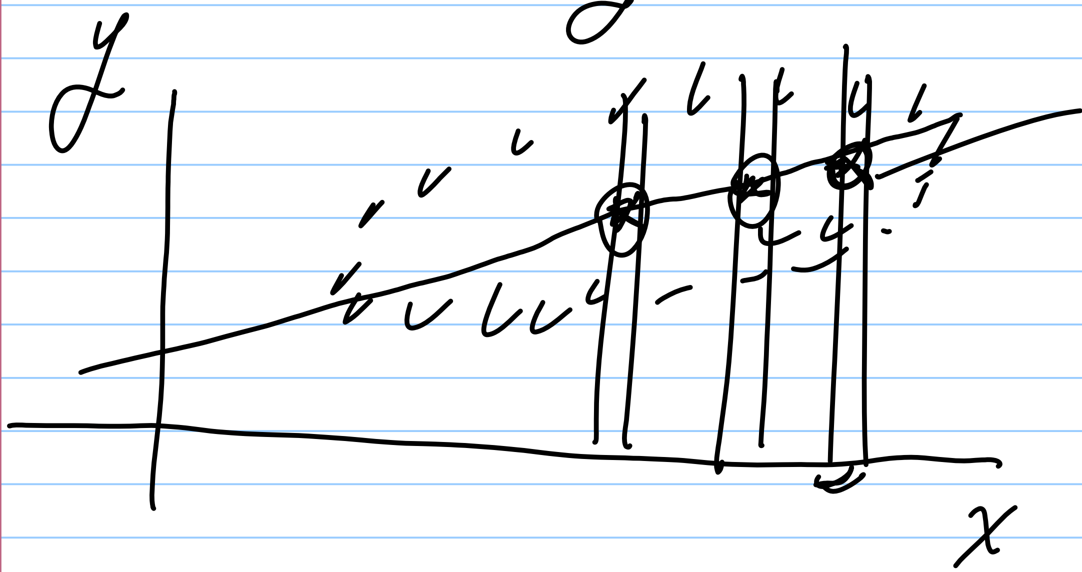
STT 200 3pm 4-5-10

Ch 7 (PARTS OF CH 8, 9)

1. SCATTERPLOT
2. COVARIANCE
3. STRAIGHT LINE ASSOC.
4. CORRELATION (ORIG. CO-RELATION)
5. 2-Dim NORMAL PLOTS
6. REGRESSION LINE (Y on X vs X on Y)
7. LEAST SQUARES LINE - EVEN WHEN PLOTS ARE NOT NORMAL.

