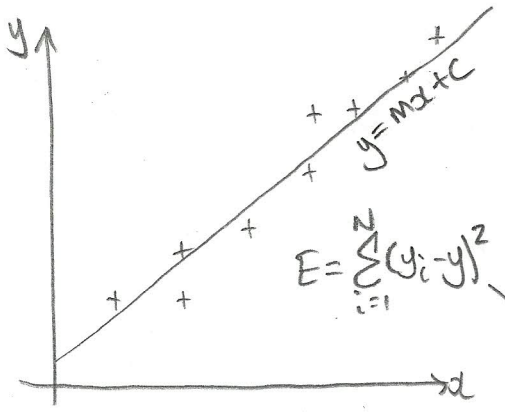
lines of best fit to (x,y) scattered data + Correlation

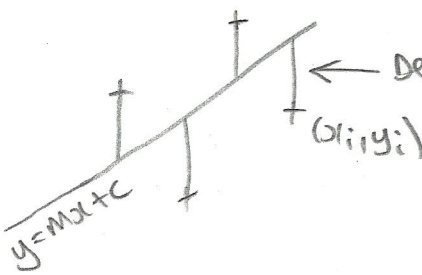


Consider a set of N pairs of x,y data
$\{x_i, y_i\}$; $i = 1 \ldots N$

what is the 'best' line $y = mx + c$ which can be associated with the data?

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

is the sum of the squares of deviations of the line of best fit from the actual y values.



To find m and c, consider a surface of $E(m,c)$

{for the best fit} → m and c are the coordinates of the minimum of the surface.

is when the overall sum of the square deviations is minimized.

Hence to find m and c we must consider solutions to

$$\frac{\partial E}{\partial m} = 0 \quad \text{and} \quad \frac{\partial E}{\partial c} = 0$$

{Partial derivatives with respect to the variables m and c correspondingly. All other parameters are held constant}

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

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

$\therefore$ when $\frac{\partial E}{\partial m} = 0 \Rightarrow$

$$m\sum_{i=1}^{N} x_i^2 + c\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i y_i = 0 \quad (2)$$

$$\frac{\partial E}{\partial c} = \sum_{i=1}^{N} (2)(-1)(y_i - mx_i - c) \quad \therefore \text{ when } \frac{\partial E}{\partial c} = 0$$

$$\sum_{i=1}^{N} y_i - m\sum_{i=1}^{N} x_i - c\sum_{i=1}^{N} = 0 \quad (3)$$

Equations ② and ③ can be solved simultaneously to yield (m, c) which minimise E^*

(* Strictly speaking we should check that $\frac{\partial^2 E}{\partial c^2} > 0$ and $\frac{\partial^2 E}{\partial m^2} > 0$ for this to be a minima) → we will check this later!

Firstly let us re-write ② and ③ and define some statistical quantities

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

$$\overline{x^2} = \frac{1}{N} \sum_{i=1}^{N} x_i^2$$

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad \text{[Mean of } \{x_i\}\text{]}$$

$$\overline{xy} = \frac{1}{N} \sum_{i=1}^{N} x_i y_i$$

So ②: $$\boxed{m\overline{x^2} + c\bar{x} - \overline{xy} = 0}$$

Now ③: $$\frac{1}{N} \sum_{i=1}^{N} y_i - m \frac{1}{N} \sum_{i=1}^{N} x_i - c \frac{1}{N} \sum_{i=1}^{N} = 0$$

$\sum_{i=1}^{N} = N$

Define $\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$

Hence ③: $$\boxed{\bar{y} - m\bar{x} - c = 0}$$

② + $\bar{x}$③: $$m\overline{x^2} - \overline{xy} + \bar{y}\bar{x} - m\bar{x}^2 = 0$$

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

$$\therefore \quad \boxed{m = \frac{\overline{xy} - \bar{y}\bar{x}}{\overline{x^2} - \bar{x}^2}}$$

m is the ratio of two statistical quantities

Covariance of $\{x, y\}$ $\quad \boxed{\mathrm{Cov}[x,y] = \overline{xy} - \bar{x}\bar{y}}$

Variance of $\{x\}$ $\quad \boxed{V[x] = \overline{x^2} - \bar{x}^2}$

Hence $\quad \boxed{m = \dfrac{\mathrm{Cov}[x,y]}{V[x]}}$

Note:
$$\mathrm{Cov}[x,y] = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

Now from ③: $\quad c = \bar{y} - m\bar{x}$

∴ in summary, the line of best fit formed by minimizing the sum of square "y" deviations (E) is given by → of data $\{x_i, y_i\}$

$$y = mx + c$$

$$m = \frac{\mathrm{Cov}[x,y]}{V[x]}$$

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

$$\mathrm{Cov}[x,y] = \overline{xy} - \bar{x}\bar{y}$$

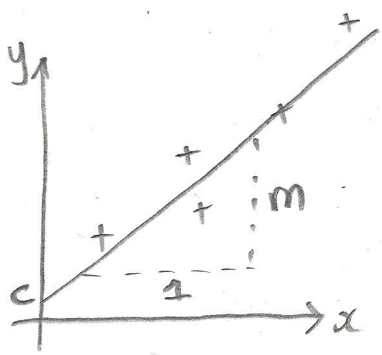
$$V[x] = \overline{x^2} - \bar{x}^2$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\overline{x^2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2$$

