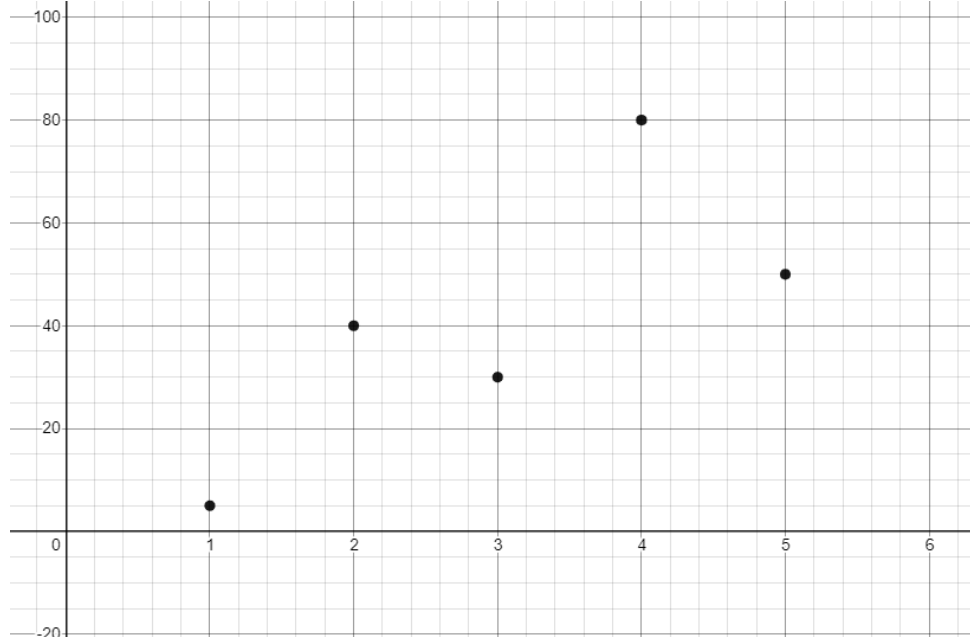
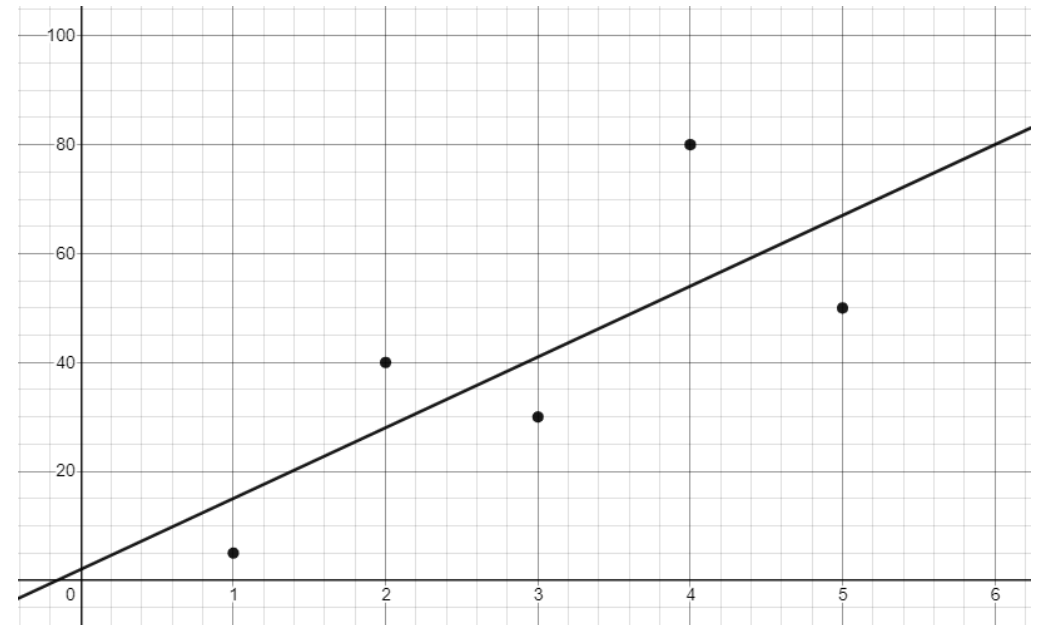