SCATTERPLOT eg x = HEIGHT (IN) OF A FATHER

y = " " " SON OF THAT FATHER



AVG HT. OF SONS BORN TO FATHER IN WINDOW

GALTON'S DATA

A. GALTON SAW THAT PLOT OF VERTICAL STRIP AVGS WAS $\approx$ ST LINE. !!

B. REGRESSION EFFECT THE AVG HT OF SONS

BORN TO FATHERS OF HEIGHT $\bar{x} + \# \sigma_x$

IS CLOSER TO $\bar{y}$ THAN $\bar{y} + \# \sigma_y$

MORE PRECISELY, FOR ELLIPTICAL SHAPED PLOTS

AVG HT OF SONS BORN TO FATHERS

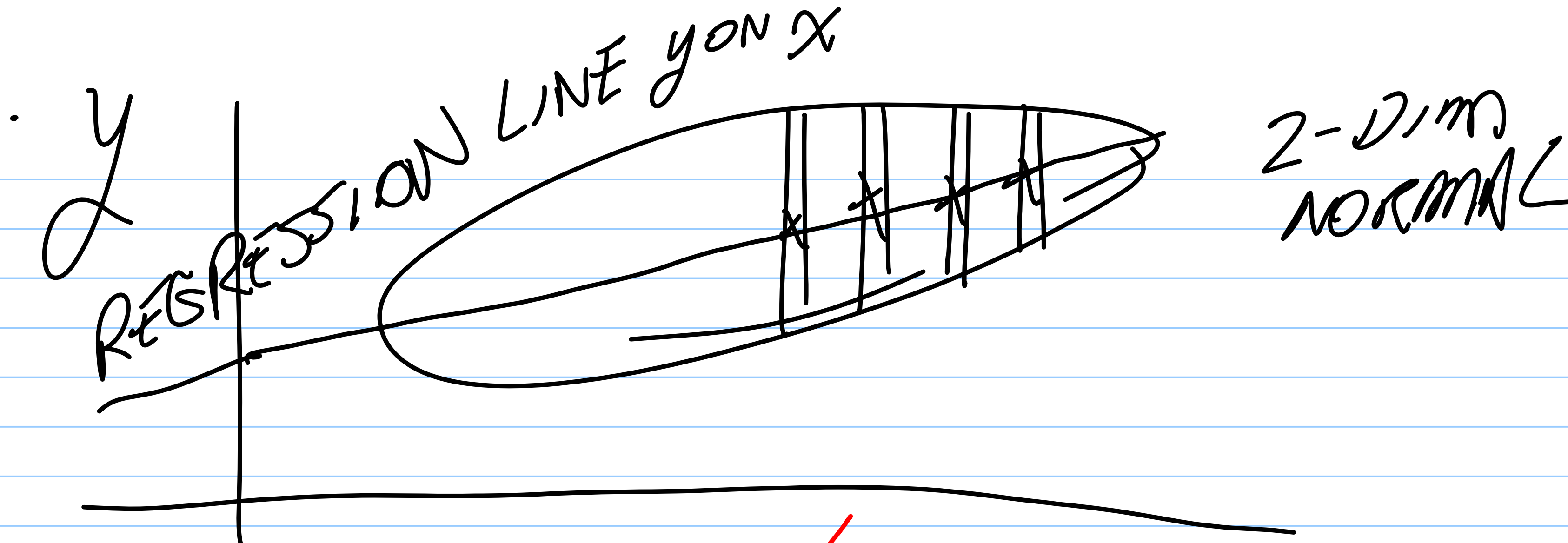
OF HT $\bar{x} + \# \sigma_x$ IS $\approx \bar{y} + \# \, r \, \sigma_y$