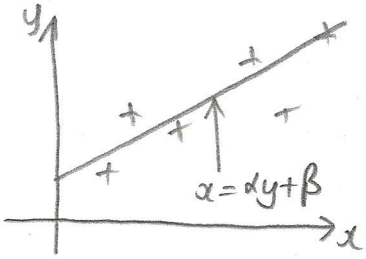
$$\overline{xy} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$



of course the entire exercise could be achieved by considering the minimum of a surface $F(\alpha,\beta)$ formed from the sum of square of deviations from line of best fit

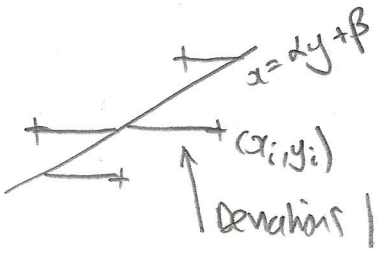
$$x = \alpha y + \beta \quad \left[\Rightarrow y = \frac{x}{\alpha} - \frac{\beta}{\alpha}\right]$$



In this case;

$$F = \sum_{i=1}^{N} (x_i - x)^2$$

$$= \sum_{i=1}^{N} (x_i - \alpha y_i - \beta)^2$$



The derivation of α, β is clearly the same as above, but with $x \Leftrightarrow y$ interchanged

ie

$$\boxed{\alpha = \frac{\overline{xy} - \bar{y}\bar{x}}{\overline{y^2} - \bar{y}^2}} = \frac{Cov[x,y]}{V[y]}$$

[obviously since $Cov[x,y] = \overline{xy} - \bar{y}\bar{x}$, $Cov[x,y] = Cov[y,x]$)

$$\boxed{\beta = \bar{x} - \alpha\bar{y}}$$

[So for horizontal deviations
$y = mx + c$
$m = V[y]/Cov[x,y]$ $\quad c = \bar{y} - m\bar{x}$]

Now, if $\{x_i, y_i\}$ are <u>perfectly correlated</u> , $x = \alpha y + \beta$ and $y = mx + c$ are the same line

ie

$$y = mx + c$$
$$y = \frac{x - \beta}{\alpha} \quad \Rightarrow \quad m = \frac{1}{\alpha} \quad \text{and} \quad c = -\frac{\beta}{\alpha}$$

Consider $m = \frac{1}{\alpha} \Rightarrow \boxed{m\alpha = 1}$

ie lines have the same gradient for vertical and horizontal deviations.

Hence if we define $\rho^2 = m\alpha$, the closeness of ρ to unity will tells use something useful about the degree of correlation of the data.

$$\rho^2 = \frac{Cov[x,y]}{V[x]} \times \frac{Cov[x,y]}{V[y]}$$

$$\therefore \quad \boxed{\rho = \frac{Cov[x,y]}{\sigma_x \sigma_y}}$$

This is called the product moment correlation coefficient.

where Standard deviations

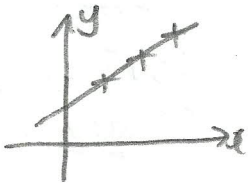
$$\sigma_x = \sqrt{V[x]}$$

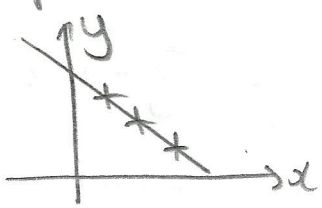
$$\sigma_y = \sqrt{V[y]}$$

Note ρ could also be -ve since $Cov[x,y]$ need not be +ve.

If $\rho = +1$ the data will be +ve correlated

If $\rho = -1$ " " " " -ve "





Extra

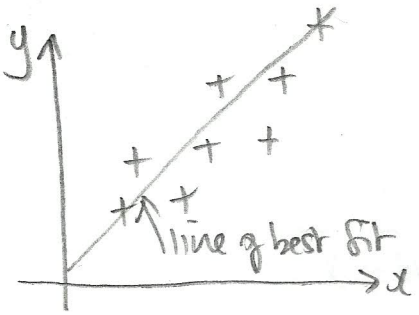
$$\frac{\partial^2 E}{\partial m^2} = 2\sum_{i=1}^{N}(x_i^2) = 2N\overline{x^2}$$

This must be >0 so $\frac{\partial E}{\partial m} = 0$ will be a minima

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

$$= 2\sum_{i=1}^{N} = 2N.$$ This is >0 so $0 = \frac{\partial E}{\partial c}$ will also be a minimum.

