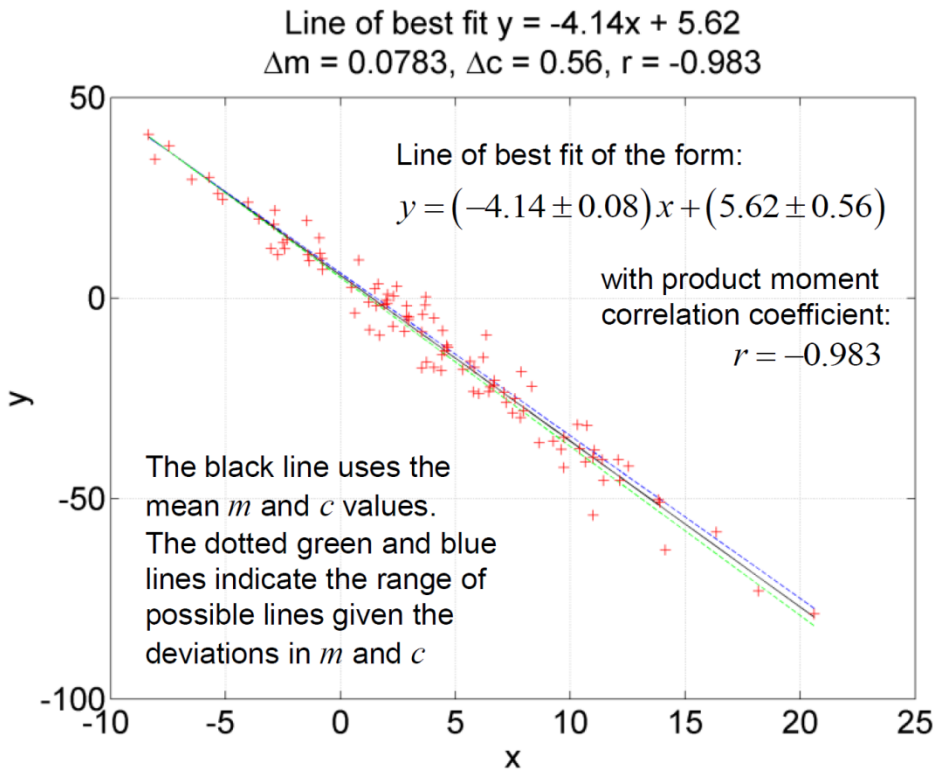
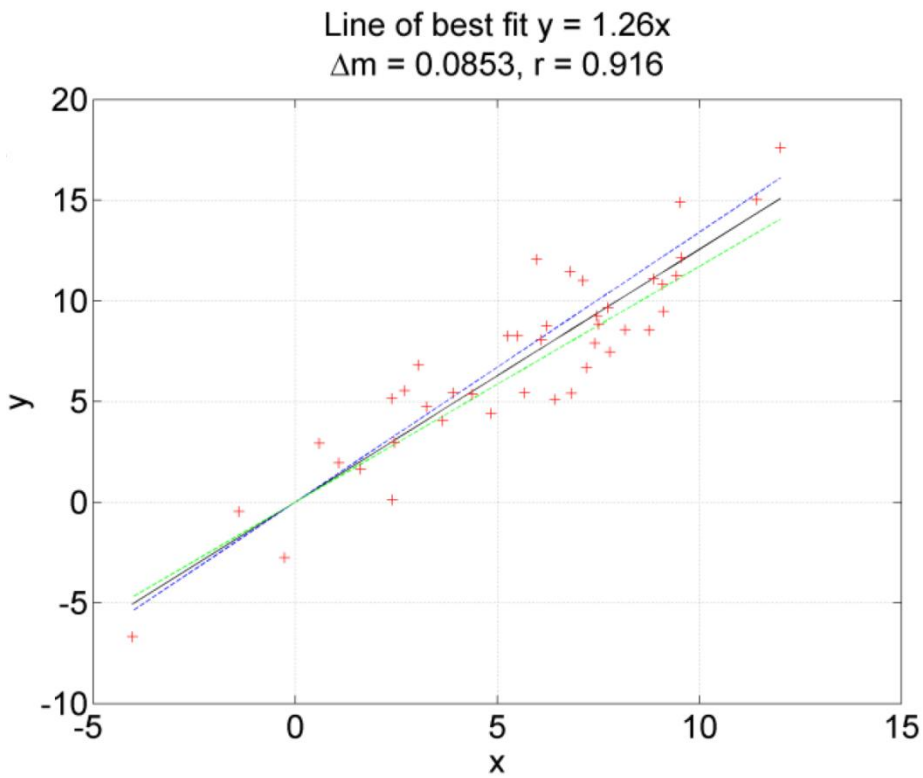
A recipe for finding lines of best fit