Least Square Regression Lines: Formula Derivation

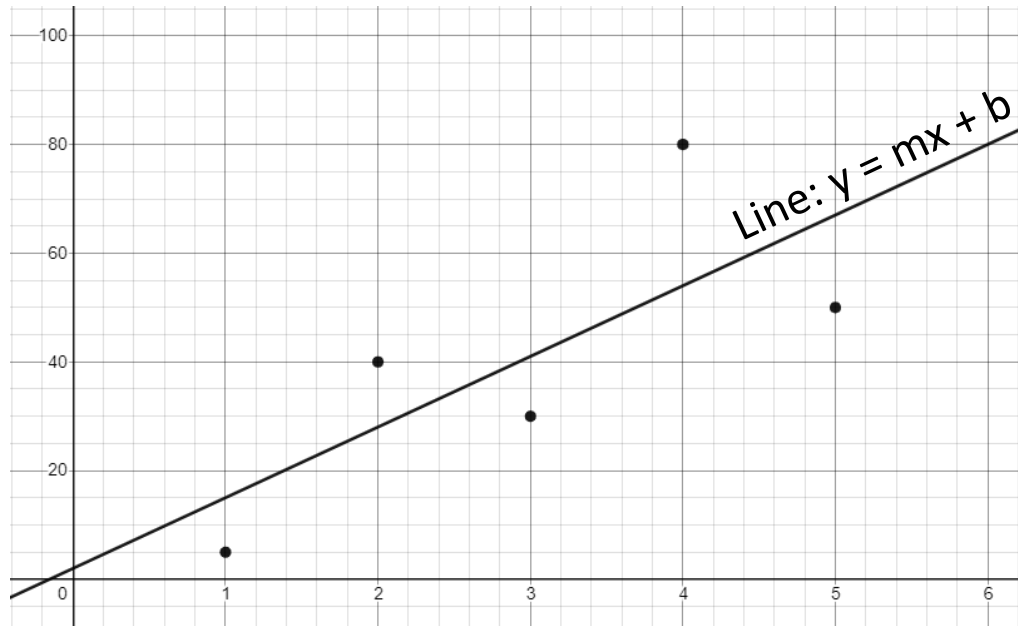
Begin with a random collection of
coordinate points.



Draw a line roughly through the center of
the field of points.

