

BPhO

Computational
Challenge

Seminar 03: Lines of Best Fit

Dr Andrew French.
December 2021.

A recipe for finding lines of best fit

$$y = mx + c$$

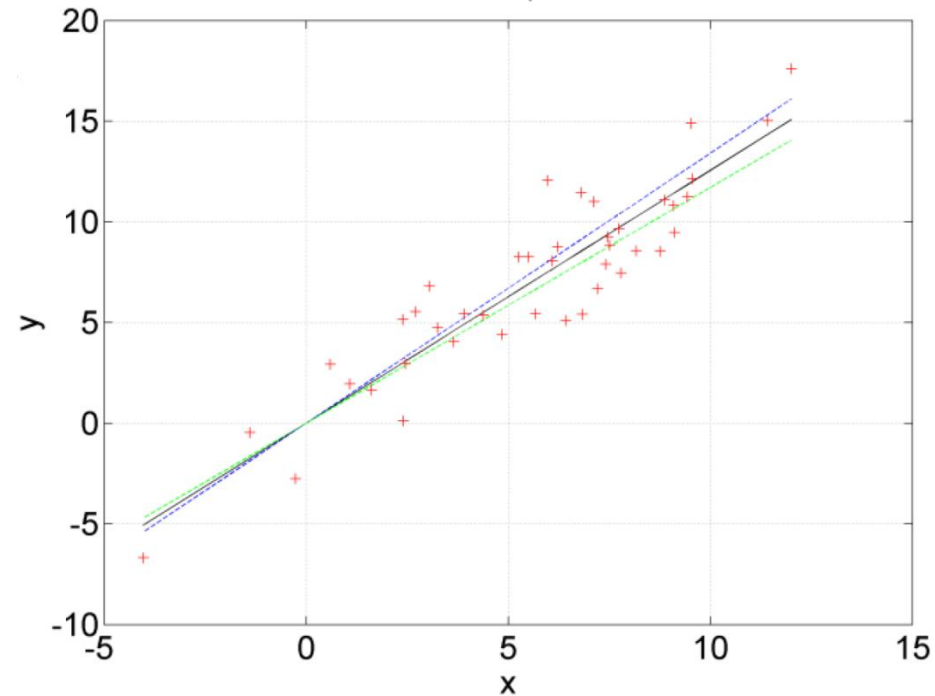
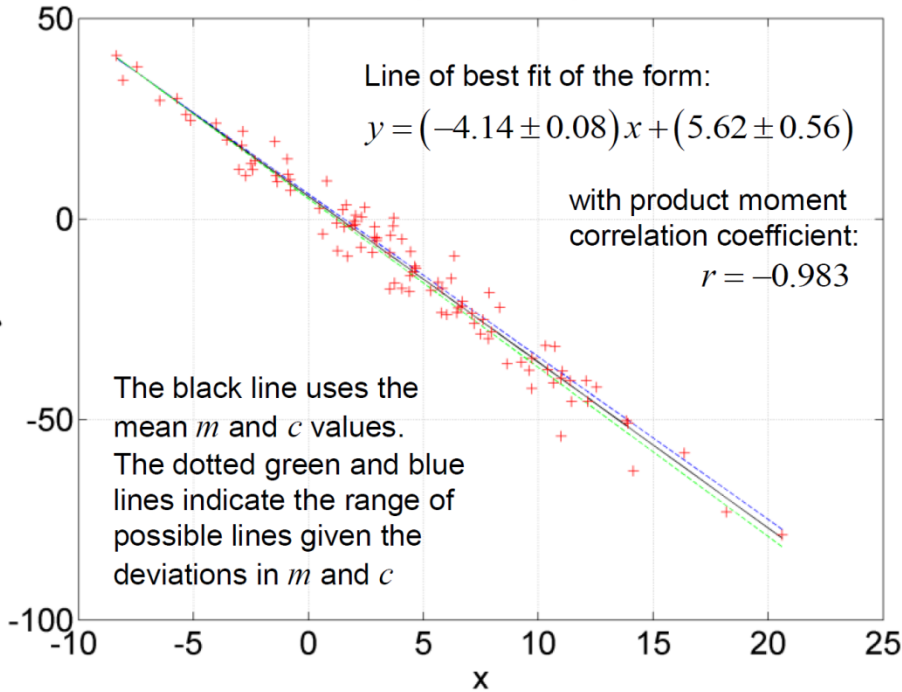
$$y = mx$$