$$y = mx + c$$

$$y = mx$$





Correlation & Linear Regression

Perhaps the most important analytical tool in the physical sciences is the ability to quantify the validity of a model relating a set of measurable parameters. The idea is as follows:

(1) Rearrange the model in such a way that it becomes a *linear equation* of the form $y = mx + c$

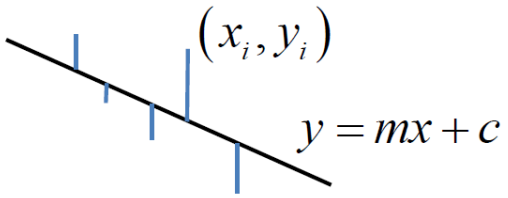
(2) Plot experimental (x,y) data on a graph and determine the **line of best fit** through the data.

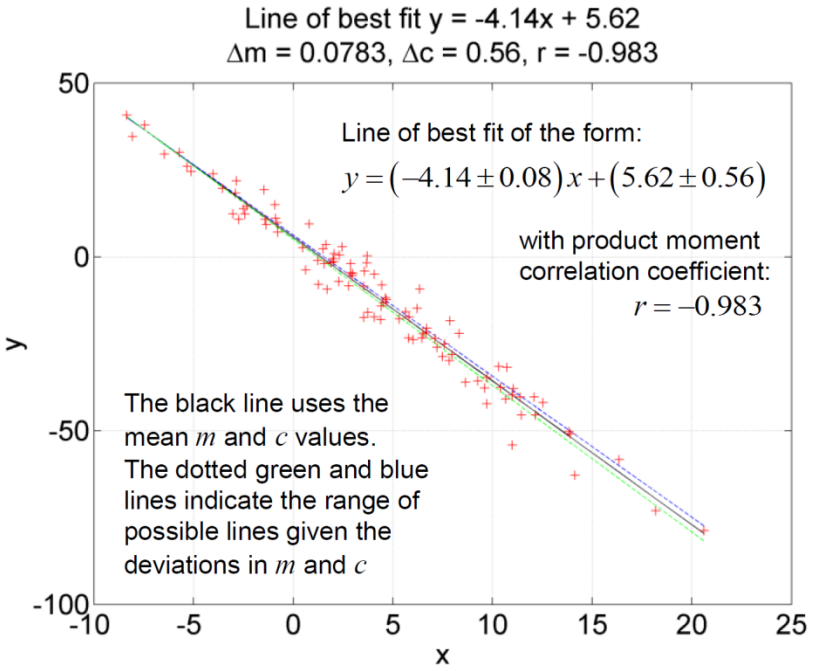
(3) Determine *gradient* m and *vertical intercept* c from the line of best fit.

(4) Determine the standard deviation of both gradient m and intercept c, and a quantitative measure of how good the fit is (this is called the **product moment correlation coefficient**).

To determine the line of best fit*, let us sum the *squared* deviations of (x,y) from the line of best fit.

$$S = \sum_{i=1}^{N} \left(y_i - mx_i - c \right)^2$$

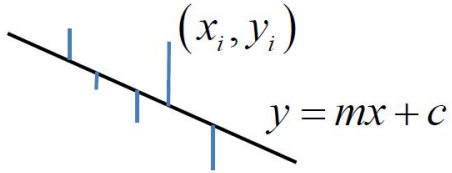




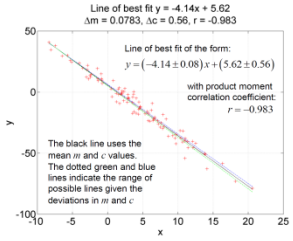
We can find S given a range of m and c values. Which pairing results in the *minimum* value of S?

