AMS 131
18 Jul 18

①

this prediction;
time: covariance
next & correlation;
time: utility

$E[h(X)] \neq h[E(X)]$ except when $h(x) = ax + b$

(~1895) Karl Pearson

Francis Galton

human genetics

Mendel (~1860) peas

Darwin (~1860) general

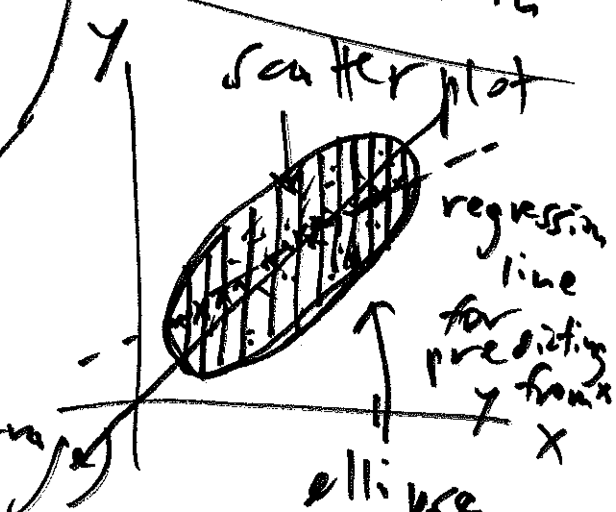
	y	x
	69.3	68.2 in
	⋮	⋮
i	y_i	x_i
		⋮
mean	~~67 in~~ 66	66 in
SD	3 in	3 in

$n = 1000$

height of father (x);
height of adult son (y)

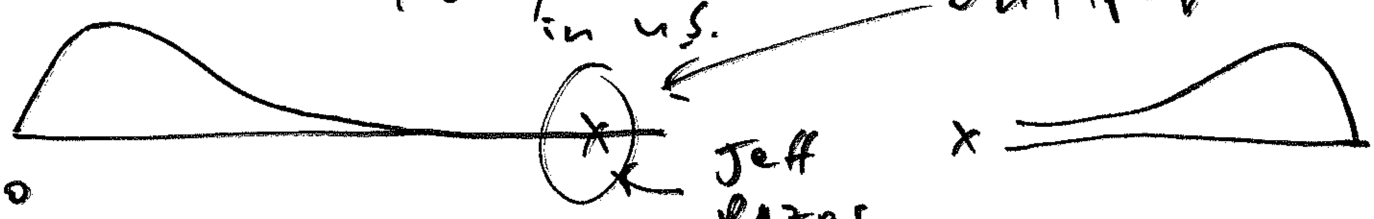
~1 in

scatter plot

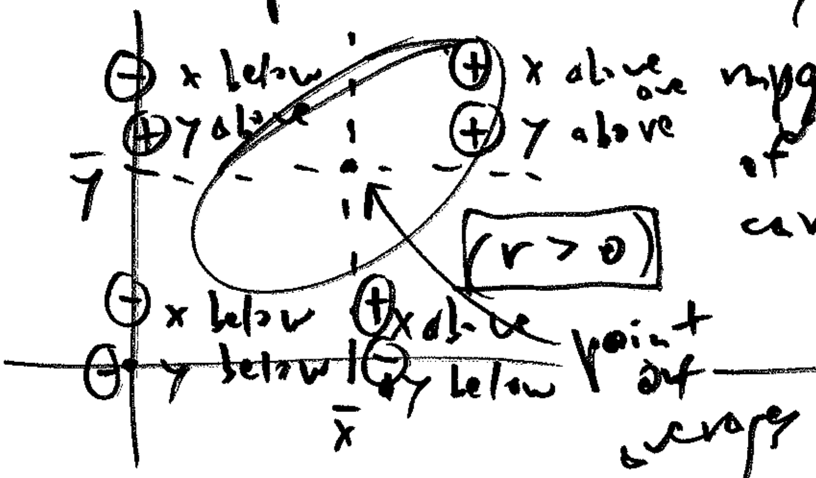


x, y are associated
(as $x \uparrow$, y tends to $\uparrow$ on average)

correlation (coefficient) r (4.53)