Overall linear regression and correlation summary



$\{x_i, y_i\}$ data samples. There are N pairs.

Define

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\overline{x^2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \qquad \overline{y^2} = \frac{1}{N}\sum_{i=1}^{N} y_i^2$$

$$\overline{xy} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i$$

Do this in a table →

x	y	x^2	y^2	xy
⋮	⋮	⋮	⋮	⋮

Then compute:

$$V[x] = \overline{x^2} - \bar{x}^2$$

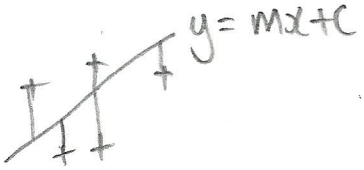
$$V[y] = \overline{y^2} - \bar{y}^2$$

$$\mathrm{Cov}[x,y] = \overline{xy} - \bar{x}\bar{y}$$
(Covariance)

Product-moment correlation coefficient

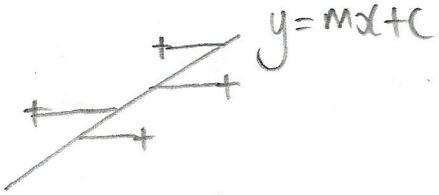
$$p = \frac{\mathrm{Cov}[x,y]}{\sqrt{V[x]\,V[y]}}$$

lines of best fit are then:



Vertical fit

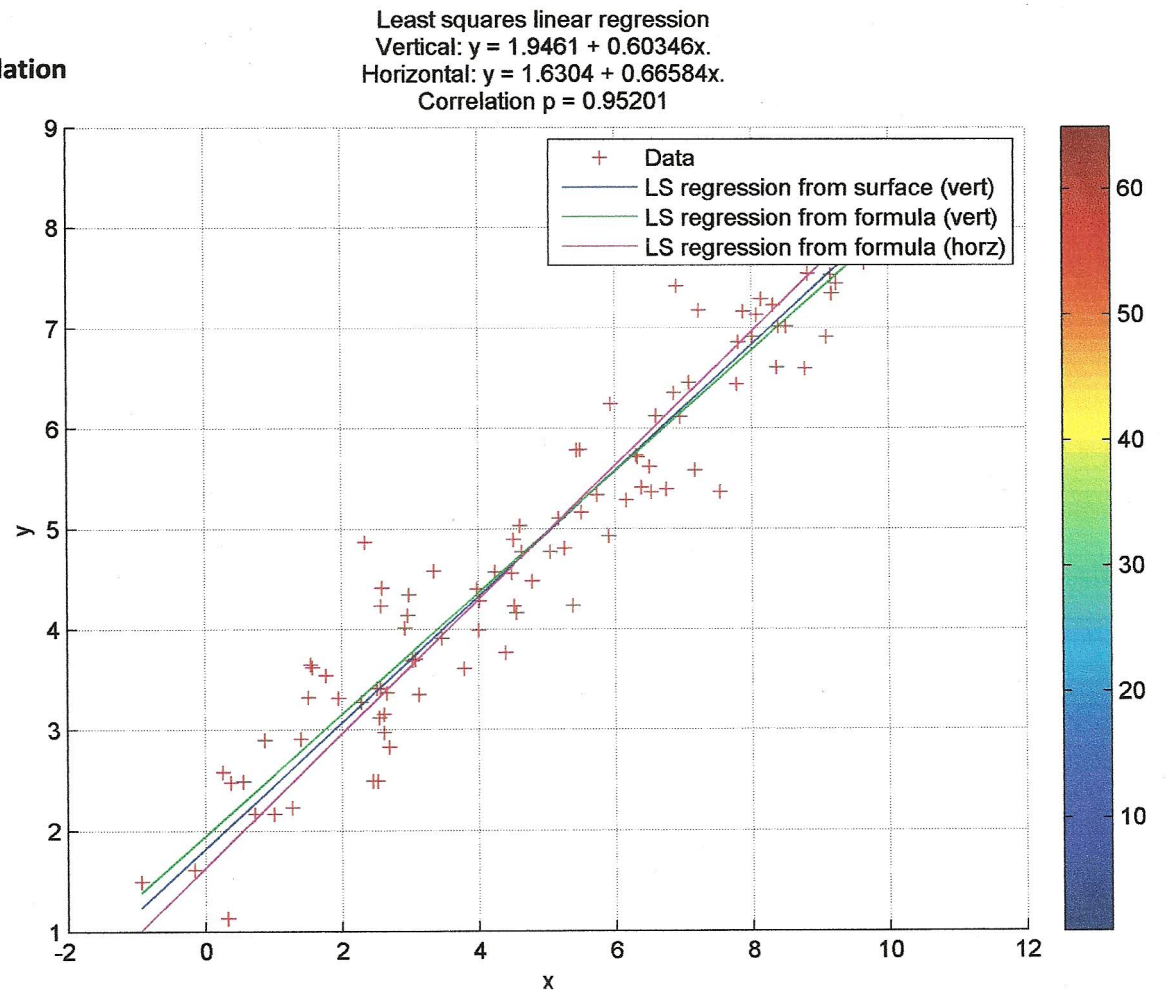
$$m = \frac{\mathrm{Cov}[x,y]}{V[x]} \qquad c = \bar{y} - m\bar{x}$$