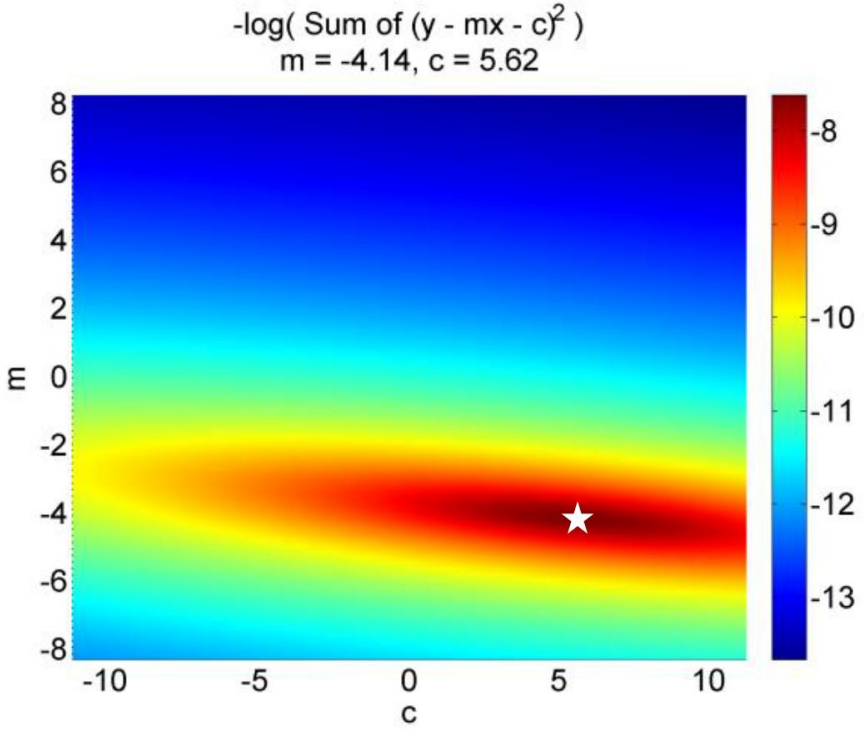
To determine the line of best fit*, let us sum the *squared* deviations of (x,y) from the line of best fit.

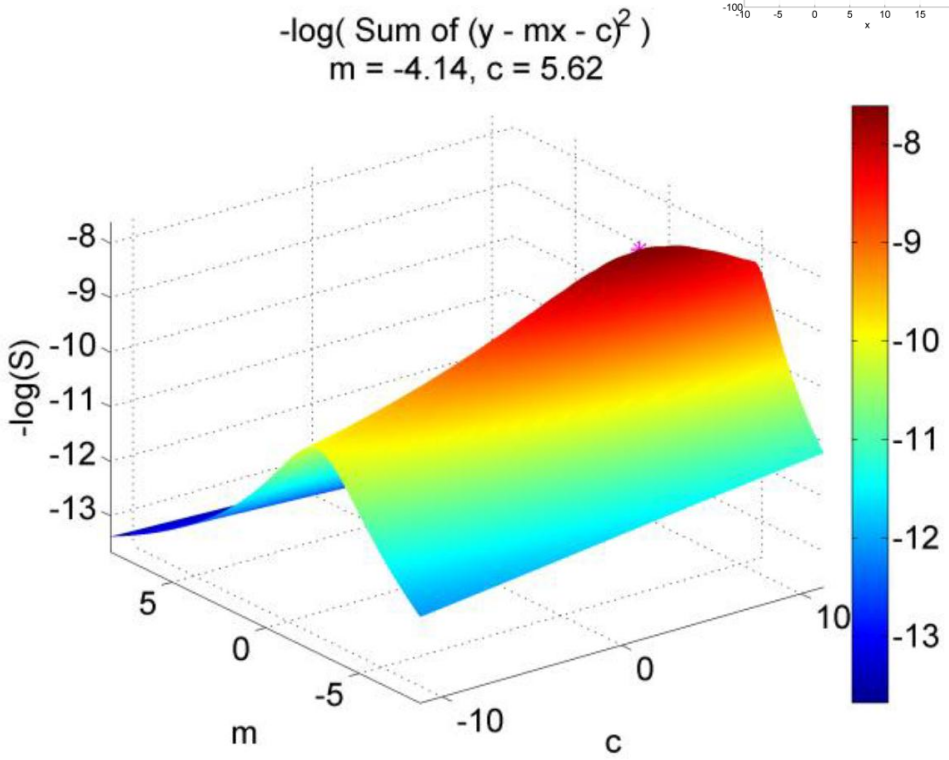
$$S = \sum_{i=1}^{N} \left(y_i - mx_i - c \right)^2$$



Using the (*negatively correlated)* data on the right, we can plot a surface of S vs m and c values We can see this has a **minimum** at a particular (m,c) coordinate. (Note for clarity the plots below are of $-\log S$, so the (m,c) coordinate corresponds to the peak, i.e. maximum, instead).







