STT 200 5:30 4-5-10

Note Title 4/5/2010

Ch 7 + Portions of 8 + 9

Scatterplot of $(x, y)$ pairs, eg $x$ = height of father, $y$ = height of son

$\bar{x}, s_x, \bar{y}, s_y, r$ (correlation)

$\approx 1890$ Galton

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{\overline{x^2}-\bar{x}^2}\sqrt{\overline{y^2}-\bar{y}^2}} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i-\bar{x}}{s_x}\cdot\frac{y_i-\bar{y}}{s_y} = \frac{1}{n-1}\sum z_{x_i} z_{y_i}$$


$z_{x_i}<0,\ z_{y_i}>0$

$z_{x_i}>0\ \ z_{y_i}>0$

$z_{x_i}<0,\ z_{y_i}<0$

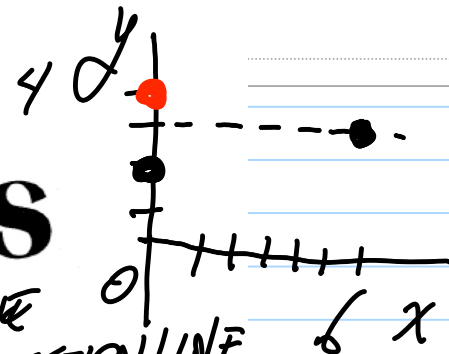
$z_{x_i}>0\ \ z_{y_i}<0$

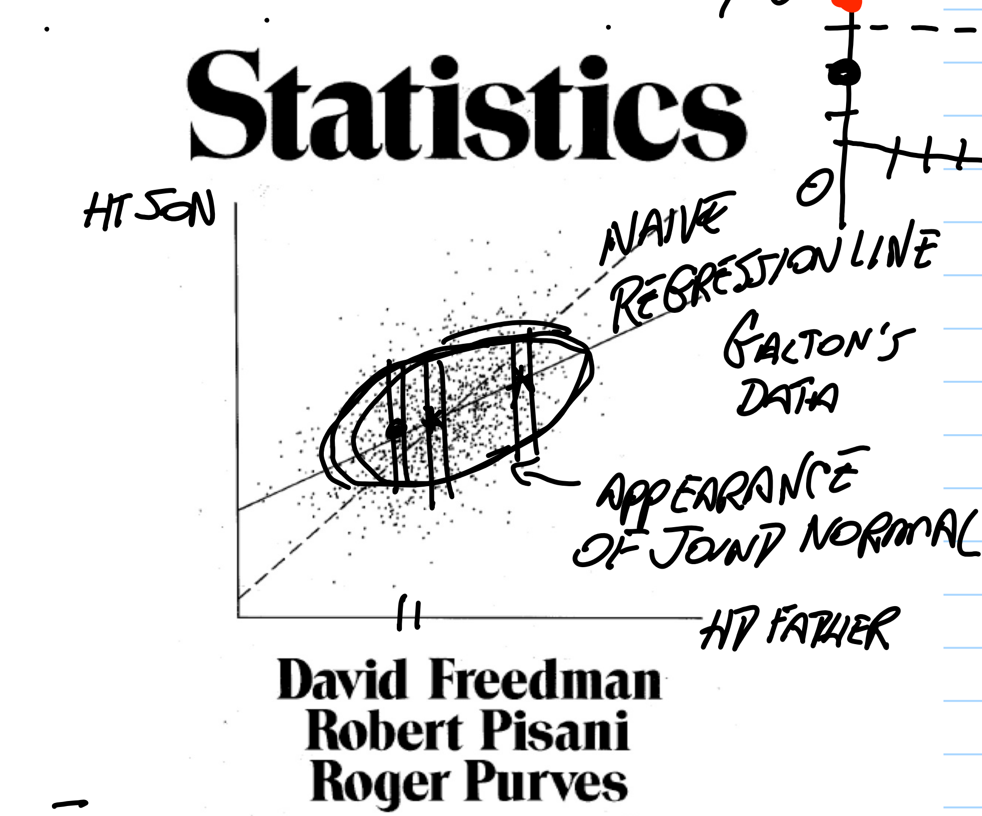
① $|r| \le 1$

② $r(ax+b, cy+d) = r(x,y)$ if $ac>0$.