Line of best fit $y = -4.14x + 5.62$
 $\Delta m = 0.0783$, $\Delta c = 0.56$, $r = -0.983$

Line of best fit $y = 1.26x$
 $\Delta m = 0.0853$, $r = 0.916$

Line of best fit of the form:
 $y = (-4.14 \pm 0.08)x + (5.62 \pm 0.56)$
with product moment
correlation coefficient:
 $r = -0.983$

The black line uses the
mean m and c values.
The dotted green and blue
lines indicate the range of
possible lines given the
deviations in m and c



Summary: Line of Best Fit for:

$$y = mx + c$$

N data point pairs (x, y)

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

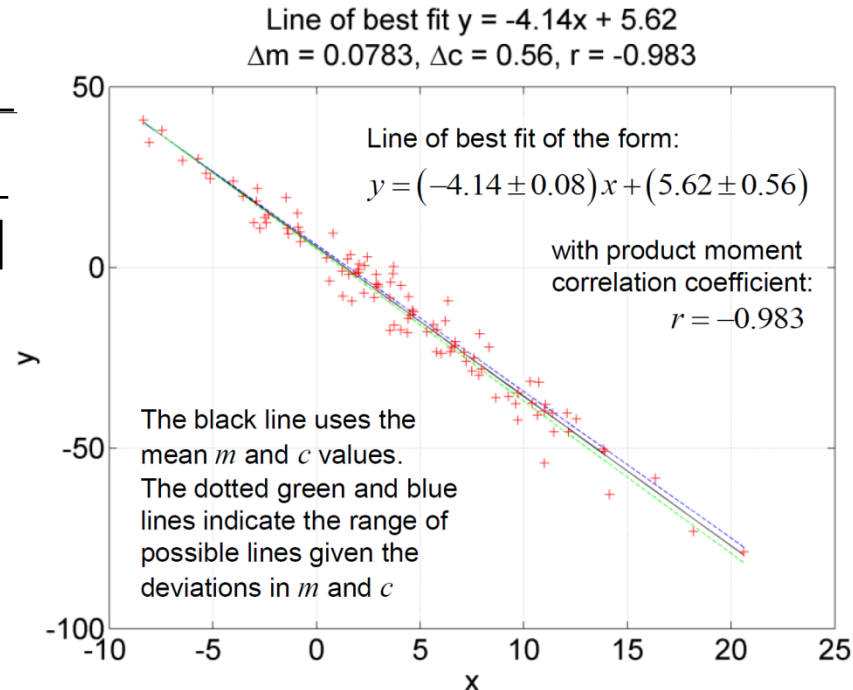
$$V[x] = \overline{x^2} - \bar{x}^2, \quad V[y] = \overline{y^2} - \bar{y}^2, \quad \text{cov}[x, y] = \overline{xy} - \bar{x}\bar{y}$$

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

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

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

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



Summary: Line of Best Fit for:

$$y = mx$$

N data point pairs (x, y)

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

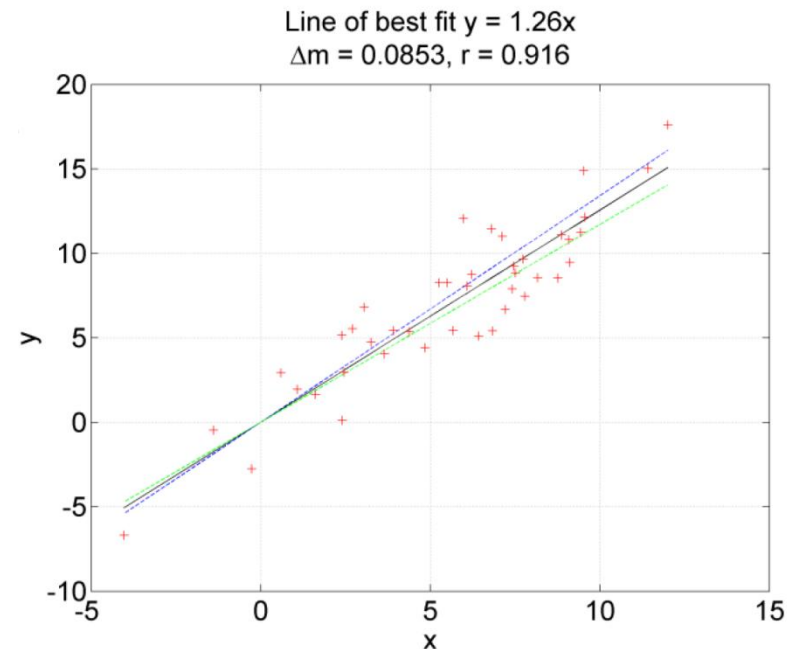
$$V[x] = \overline{x^2} - \bar{x}^2, \quad V[y] = \overline{y^2} - \bar{y}^2, \quad \text{cov}[x, y] = \overline{xy} - \bar{x}\bar{y}$$