The minimum of S can be found by differentiating S with respect to m and c, and setting these expressions equal to zero. Since S is a function of two variables we must use *partial derivatives.*

$$S = \sum_{i=1}^{N} \left(y_i - mx_i - c \right)^2 \qquad S = \sum_{i=1}^{N} \left(y_i - mx_i - c \right)^2$$

$$\frac{\partial S}{\partial m} = 2\sum_{i=1}^{N} \left(y_i - mx_i - c \right)\left(-x_i \right) \qquad \frac{\partial S}{\partial c} = 2\sum_{i=1}^{N} \left(y_i - mx_i - c \right)\left(-1 \right)$$

$$\therefore \frac{\partial S}{\partial m} = 0 \Rightarrow \sum_{i=1}^{N} x_i \left(y_i - mx_i - c \right) = 0 \qquad \therefore \frac{\partial S}{\partial c} = 0 \Rightarrow \sum_{i=1}^{N} \left(y_i - mx_i - c \right) = 0$$

$$\therefore \sum_{i=1}^{N} x_i y_i - m\sum_{i=1}^{N} x_i^2 - c\sum_{i=1}^{N} x_i = 0 \qquad \therefore \sum_{i=1}^{N} y_i - m\sum_{i=1}^{N} x_i - cN = 0$$

Hence:

$$\sum_{i=1}^{N} x_i y_i - m\sum_{i=1}^{N} x_i^2 - c\sum_{i=1}^{N} x_i = 0 \quad \therefore \overline{xy} - m\overline{x^2} - c\overline{x} = 0$$

$$\sum_{i=1}^{N} y_i - m\sum_{i=1}^{N} x_i - cN = 0 \quad \therefore \overline{y} - m\overline{x} - c = 0$$

Therefore:

$$c = \overline{y} - m\overline{x}$$

$$\therefore \overline{xy} - m\overline{x^2} - \left(\overline{y} - m\overline{x} \right)\overline{x} = 0$$

$$\therefore m\left(\overline{x}^2 - \overline{x^2} \right) + \overline{xy} - \overline{yx} = 0$$

$$\therefore m = \frac{\overline{xy} - \overline{yx}}{\overline{x^2} - \overline{x}^2} = \frac{\mathrm{cov}[x, y]}{V[x]}$$

If we repeat the analysis for the line: $x = My + d \Rightarrow M = \dfrac{\mathrm{cov}[x, y]}{V[y]}$

If this was the *same line but rearranged*:

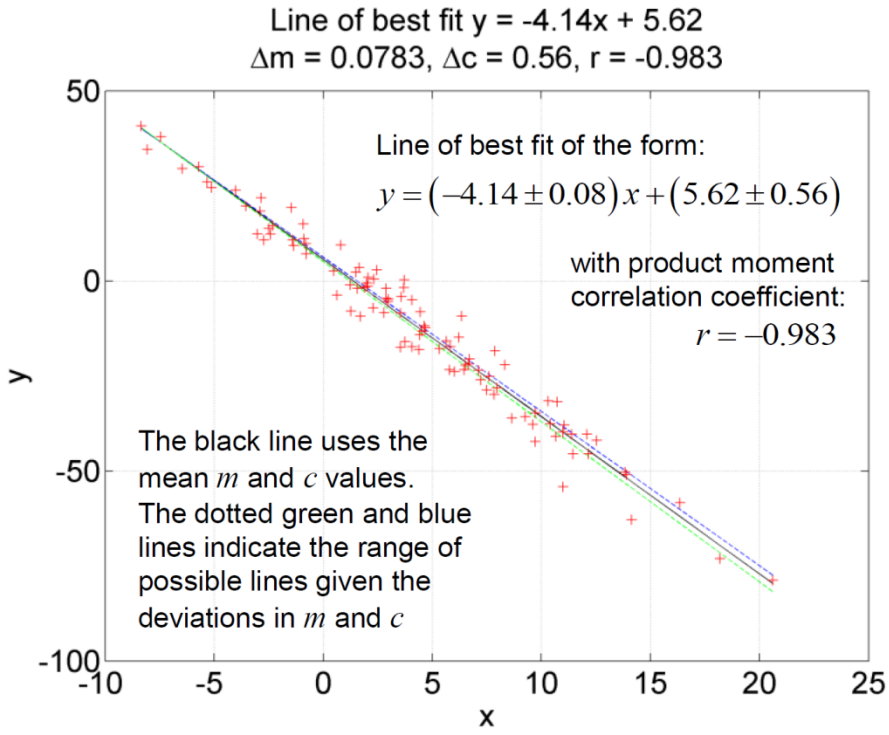
$M = \frac{1}{m} \quad \therefore mM = 1$