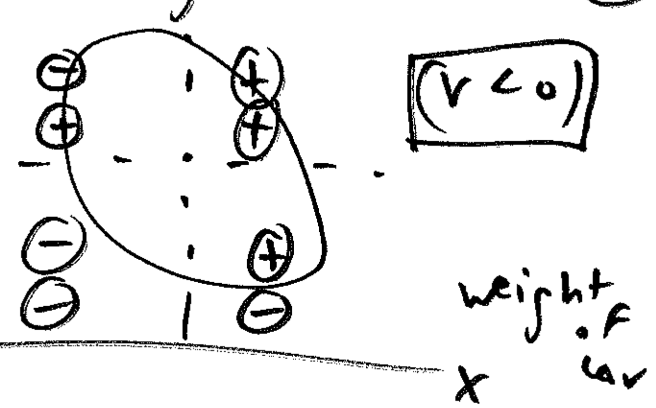


positive assoc.



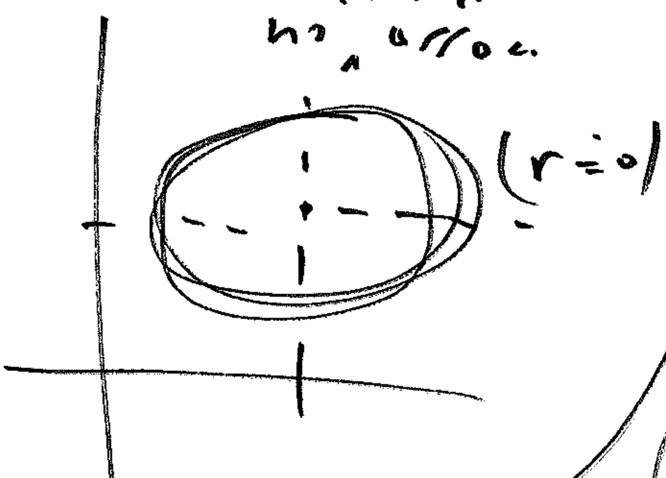
avg of cov

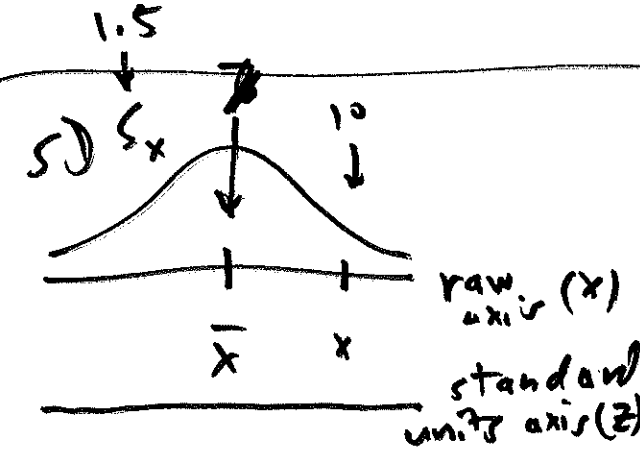
negative assoc ②



weight of cov

(linear) no assoc.





(how far is x from $\bar{x}$, in units of SD s_x?)

$$z = \frac{x - \bar{x}}{s_x}$$

converting to standard units

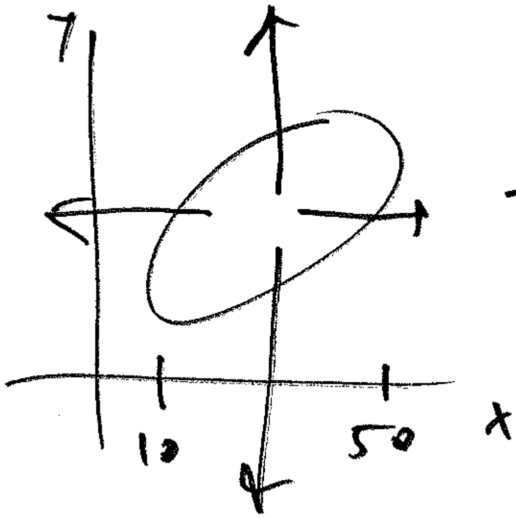
$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x^*}\right)\cdot\left(\frac{y_i - \bar{y}}{s_y^*}\right) = r$$