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

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

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

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



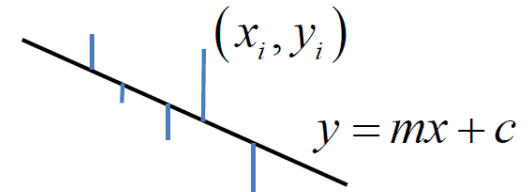
Correlation & Linear Regression

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.

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



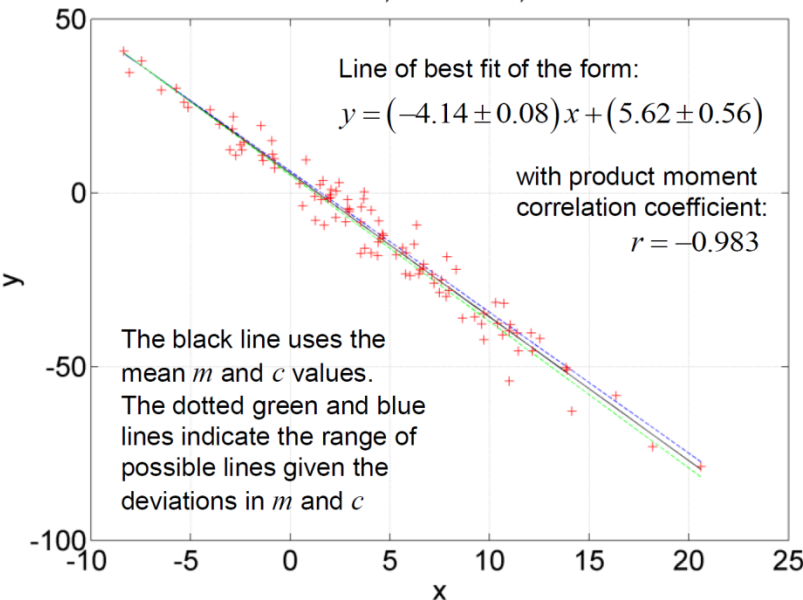
Line of best fit $y = -4.14x + 5.62$
 $\Delta m = 0.0783, \Delta c = 0.56, r = -0.983$

Line of best fit of the form:
 $y = (-4.14 \pm 0.08)x + (5.62 \pm 0.56)$

with product moment
correlation coefficient:
 $r = -0.983$

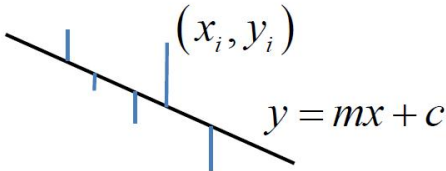
The black line uses the
mean m and c values.
The dotted green and blue
lines indicate the range of
possible lines given the
deviations in m and c

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

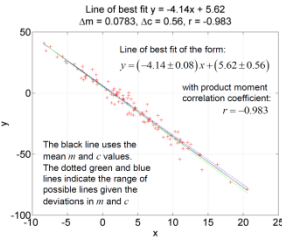


To determine the line of best fit*, let us sum the *squared* deviations of (x,y) from the line of best fit.