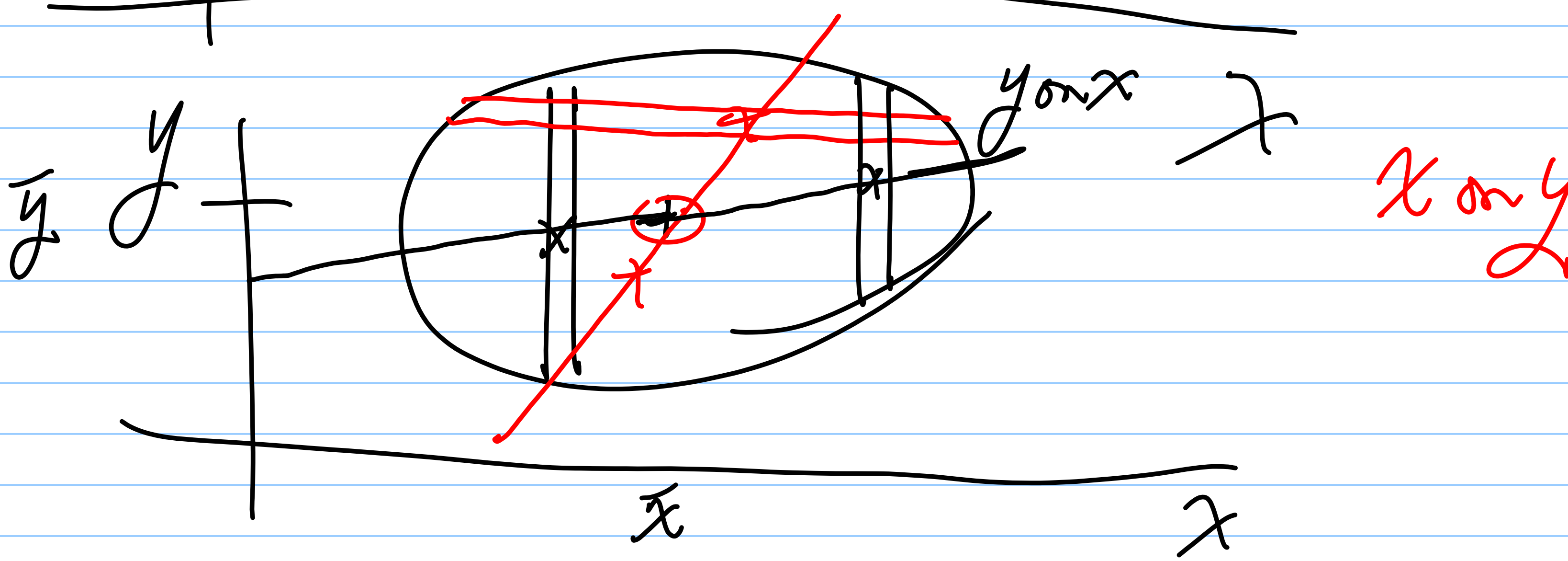
$|r| < 1$ CORRELATION

REGRESSION TO MEDIOCRITY (TO MEAN)

"NOTION OF CO-RELATION" GAVE WAY

TO TECHNICAL "CORRELATION".

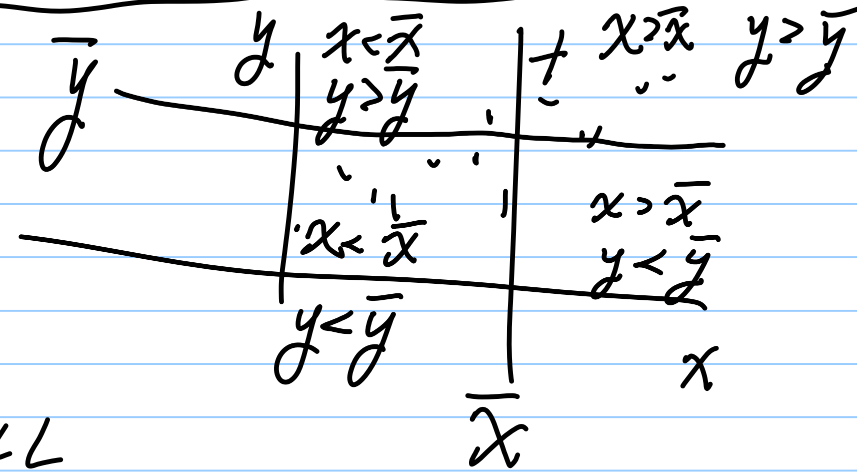

REGRESSION LINE y on x
y
2-DIM NORMAL


$\bar{y}$
y
y on x
x
x on y
$\bar{x}$
x

TAKE AWAY: REGR LINE x ON y IS NOT THE SAME AS REGR LINE y ON x — REGRESSION IS NOT SYMMETRI IN x, y.

WHAT IS CORRELATION?

SCORE A POINT (x, y) BY $(x-\bar{x})(y-\bar{y})$



COULD AVG THESE OVER ALL POINTS (x, y) $\Rightarrow$ $\overline{(x-\bar{x})(y-\bar{y})}$ COVARIANCE

HAD TO GET RID OF SCALE

TRY $\overline{(x-\bar{x})(y-\bar{y})} \Big/ \sqrt{\overline{(x-\bar{x})^2}} \sqrt{\overline{(y-\bar{y})^2}}$

CALLED CORRELATION r.

$$r = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y}) / n}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 / n} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 / n}}$$

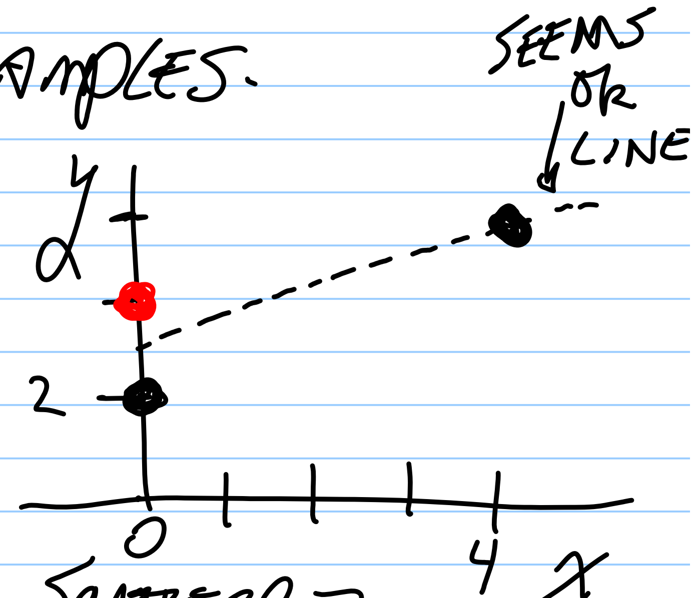
$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\,\sqrt{\overline{y^2} - \bar{y}^2}}$$

NUMERICALLY SUBJECT TO ROUNDING ERRORS.