$$s_x^* = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

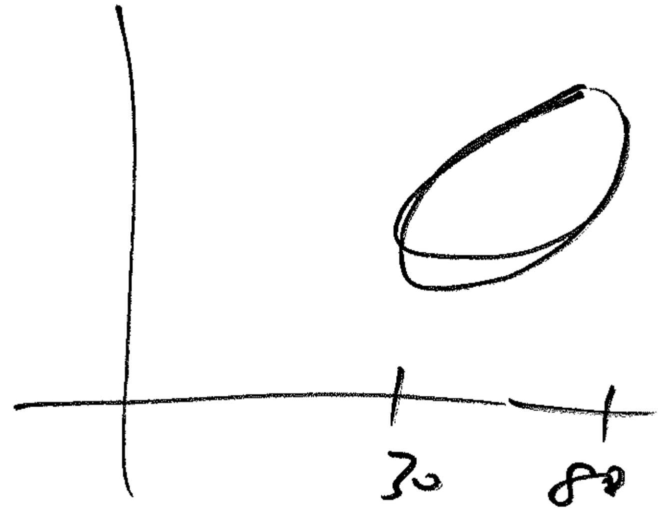
data

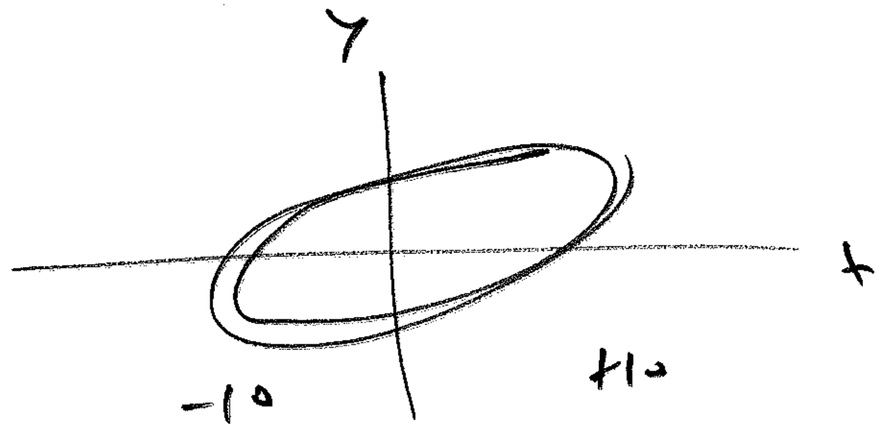
$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$r = \frac{1}{n}\sum_{i=1}^{n}\frac{\boxed{x_i \cdot y_i} - \bar{x} y_i - \bar{y} x_i + \bar{x}\bar{y}}{s_x^* \cdot s_y^*}$$