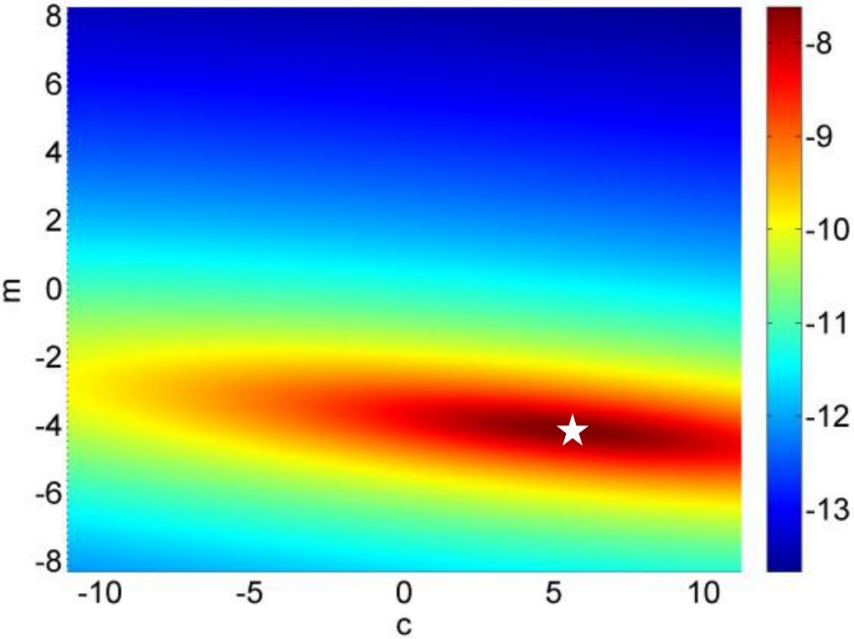
$$S = \sum_{i=1}^N (y_i - mx_i - c)^2$$



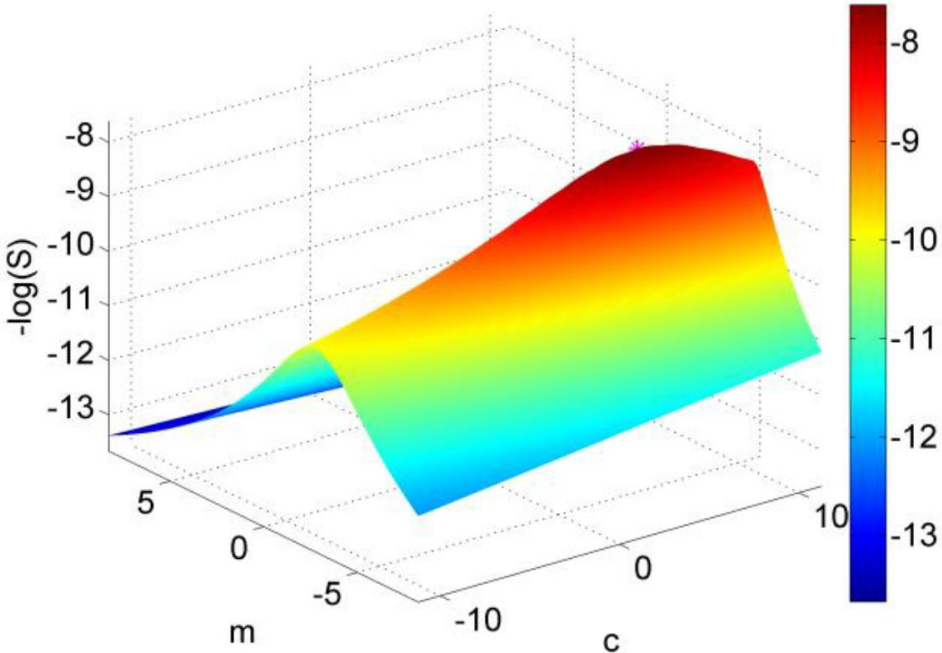
Using the (*negatively correlated*) data on the right, we can plot a surface of S vs m and c values. We can see this has a **minimum** at a particular (m,c) coordinate. (Note for clarity the plots below are of $-\log S$, so the (m,c) coordinate corresponds to the peak, i.e. maximum, instead).



$-\log(\text{Sum of } (y - mx - c)^2)$
 $m = -4.14, c = 5.62$



$-\log(\text{Sum of } (y - mx - c)^2)$
 $m = -4.14, c = 5.62$



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*.

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

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

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

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

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

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

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

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

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:

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

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

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

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

If we repeat the analysis for the line: $x = My + d \Rightarrow M = \frac{\text{cov}[x, y]}{V[y]}$
 If this was the *same line but rearranged*:

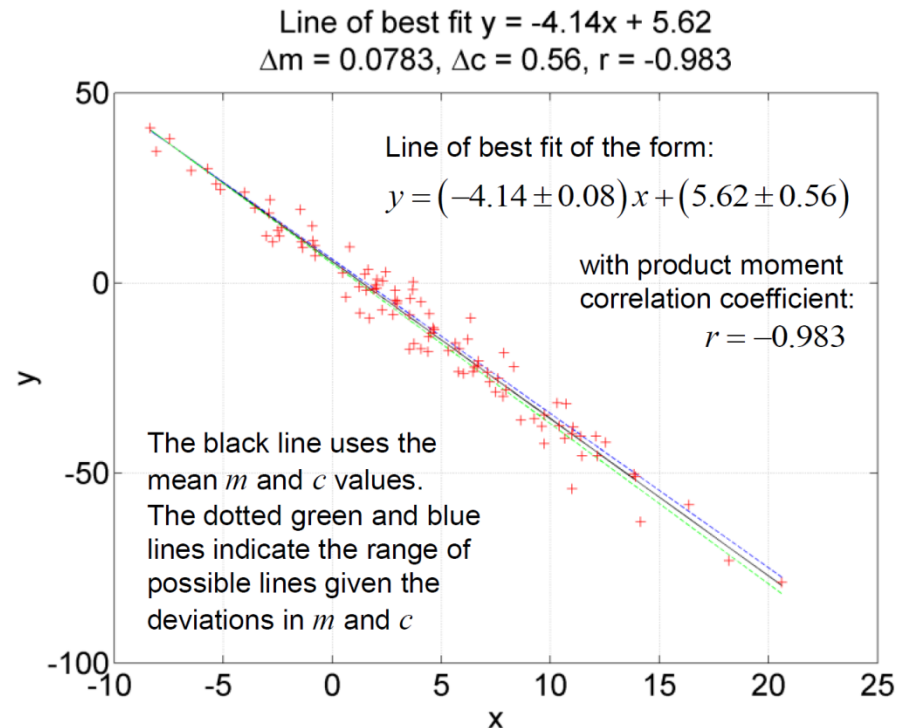
$$M = \frac{1}{m} \quad \therefore mM = 1$$

$$y = mx + c$$

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

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

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:

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

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

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

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.

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

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

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

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

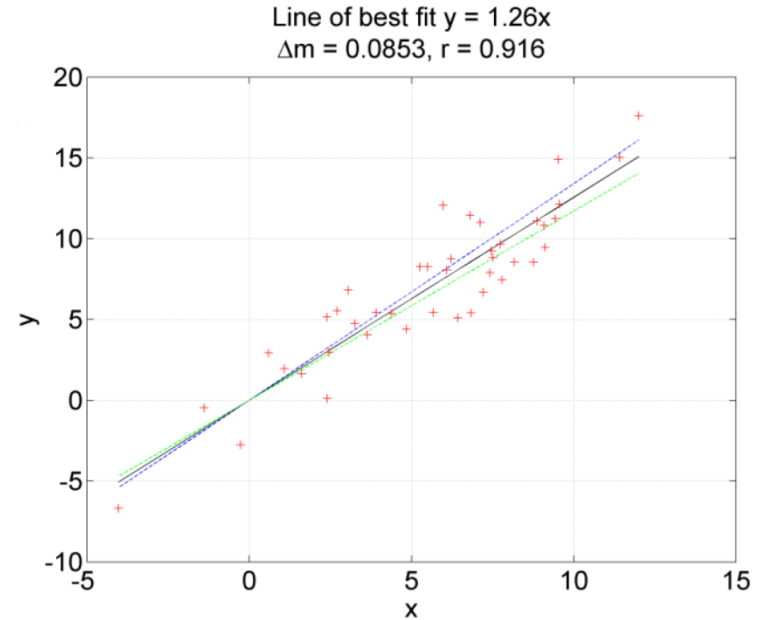
$$\therefore m = \frac{\overline{xy}}{\overline{x^2}}$$

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).

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

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

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



Summary: Line of Best Fit for:

$$y = mx + c$$

N data point pairs (x, y)