$y = mx + c$

Hence define a **product moment correlation coefficient:**

$$r = \frac{\mathrm{cov}[x, y]}{\sqrt{V[x]V[y]}}$$

This will be +1 for a perfect positive correlation and -1 for a perfect negative correlation (i.e. $S = 0$ in both cases).



It is possible to show* that the standard deviations (i.e. 'errors') in m and c are:

$$\Delta m = \frac{s}{\sqrt{N}} \frac{1}{\sqrt{V[x]}}$$

$$\Delta c = \frac{s}{\sqrt{N}} \sqrt{1 + \frac{\bar{x}^2}{V[x]}}$$

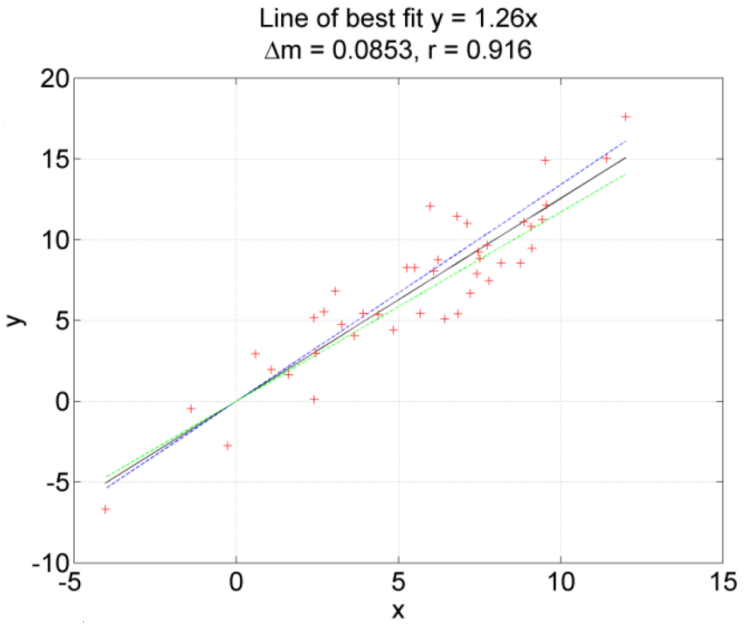
$$s = \sqrt{\frac{1}{N-2} \sum_{i=1}^{N} \left(y_i - m x_i - c \right)^2}$$

This is very useful in the physical sciences, as the errors in m and c will often be the uncertainties in model parameters (e.g. the strength of gravity...)

s is the *unbiased estimator* of the standard deviation in the y values from the line of best fit. The N-2 factor is due to two parameters (m and c) being used in the calculation, which are of course derived from the sample data themselves as shown above.

*http://mathworld.wolfram.com/LeastSquaresFitting.html

In many situations a **direct proportion** is asserted between y and x. The computation of the line of best fit (which passes through $(0,0)$ follows a similar argument to the one above.



$$S = \sum_{i=1}^{N} \left(y_i - mx_i \right)^2$$

$$\frac{\partial S}{\partial m} = 2\sum_{i=1}^{N} \left(y_i - mx_i \right)\left(-x_i \right)$$

$$\therefore \frac{\partial S}{\partial m} = 0 \Rightarrow \sum_{i=1}^{N} x_i \left(y_i - mx_i \right) = 0$$

$$\therefore \sum_{i=1}^{N} x_i y_i - m\sum_{i=1}^{N} x_i^2 = 0$$

$$\therefore m = \frac{\overline{xy}}{\overline{x^2}}$$

The product moment correlation coefficient is the same as before but the standard deviation in m is slightly different since only *one* parameter is used in the computation of s (i.e. m).