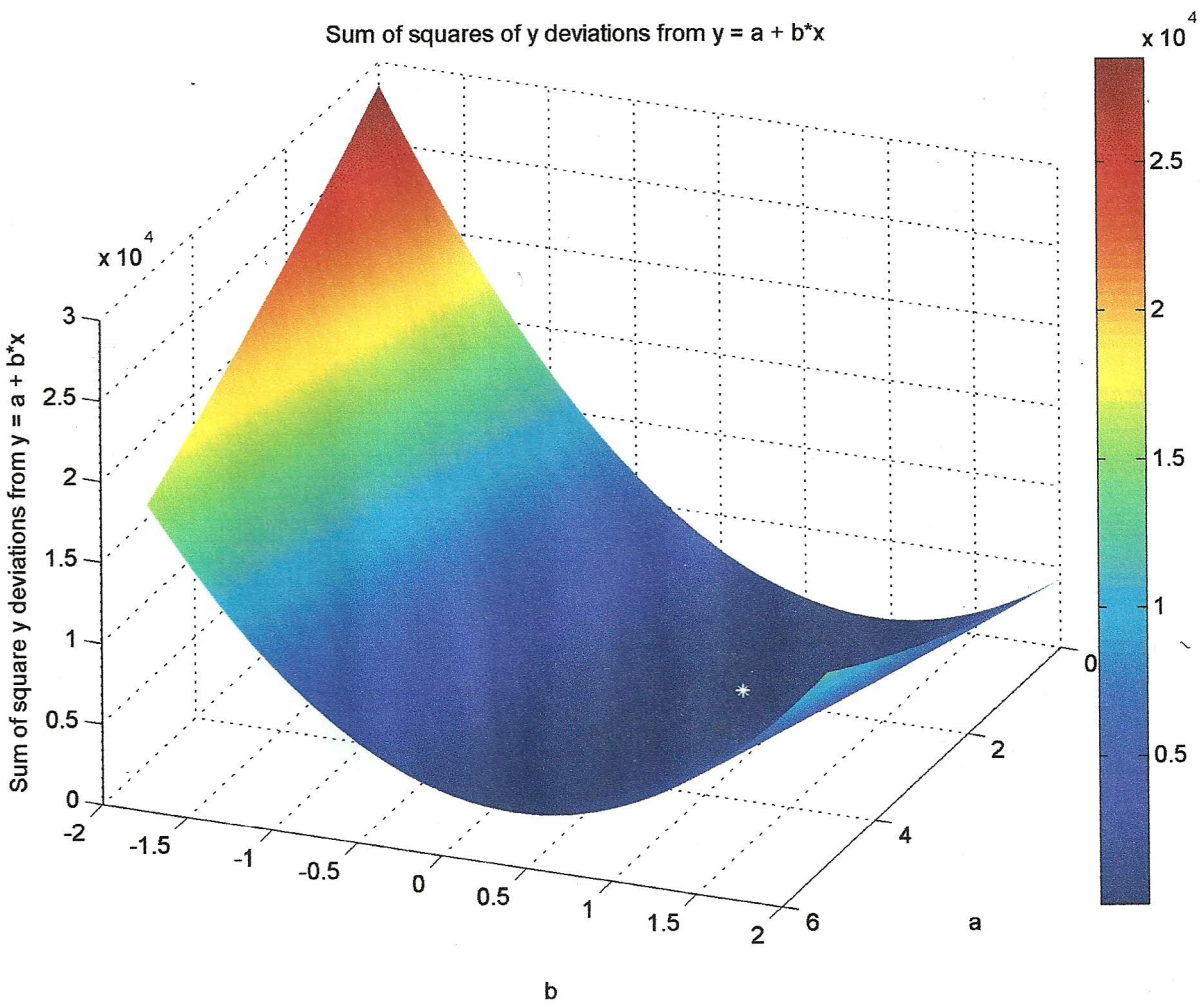


Horizontal fit

$$m = \frac{V[y]}{\mathrm{Cov}[x,y]} \qquad c = \bar{y} - m\bar{x}$$

+ve correlation





-ve correlation