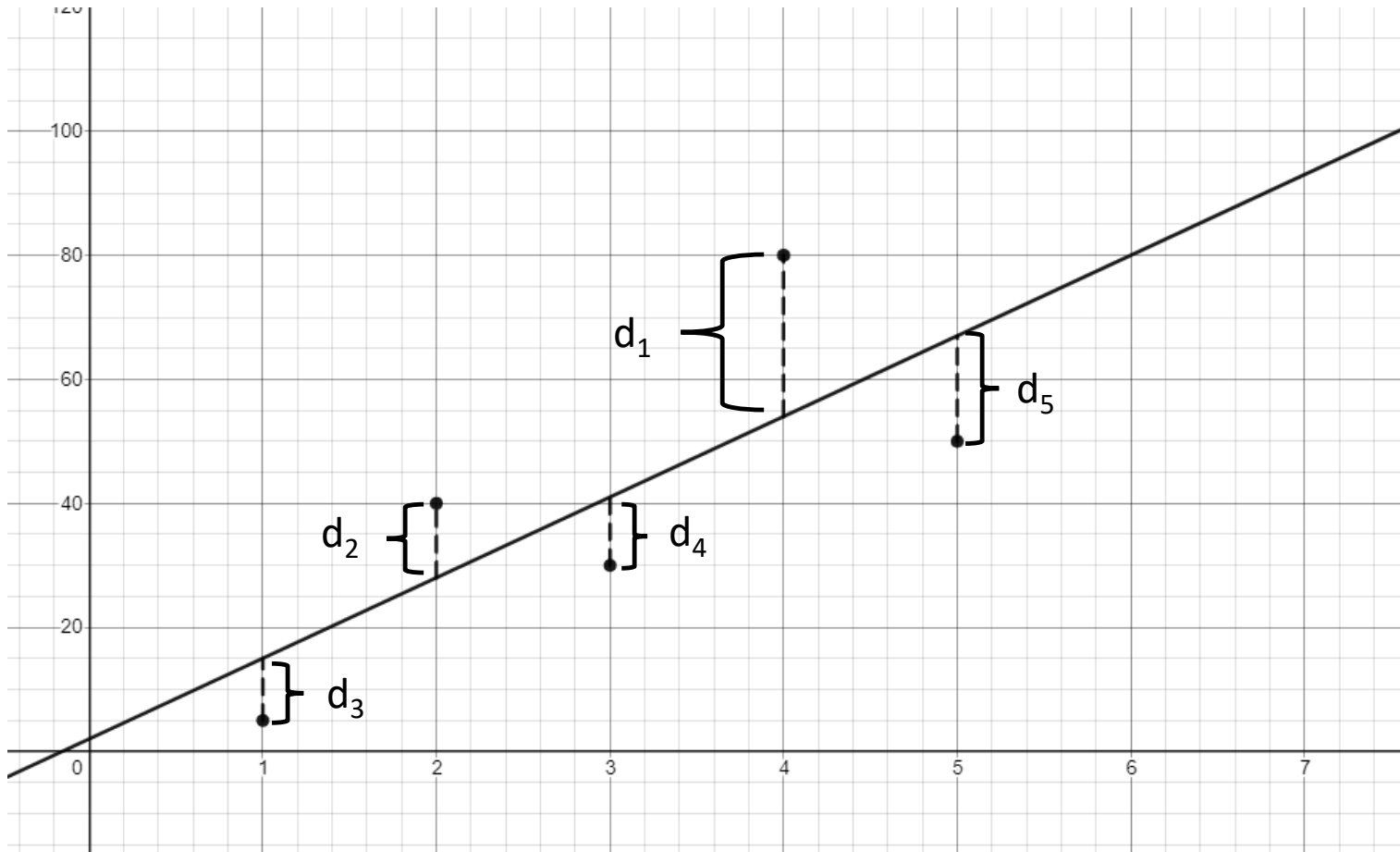


This line, since we don't know its equation yet, can be given the generic equation of $y = mx + b$, for a slope of m and a y – intercept of b .



So, our goal for this process is to determine numeric values of m and b that will give this line the best fit to the data points.

Since we have two variables to solve for, m and b ; not x and y since the final equation should be y in terms of x , we will need to create two distinct equations based on this situation.



The first equation is that the total signed distance from each data point to the line should have a sum of zero.

