_{mean} = 0 - 2 = -2$$
$$x2 - x_{mean} = 1 - 2 = -1$$
$$x3 - x_{mean} = 2 - 2 = 0$$
$$x3 - x_{mean} = 3 - 2 = 1$$
$$x4 - x_{mean} = 4 - 2 = 2$$

Do the same for y values:

$$y_{mean} = \frac{4 + 12 + 28 + 52 + 80}{5} = \mathbf{35.2}$$

Additionally, calculate x-x mean for each value of x:

$$y1 - y = 4 - 35.2 = -31.2$$
$$\ldots$$
$$y4 - y_{mean} = 80 - 35.2 = 44.8$$

*Sum of all these multiplied together will give us numerator which equals to **192.** Sum of (x-x mean)^2 values will give us **10**. Therefore, slope results in **19.2.***

To calculate the y-intercept we use:

$$c = y_{mean} - m * x_{mean}$$
$$= 35.1 - 19.2 * 3 = \mathbf{-22.4}$$

Overall formula equals to:

$$y = 19.2x - 22.5$$

Notice the difference in the plot on the right.