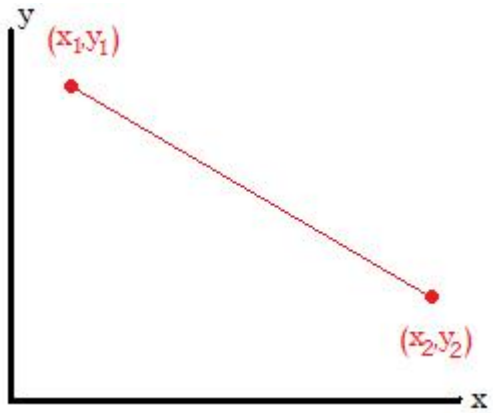
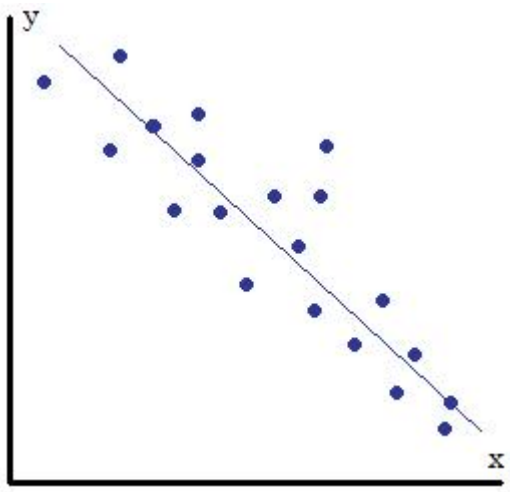
Calculating Line Regression by Hand

When there are more than 2 points of data it is usually impossible to find a line that goes exactly through all the points. But, usually we can find a line (or curve) that is a good approximation to the data. For most science fair projects, a line of best fit is what is needed, and that's what we will be finding on this page.



$$y = y_1 + (x-x_1)\frac{y_2 - y_1}{x_2 - x_1}$$



???

If you have a calculator, bring it out because it will be helpful in adding up all of these numbers. Here's our sample data given below:

x	4.1	6.5	12.6	25.5	29.8	38.6	46	52.8	59.6	66.3	74.7
y	2.2	4.5	10.4	23.1	27.9	36.8	44.3	50.7	57.5	64.1	72.6

It is important for us to keep our numbers straight, so we have created a few variables below which we defined to the right.

- x_{sum} - The sum of all the values in the x column.
- y_{sum} - The sum of all the values in the y column.
- xy_{sum} - The sum of the products of the x_n and y_n that are recorded at the same time (vertical on this chart).
- x^2_{sum} - The total of each value in the x column squared and then added together.
- y^2_{sum} - The total of each value in the y column squared and then added together.
- N - The total number of elements (or trials in your experiment).

For our example, here's how you would calculate these:

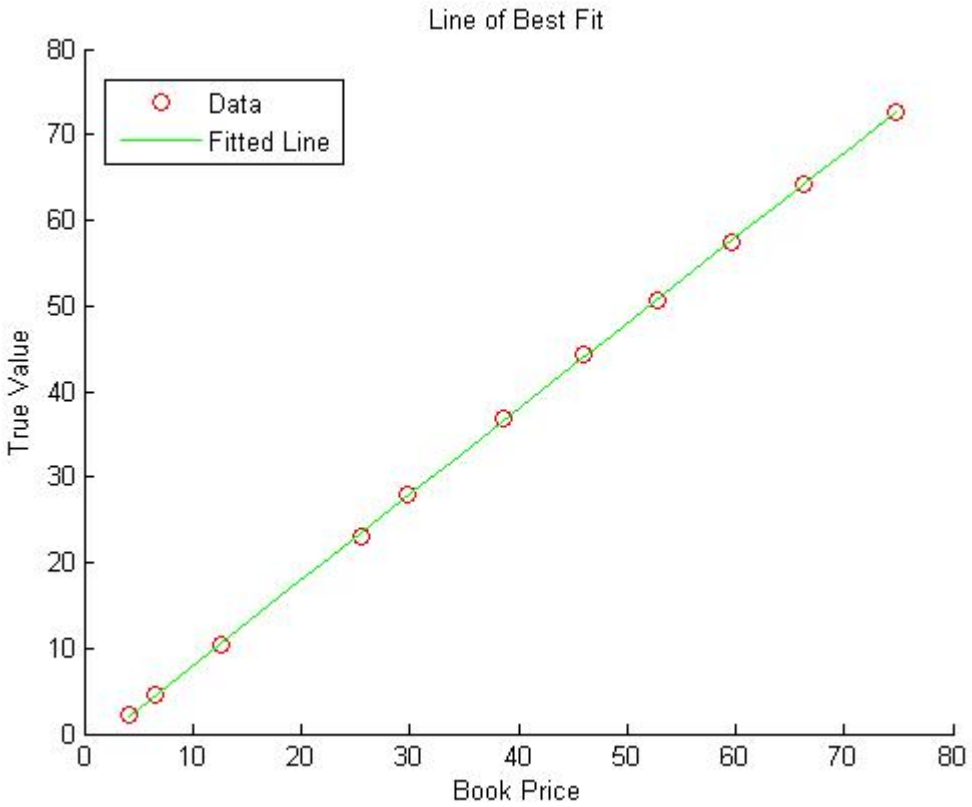
- x_{sum} = 4.1 + 6.5 + 12.6 + 25.5 + 29.8 + 38.6 + 46 + 52.8 + 59.6 + 66.3 + 74.7 = 416.5
- y_{sum} = 2.2 + 4.5 + 10.4 + 23.1 + 27.9 + 36.8 + 44.3 + 50.7 + 57.5 + 64.1 + 72.6 = 394.1
- xy_{sum} = 4.1*2.2 + 6.5*4.5 + 12.6*10.4 + 25.5*23.1 + 29.8*27.9 + 38.6*36.8 + 46*44.3 + 52.8*50.7 + 59.6*57.5 + 66.3*64.1 + 74.7*72.6 = 20825
- x^2_{sum} = $4.1^2 + 6.5^2 + 12.6^2 + 25.5^2 + 29.8^2 + 38.6^2 + 46^2 + 52.8^2 + 59.6^2 + 66.3^2 + 74.7^2$ = 21678
- y^2_{sum} = $2.2^2 + 4.5^2 + 10.4^2 + 23.1^2 + 27.9^2 + 36.8^2 + 44.3^2 + 50.7^2 + 57.5^2 + 64.1^2 + 72.6^2$ = 20018
- N = 11

The best form for our line is slope-intercept form, which looks like $y = mx + b$. Therefore, it is only necessary to compute m and b to determine the best fit line. Those values can be computed by the following equations:

$$m = \frac{(Nxy_{sum}) - (x_{sum}y_{sum})}{(Nx^2_{sum}) - (x_{sum}x_{sum})} \qquad b = \frac{(x^2_{sum}y_{sum}) - (x_{sum}xy_{sum})}{(Nx^2_{sum}) - (x_{sum}x_{sum})}$$

After plugging in the values that we found, we get: m = .99992 and b = -2.0067.

This means that the equation of the line is y = .99992x + -2.0067, or y = .99992x - 2.0067. We graphed the results using Matlab, but you can even make a graph by hand.



Example best fit line

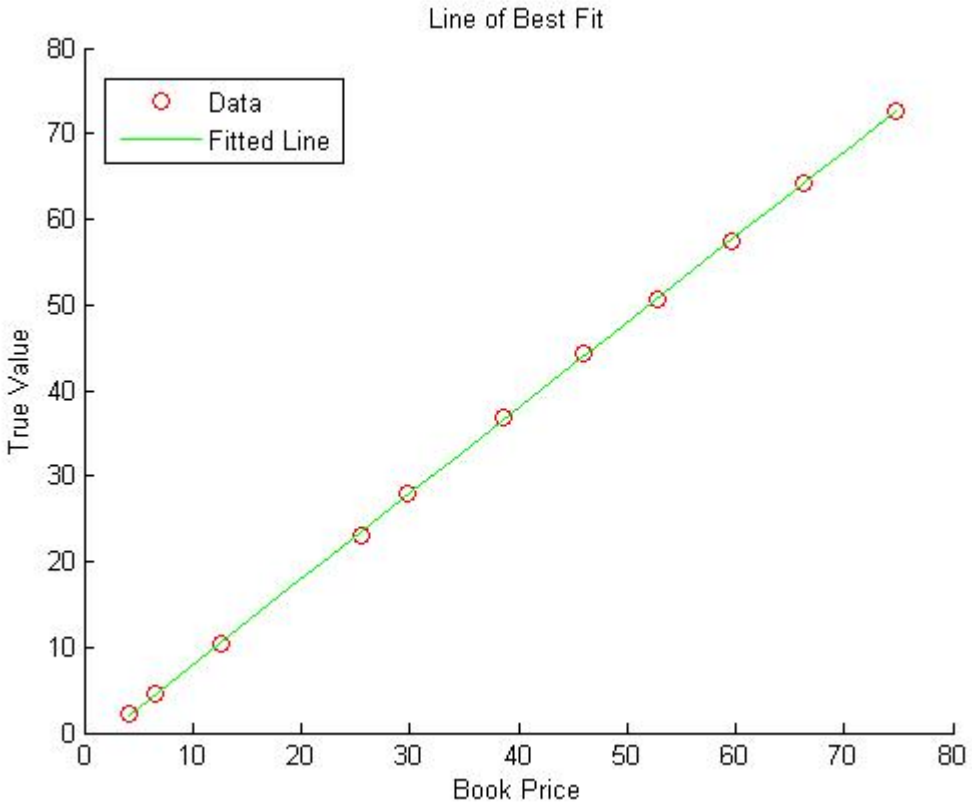
For a more detailed mathematical explanation of this topic, see **MathWorld**. Beware, calculus is needed.