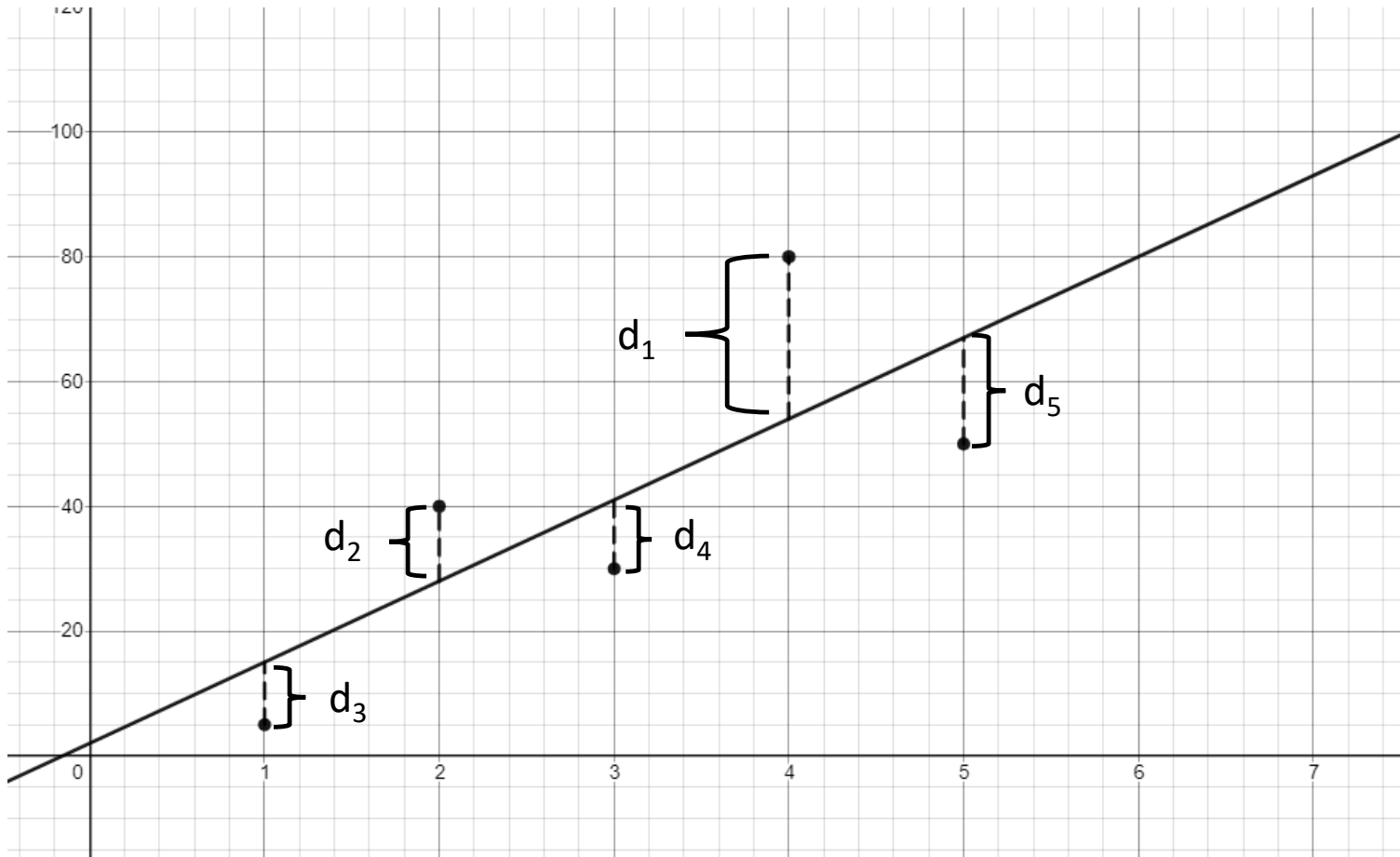
Essentially the total distance to the line of all points above it should equal the total distance for all points below it.

So: $d_1 + d_2 - d_3 - d_4 - d_5 = 0$
or

$$(y_1 - (m \cdot x_1 + b)) + (y_2 - (m \cdot x_2 + b)) + (y_3 - (m \cdot x_3 + b)) + (y_4 - (m \cdot x_4 + b)) + (y_5 - (m \cdot x_5 + b)) = 0$$

$$(y_1 - m \cdot x_1 - b) + (y_2 - m \cdot x_2 - b) + \underbrace{(y_3 - m \cdot x_3 - b) + (y_4 - m \cdot x_4 - b) + (y_5 - m \cdot x_5 - b)} = 0$$

Since these y-values are less than the y-value of the line (at the same x-value) these terms will be negative.



The second equation is that the total positive distance from each data point to the line should be as minimal as possible.

In order to make certain that each distance is going to have a positive measure, the each distance will be squared. [If the squared distance is minimal, the original distance will also be minimal.]

So: $(d_1)^2 + (d_2)^2 + (d_3)^2 + (d_4)^2 + (d_5)^2 = d_{\min}$
or

$$(y_1 - (m \cdot x_1 + b))^2 + (y_2 - (m \cdot x_2 + b))^2 + (y_3 - (m \cdot x_3 + b))^2 + (y_4 - (m \cdot x_4 + b))^2 + (y_5 - (m \cdot x_5 + b))^2 = d_{\min}$$

$$(y_1 - m \cdot x_1 - b)^2 + (y_2 - m \cdot x_2 - b)^2 + (y_3 - m \cdot x_3 - b)^2 + (y_4 - m \cdot x_4 - b)^2 + (y_5 - m \cdot x_5 - b)^2 = d_{\min}$$

What happens now:

Pick 5 random coordinate points (it doesn't matter what values they hold), and substitute them into the two equations. Solve, first for b in terms of m , then for the m that will create the minimal total distance, and finally, use that to solve for b .