$$\Delta m = \frac{s}{\sqrt{N}} \frac{1}{\sqrt{V[x]}}$$

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \left(y_i - mx_i \right)^2}$$

$$r = \frac{\mathrm{cov}[x,y]}{\sqrt{V[x]V[y]}}$$

Summary: Line of Best Fit for: $y = mx + c$

N data point pairs (x,y)

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad \overline{x^2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2, \quad \overline{y^2} = \frac{1}{N}\sum_{i=1}^{N} y_i^2, \quad \overline{xy} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i$$

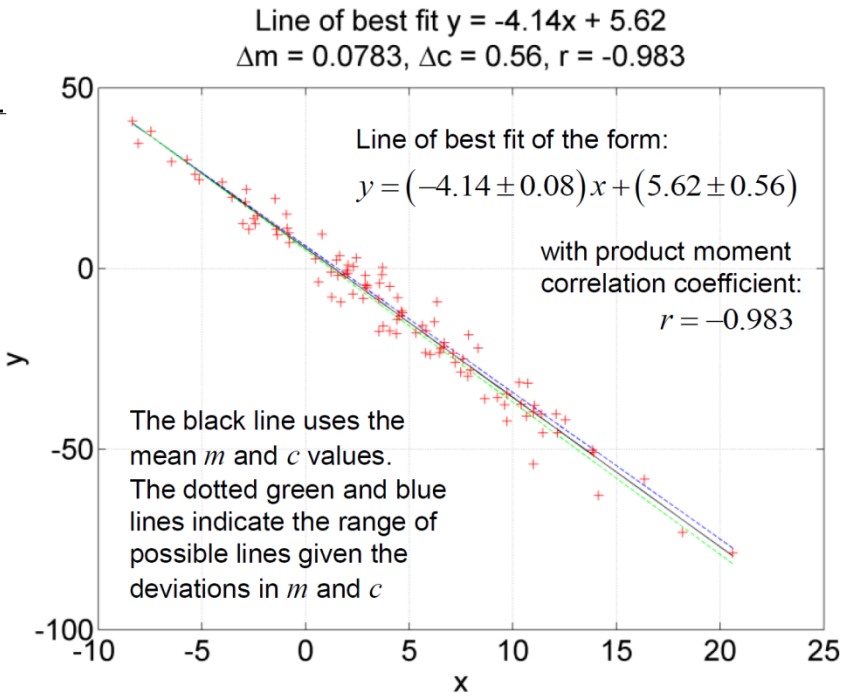
$$V[x] = \overline{x^2} - \bar{x}^2, \quad V[y] = \overline{y^2} - \bar{y}, \quad \mathrm{cov}[x,y] = \overline{xy} - \bar{x}\bar{y}$$

$$m = \frac{\overline{xy} - \bar{y}\bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{\mathrm{cov}[x,y]}{V[x]}, \quad c = \bar{y} - m\bar{x}$$

$$r = \frac{\mathrm{cov}[x,y]}{\sqrt{V[x]V[y]}}$$

$$\Delta m = \frac{s}{\sqrt{N}}\frac{1}{\sqrt{V[x]}}, \quad \Delta c = \frac{s}{\sqrt{N}}\sqrt{1 + \frac{\bar{x}^2}{V[x]}}$$

$$s = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}\left(y_i - mx_i - c\right)^2}$$



Summary: Line of Best Fit for: $y = mx$

N data point pairs (x,y)

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad \overline{x^2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2, \quad \overline{y^2} = \frac{1}{N}\sum_{i=1}^{N} y_i^2, \quad \overline{xy} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i$$

$$V[x] = \overline{x^2} - \bar{x}^2, \quad V[y] = \overline{y^2} - \bar{y}, \quad \mathrm{cov}[x,y] = \overline{xy} - \overline{xy}$$

$$m = \frac{\overline{xy}}{\overline{x^2}}$$

$$r = \frac{\mathrm{cov}[x,y]}{\sqrt{V[x]V[y]}}$$

$$\Delta m = \frac{s}{\sqrt{N}}\frac{1}{\sqrt{V[x]}}$$

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - mx_i\right)^2}$$