This line looks like it accurately depicts the data, but can we be sure simply by looking at it?

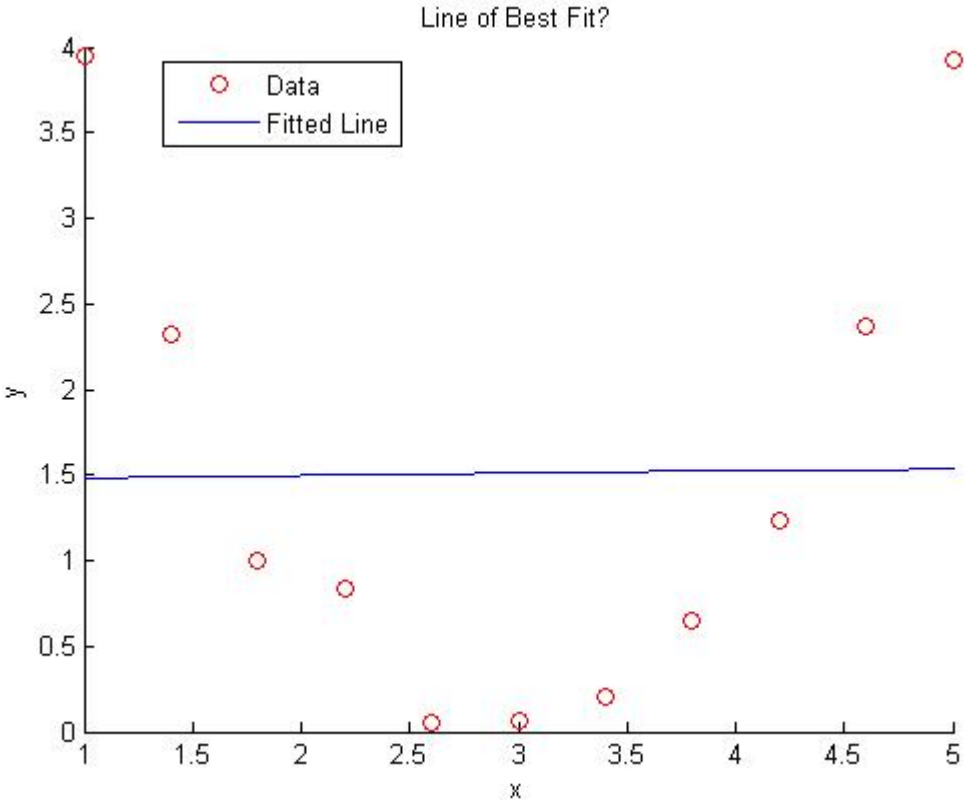
The Correlation

Correlation is a measue of how close the line fits the points that you found in your experiment. Anyone can say whether they think the line is a good fit or not, but to measure it exactly, we use the correlation coefficient: R^2.

Look at the following two plots. In the first, the line fits very well, and in the second, then line fits very poorly.



$R^2 = .99992$



$R^2 = .000142$

Notice the value R^2 differs in both graphs. The lines that fit the points best have a value of R. R is the correlation coefficient.

$$R^2 = \frac{(Nxy_{sum} - x_{sum}y_{sum})^2}{(Nx^2_{sum} - x_{sum}x_{sum})(Ny^2_{sum} - y_{sum}y_{sum})}$$

From comparing the graphs to the R values, you can probably see that the closer R is to 1, the better the line fits your data. When R is far from 1, your line will not represent the data at all. This is easily seen above, and for more information please see **MathWorld**.

To see how to quickly find the equation of the best fit line and the correlation coefficient using Microsoft Excel (or Open Office Software), visit our **Excel Line Regression** webpage.