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

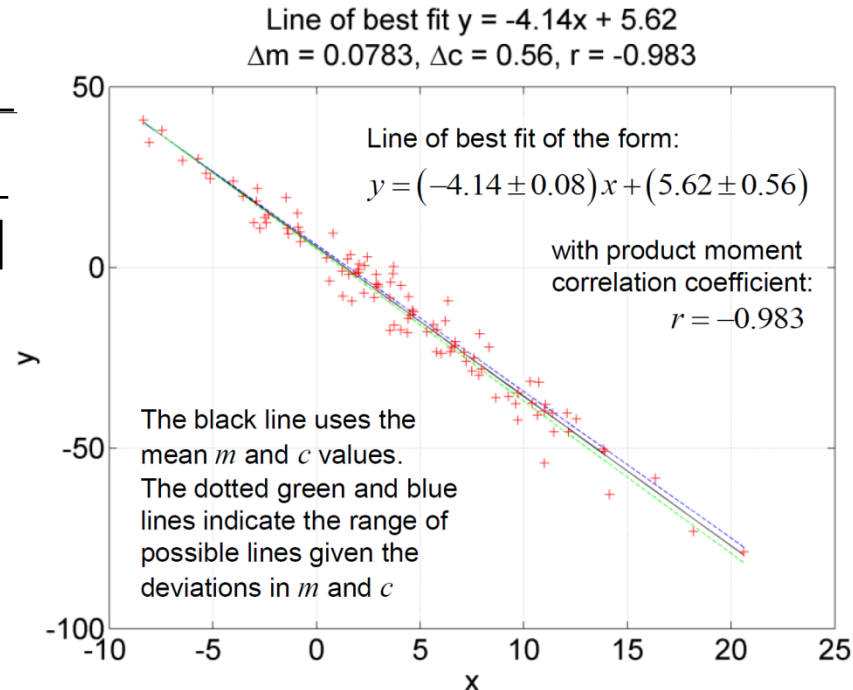
$$V[x] = \overline{x^2} - \bar{x}^2, \quad V[y] = \overline{y^2} - \bar{y}^2, \quad \text{cov}[x, y] = \overline{xy} - \bar{x}\bar{y}$$

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

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

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

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



Summary: Line of Best Fit for:

$$y = mx$$

N data point pairs (x, y)

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

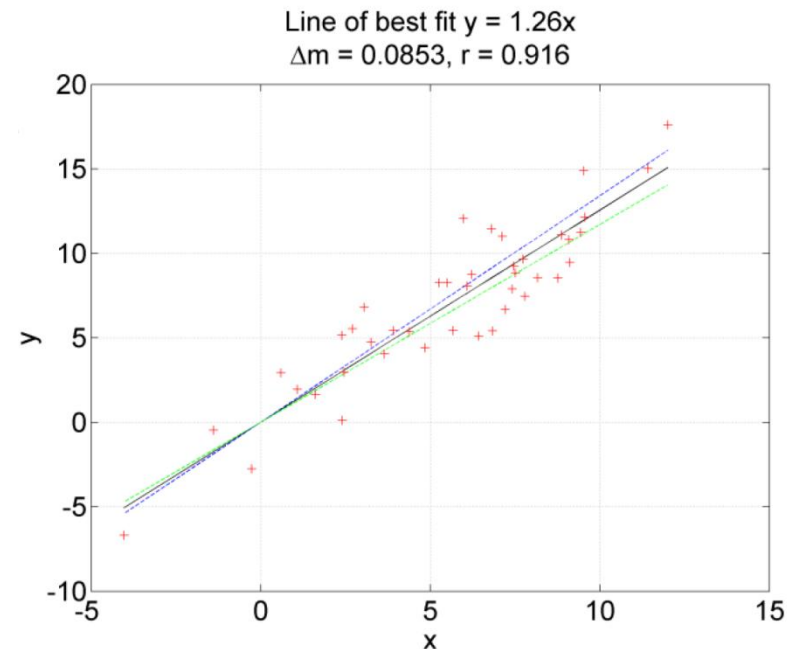
$$V[x] = \overline{x^2} - \bar{x}^2, \quad V[y] = \overline{y^2} - \bar{y}^2, \quad \text{cov}[x, y] = \overline{xy} - \bar{x}\bar{y}$$

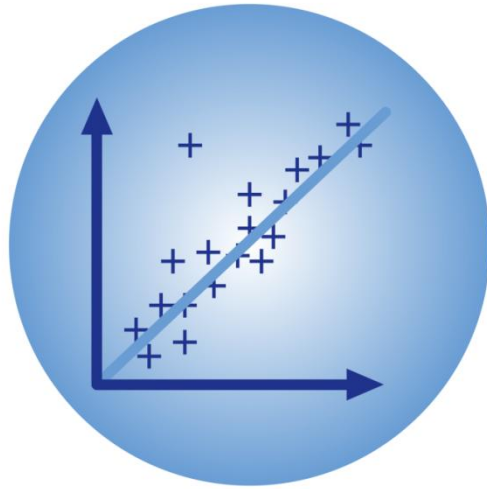
$$m = \frac{\overline{xy}}{\overline{x^2}}$$

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

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

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





BPhO

Computational Challenge

- Suggested homework
- Q&A