For every step of the process, keep track of where the numbers you are calculating came from and how they can be recreated from the original data points.

This is how we will come up the formulas for m and b to create a Least Square Regression Line.

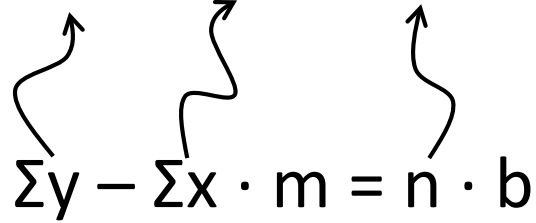
Solution:

My five points: (1, 15), (2, 35), (3, 50), (4, 70), (5, 90)

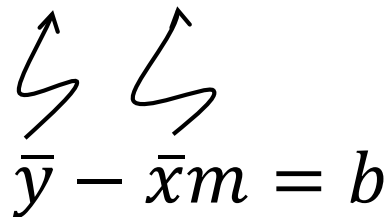
$$(15 - m \cdot 1 - b) + (35 - m \cdot 2 - b) + (50 - m \cdot 3 - b) + (70 - m \cdot 4 - b) + (90 - m \cdot 5 - b) = 0$$

$$(15 - m - b) + (35 - 2m - b) + (50 - 3m - b) + (70 - 4m - b) + (90 - 5m - b) = 0$$

$$260 - 15m = 5b$$


$$\Sigma y - \Sigma x \cdot m = n \cdot b$$

$$52 - 3m = b$$


$$\bar{y} - \bar{x}m = b$$

$$(y_1 - m \cdot x_1 - b)^2 + (y_2 - m \cdot x_2 - b)^2 + (y_3 - m \cdot x_3 - b)^2 + (y_4 - m \cdot x_4 - b)^2 + (y_5 - m \cdot x_5 - b)^2 = d_{\min}$$

$$(y_1 - m \cdot x_1 - b)^2 = (y_1 - m \cdot x_1 - (52 - 3m))^2 = (y_1 - m \cdot x_1 - 52 + 3m)^2 =$$
$$(15 - m \cdot 1 - 52 + 3m)^2 = (-37 + 2m)^2 = 1369 - 148m + 4m^2$$

$$(y_2 - m \cdot x_2 - b)^2 = (y_2 - m \cdot x_2 - (52 - 3m))^2 = (y_2 - m \cdot x_2 - 52 + 3m)^2 =$$
$$(35 - m \cdot 2 - 52 + 3m)^2 = (-17 + m)^2 = 289 - 34m + m^2$$

$$(y_3 - m \cdot x_3 - b)^2 = (y_3 - m \cdot x_3 - (52 - 3m))^2 = (y_3 - m \cdot x_3 - 52 + 3m)^2 =$$
$$(50 - m \cdot 3 - 52 + 3m)^2 = (-2)^2 = 4$$

$$(y_4 - m \cdot x_4 - b)^2 = (y_4 - m \cdot x_4 - (52 - 3m))^2 = (y_4 - m \cdot x_4 - 52 + 3m)^2 =$$
$$(70 - m \cdot 4 - 52 + 3m)^2 = (18 - m)^2 = 324 - 36m + m^2$$

$$(y_5 - m \cdot x_5 - b)^2 = (y_5 - m \cdot x_5 - (52 - 3m))^2 = (y_5 - m \cdot x_5 - 52 + 3m)^2 =$$
$$(90 - m \cdot 5 - 52 + 3m)^2 = (38 - 2m)^2 = 1444 - 152m + 4m^2$$

$$(y_1 - m \cdot x_1 - b)^2 + (y_2 - m \cdot x_2 - b)^2 + (y_3 - m \cdot x_3 - b)^2 + (y_4 - m \cdot x_4 - b)^2 + (y_5 - m \cdot x_5 - b)^2 = d_{\min}$$

$$1369 - 148m + 4m^2 + 289 - 34m + m^2 + 4 + 324 - 36m + m^2 + 1444 - 152m + 4m^2$$

$$(y - \bar{y})^2$$

$$2 \cdot (x - \bar{x})(y - \bar{y})$$

$$(x - \bar{x})^2$$

$$(y_1 - m \cdot x_1 - b)^2 + (y_2 - m \cdot x_2 - b)^2 + (y_3 - m \cdot x_3 - b)^2 + (y_4 - m \cdot x_4 - b)^2 + (y_5 - m \cdot x_5 - b)^2 = d_{\min}$$