NOTE: BONUS HANDOUT 4-7-10
"LARGER n"

CORRECTION → ~~40~~ 1/80




CALCULATION

| | $x$ | $y$ | $\overline{x^2}$ | $\overline{y^2}$ | $xy$ |
|---|---|---|---|---|---|
| | 0 | 2 | 0 | 4 | $0 \cdot 2 = 0$ |
| | 0 | 4 | 0 | 16 | $= 0$ |
| | 6 | 3 | 36 | 9 | $= 18$ |
| TOT | 6 | 9 | 36 | 29 | 18 |
| AVG | 2 | 3 | 12 | 29/3 | 6 |

$\bar{x} = 2 \quad \bar{y} = 3$

$\hat{\sigma}_x = \sqrt{12 - 4} = \sqrt{8}$

$\hat{\sigma}_y = \sqrt{29/3 - 3^2} = \sqrt{2/3}$

$r = \dfrac{6 - 2 \cdot 3}{\sqrt{8}\,\sqrt{2/3}} = 0$

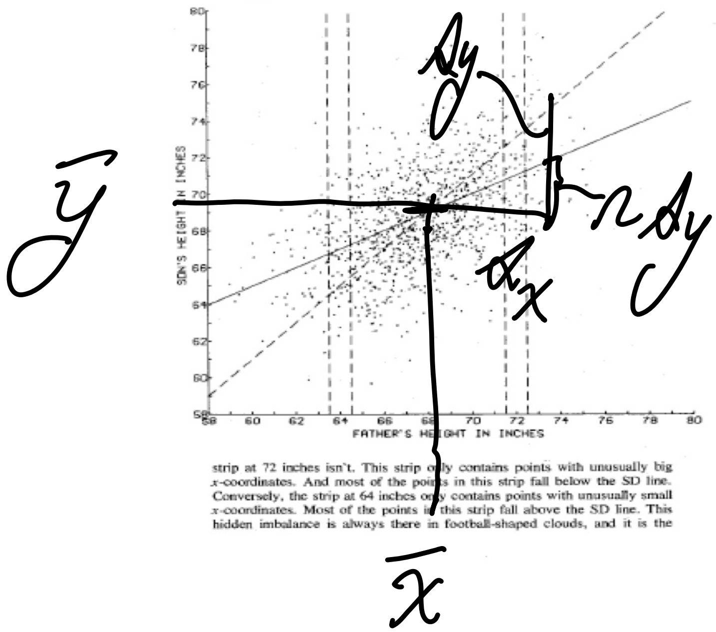
Note Title

GAUSS:

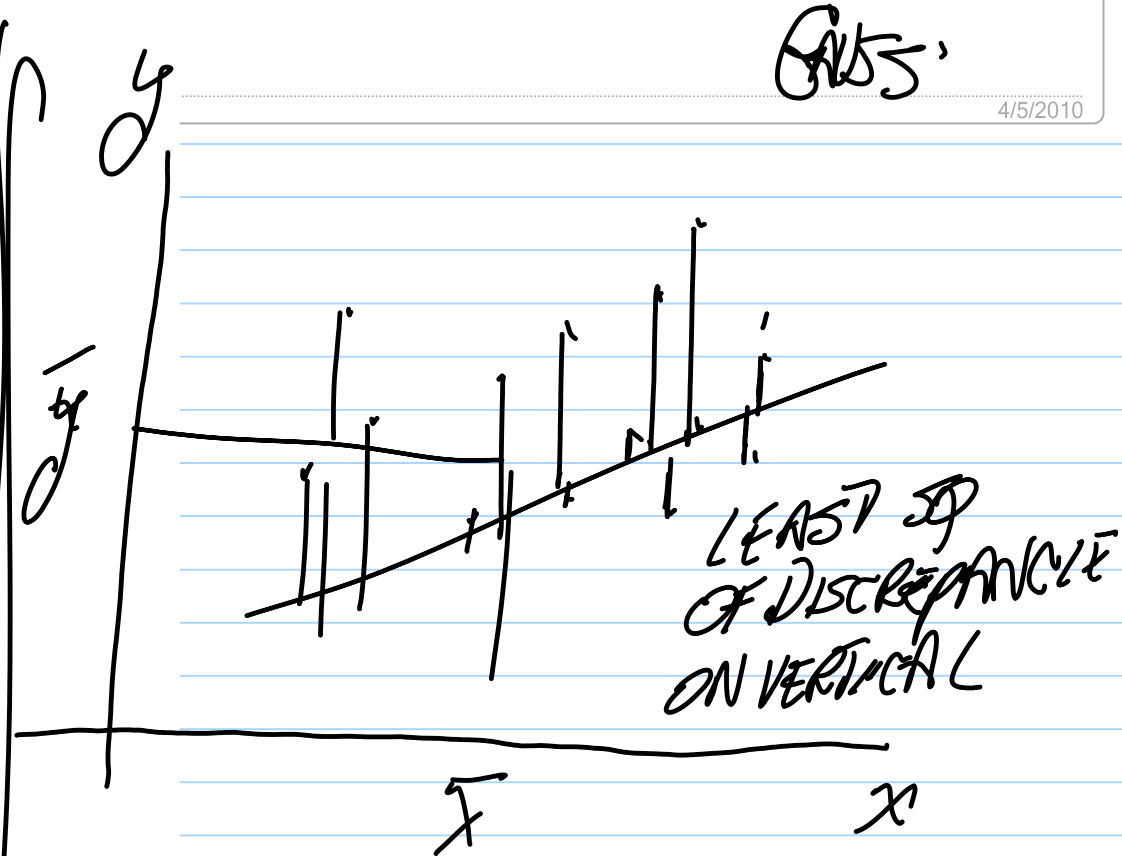
4/5/2010



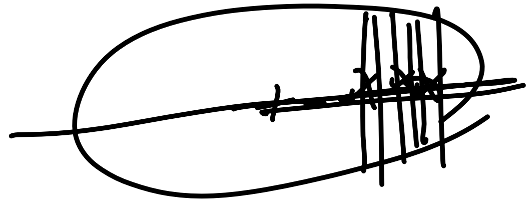
Figure 7. The regression effect. The scatter diagram shows the heights of 1,078 fathers and sons. The sons averaged 1 inch taller than the fathers. If the son was 1 inch taller than his father, this family is plotted along the dashed line. The points in the strip over 72 inches correspond to the families where the father was 72 inches tall, to the nearest inch. Most of these points are below the dashed line, and the average height of these sons is only 71 inches. The points in the strip over 64 inches correspond to families where the father was 64 inches tall, to the nearest inch. Most of these points are above the dashed line, and the average height of these sons is 67 inches. The solid regression line picks off the centers of all the vertical strips, and is much flatter than the dashed line.




strip at 72 inches isn't. This strip only contains points with unusually big $x$-coordinates. And most of the points in this strip fall below the SD line. Conversely, the strip at 64 inches only contains points with unusually small $x$-coordinates. Most of the points in this strip fall above the SD line. This hidden imbalance is always there in football-shaped clouds, and it is the



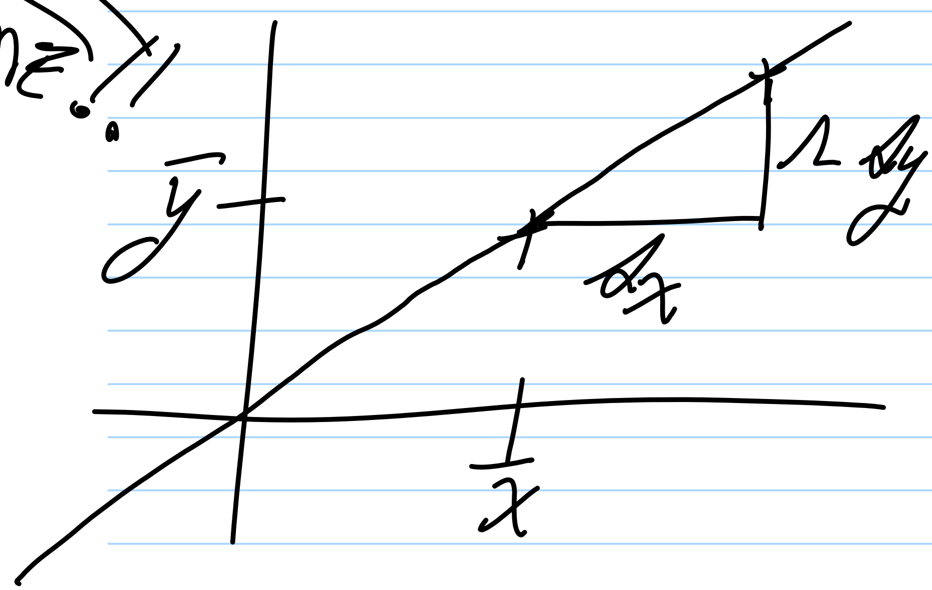
Note Title 4/5/2010



REGR

LEAST SQ LINE (GAUSS)

SAME!!