USE TO EASE CALC IN SIMPLE EXAMPLES.

EXAMPLE (CALC CORRELATION r)

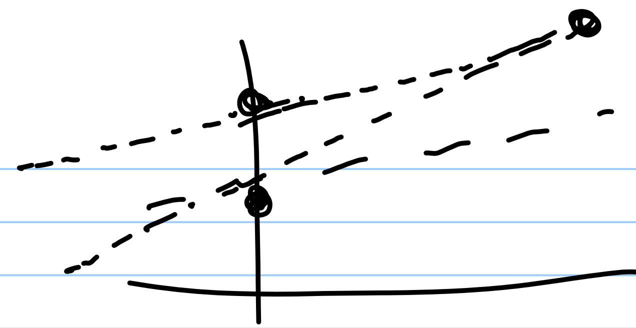
	x	y	x^2	y^2	xy
●	0	2	0	2^2	0
●	0	4	0	4^2	0
●	4	6	16	6^2	24



SCATTERPLOT

TOT 4 12 16 56 24

AVG 4/3 12/3 16/3 56/3 24/3



$\bar{x} = 4/3$ $\quad \bar{y} = 12/3$

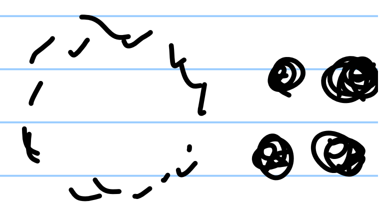
$\overline{x^2} = 16/3$ $\quad \overline{y^2} = 56/3$

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\sqrt{\overline{y^2} - \bar{y}^2}}$$

$$= \frac{24/3 - \frac{4}{3}\frac{12}{3}}{\sqrt{16/3 - (4/3)^2}\sqrt{56/3 - (12/3)^2}} =$$

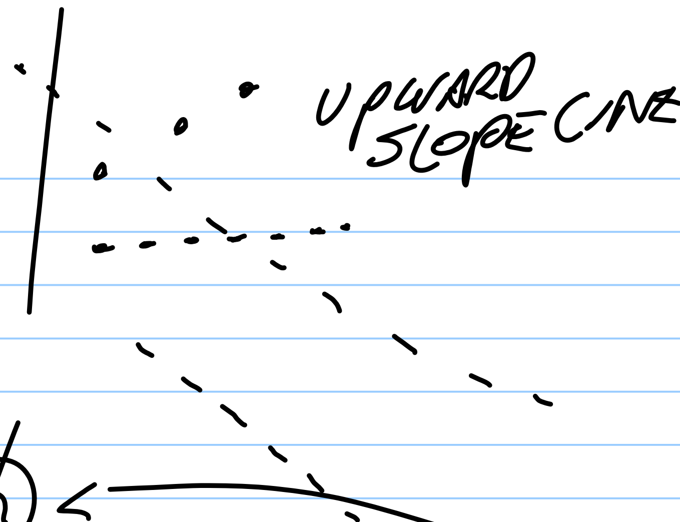
$\overline{x^2}$ $\quad \bar{x}^2$

$r = 0$



MATHEMATICALLY $|r| \le 1$ $\quad r = .866025$

$r = 1$ if & only if



$r = -1$ " " "

$r(x,y)$

$$r(3x+2,\ 7y-9) = \frac{\sum (3x+2 - \overline{3x+2})(7y-9 - \overline{7y-9})}{\sqrt{\sum (3x+2 - \overline{3x+2})^2}\ \sqrt{\ }}$$