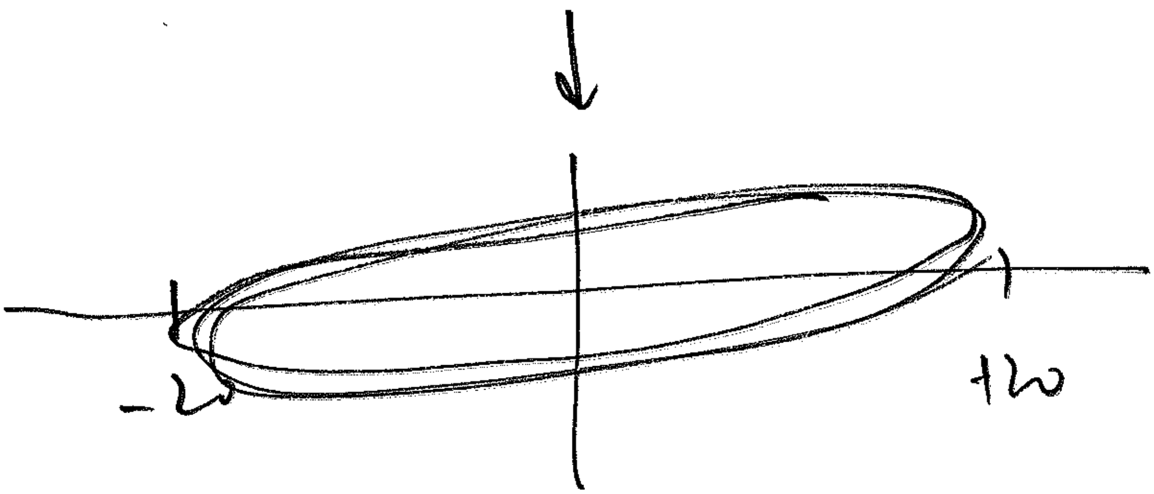


add 20 to x



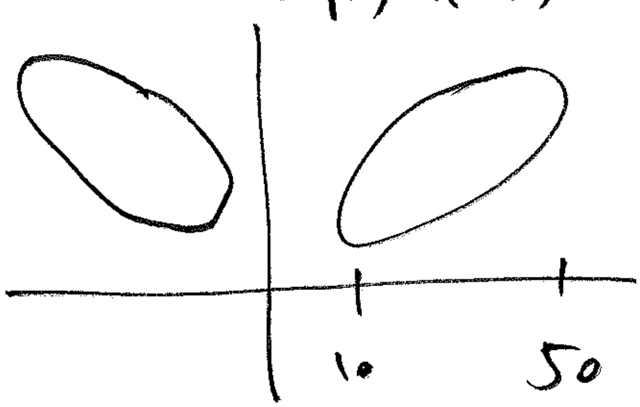


multiply x by 2

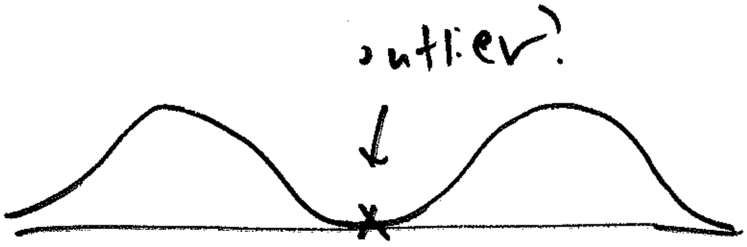


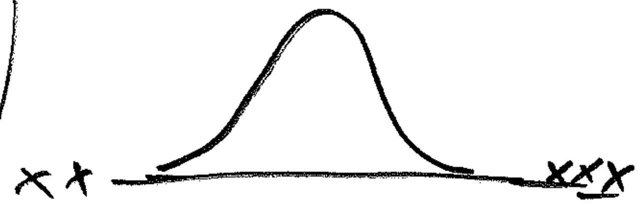
multiply x by -1

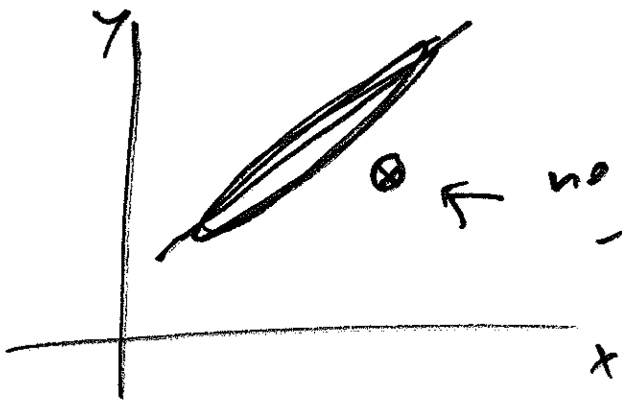
new
r = -0.6



old
r = + 0.6

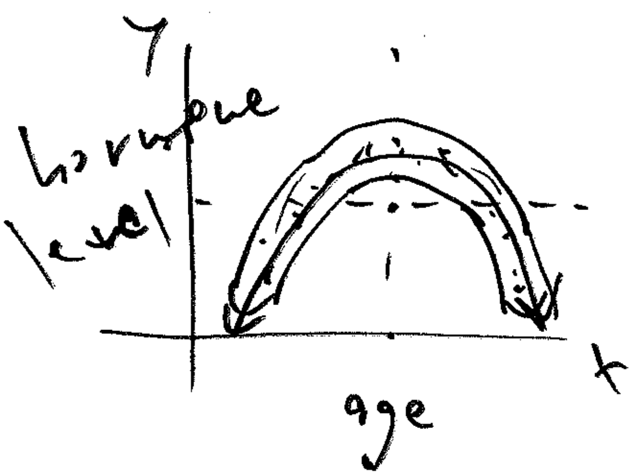






← not outlier in x
y

is outlier in (x, y)



x, y strongly
associated
and yet $r \doteq 0$