There is a "ghost" 1

And a "ghost" 0

There is another "ghost" 1

$$1369 - 148m + 4m^2 + 289 - 34m + 1m^2 + 4 + 0 + 324 - 36m + 1m^2 + 1444 - 152m + 4m^2$$

$$(y - \bar{y})^2$$

$$2 \cdot (x - \bar{x})(y - \bar{y})$$

$$(x - \bar{x})^2$$

$$(y_1 - m \cdot x_1 - b)^2 + (y_2 - m \cdot x_2 - b)^2 + (y_3 - m \cdot x_3 - b)^2 + (y_4 - m \cdot x_4 - b)^2 + (y_5 - m \cdot x_5 - b)^2 = d_{\min}$$

Same "ghost" 0

$$1369 - 148m + 4m^2 + 289 - 34m + m^2 + 4 + 0x + 324 - 36m + m^2 + 1444 - 152m + 4m^2$$

$$(y - \bar{y})^2$$

$$2 \cdot (x - \bar{x})(y - \bar{y})$$

$$(x - \bar{x})^2$$

$$(y_1 - m \cdot x_1 - b)^2 + (y_2 - m \cdot x_2 - b)^2 + (y_3 - m \cdot x_3 - b)^2 + (y_4 - m \cdot x_4 - b)^2 + (y_5 - m \cdot x_5 - b)^2 = d_{\min}$$

$$1369 - 148m + 4m^2 + 289 - 34m + m^2 + 4 + 324 - 36m + m^2 + 1444 - 152m + 4m^2$$

$$3430 - 370m + 10m^2 = d_{\min}$$

$\sum (y - \bar{y})^2$ $2 \cdot \sum (x - \bar{x})(y - \bar{y})$ $\sum (x - \bar{x})^2$

$$(y_1 - m \cdot x_1 - b)^2 + (y_2 - m \cdot x_2 - b)^2 + (y_3 - m \cdot x_3 - b)^2 + (y_4 - m \cdot x_4 - b)^2 + (y_5 - m \cdot x_5 - b)^2 = d_{\min}$$

$$1369 - 148m + 4m^2 + 289 - 34m + m^2 + 4 + 324 - 36m + m^2 + 1444 - 152m + 4m^2$$

$$3430 - 370m + 10m^2 = d_{\min}$$

Since this is a quadratic (in terms of m), the minimum would occur at the vertex.

$$m = \frac{-b}{2a} \text{ (Note: the } b \text{ on this line is in reference to } ax^2 + bx + c, \text{ not the } y = mx + b \text{ used earlier.)}$$