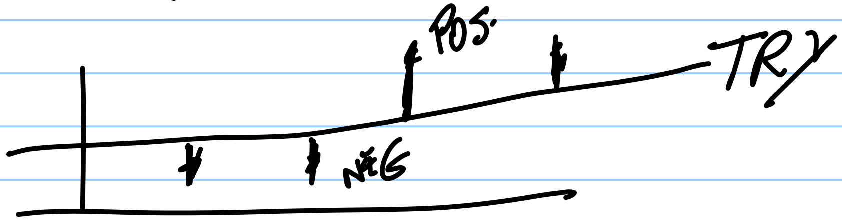
"FIX" ✓

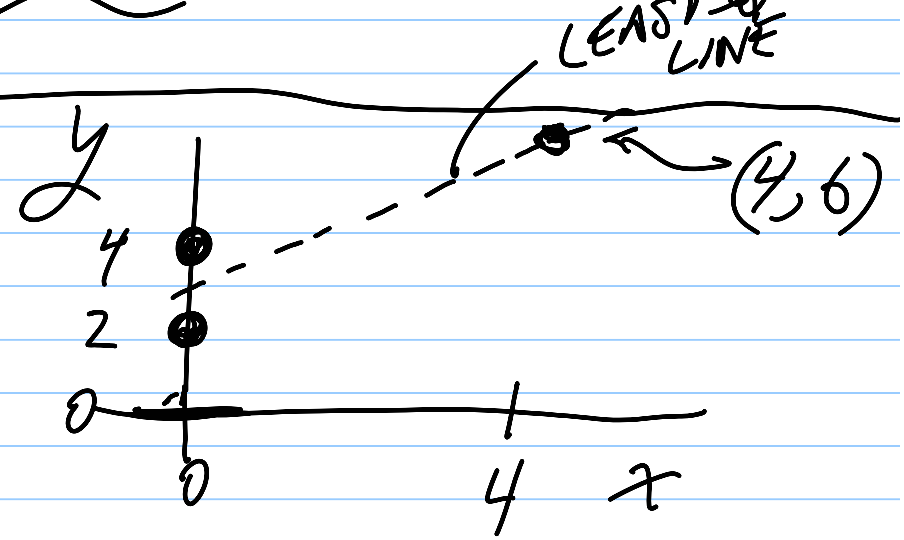
ANS $r(ax+b,\ cy+d) = \begin{cases} r(x,y) & \text{if } ab>0 \\ -r(x,y) & \text{if } ab<0 \end{cases}$

WE'LL SEE THAT r^2 IS A BETTER MEASURE OF STRENGTH OF CORRELATION THAN IS r.

BACK TO PLOT ABOVE

GAUSS LEAST SQUARES LINE

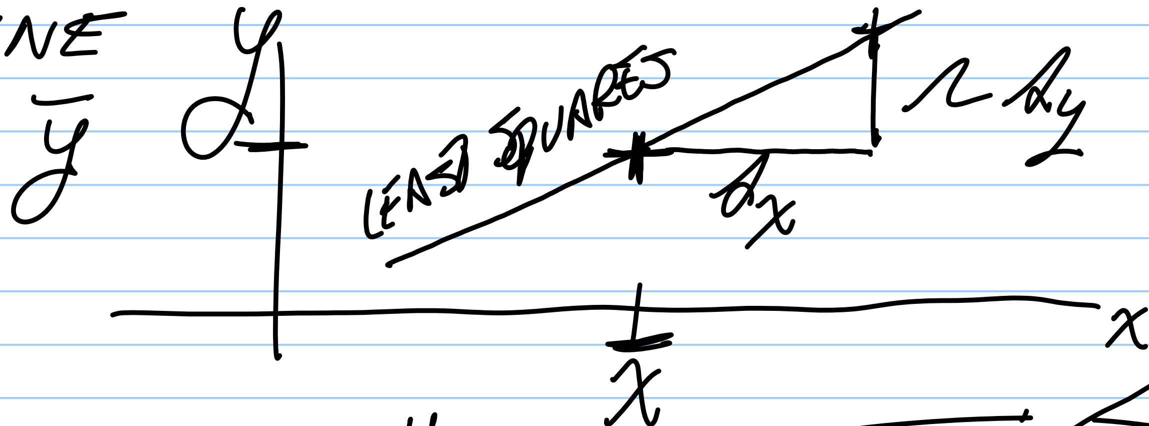




CHOOSE LINE THAT MAKES SMALLEST THE SUM OF SQUARES OF VERTICAL DEVIATION FROM LINE.

FOR ANY PLOT - THE LEAST SQUARES LINE

x = HT FATHER
y = HT SON



FOR NORMAL PLOT