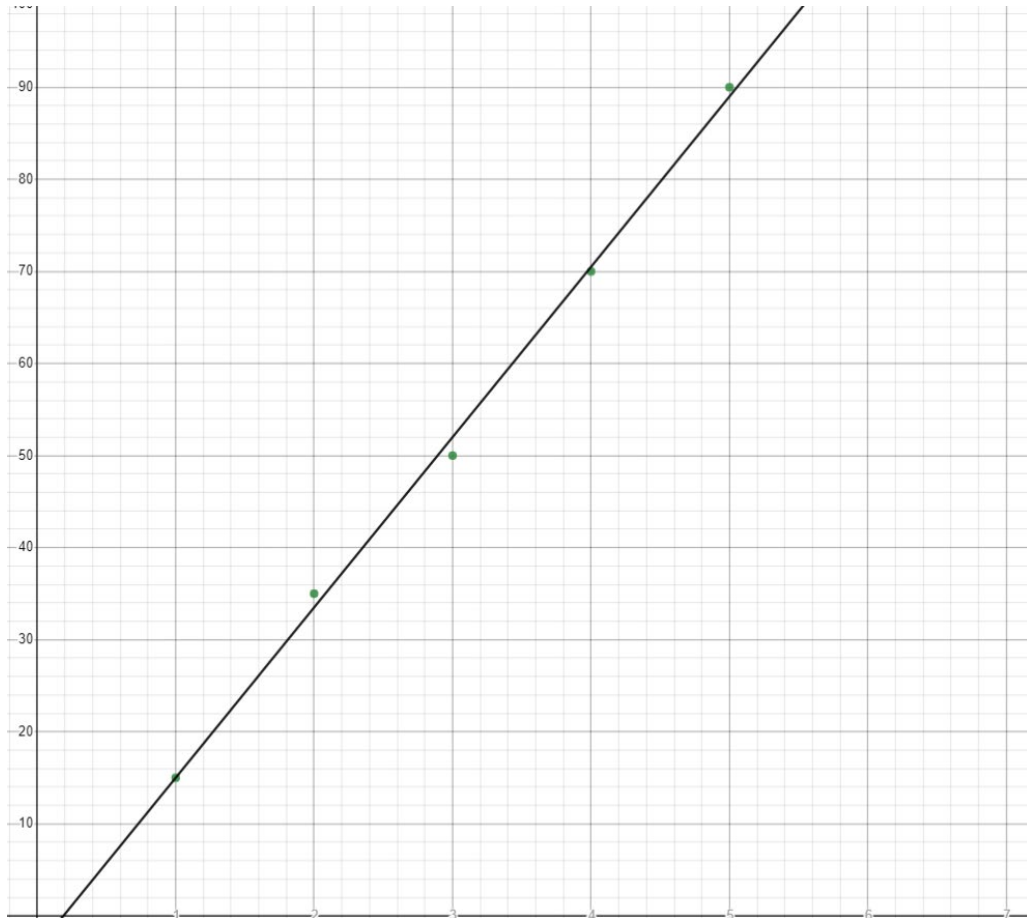
$$m = \frac{-(-370)}{2 \cdot 10} = \frac{370}{20} = 18.5$$

And, using the formula for y -intercept calculated earlier (b is now back to the $y = mx + b$ version.)

$$b = 52 - 3m = 52 - 3(18.5) = -3.5$$

So, the Least Squares Regression Line is: $\hat{y} = 18.5x - 3.5$

The Result (for the specific example)



(1, 15), (2, 35), (3, 50), (4, 70), (5, 90)

$$\hat{y} = 18.5x - 3.5$$

$$\sum (y - \bar{y})^2 - 2m \sum (x - \bar{x})(y - \bar{y}) + m^2 \sum (x - \bar{x})^2 = d$$

So, for a minimum d-value:

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

Now, to simplify this into terms we already know:

$$s_x^2 = \frac{\sum (x - \bar{x})^2}{n - 1} \rightarrow (n - 1)s_x^2 = \sum (x - \bar{x})^2$$

So:

$$m = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n - 1)s_x^2}$$

$$\begin{aligned}
m &= \frac{\sum(x - \bar{x})(y - \bar{y})}{(n - 1)s_x^2} = \frac{1}{n - 1} \cdot \frac{1}{s_x} \frac{\sum(x - \bar{x})(y - \bar{y})}{s_x} \\
&= \frac{1}{n - 1} \cdot \frac{1}{s_x} \cdot \frac{s_y}{s_y} \frac{\sum(x - \bar{x})(y - \bar{y})}{s_x} = \frac{1}{n - 1} \cdot \frac{s_y}{s_x} \cdot \frac{\sum(x - \bar{x})(y - \bar{y})}{s_x s_y} = \\
&\quad \frac{1}{n - 1} \cdot \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right) \cdot \frac{s_y}{s_x} = r \cdot \frac{s_y}{s_x}
\end{aligned}$$

So, for $y = mx + b$:

$$m = r \cdot \frac{s_y}{s_x}$$

and

$$b = \bar{y} - \bar{x}m$$

....just to keep statisticians and analysts (disciplines such as algebra and calculus) sitting at separate tables in the lunch room, we use: $\hat{y} = a + bx$ for a slope of b and a y-intercept of a .