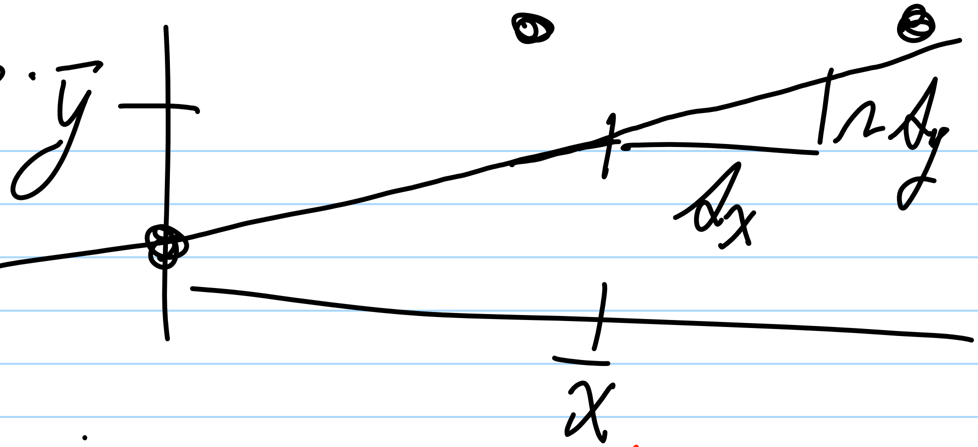
3. For the (x, y) data

| x | y | x deviations = $(x - \bar{x})$ | y deviations = $(y - \bar{y})$ | product |
|---|---|---|---|---|
| 31 | 78 | | | |
| 26 | 46 | | | |
| 28 | 69 | | | |

a. Fill out the above table to complete the table find the correlation r by hand as is s
173.

CALC r

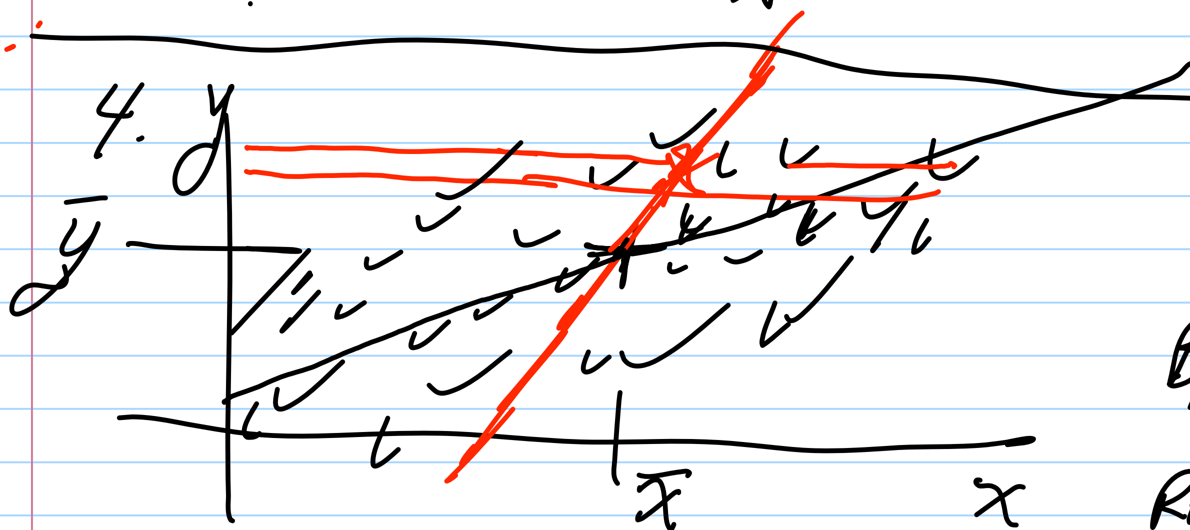
3b.




get $\bar{x}, s_x, \bar{y}, s_y, r$

$\hat{\sigma}_x$ $\hat{\sigma}_y$

4.



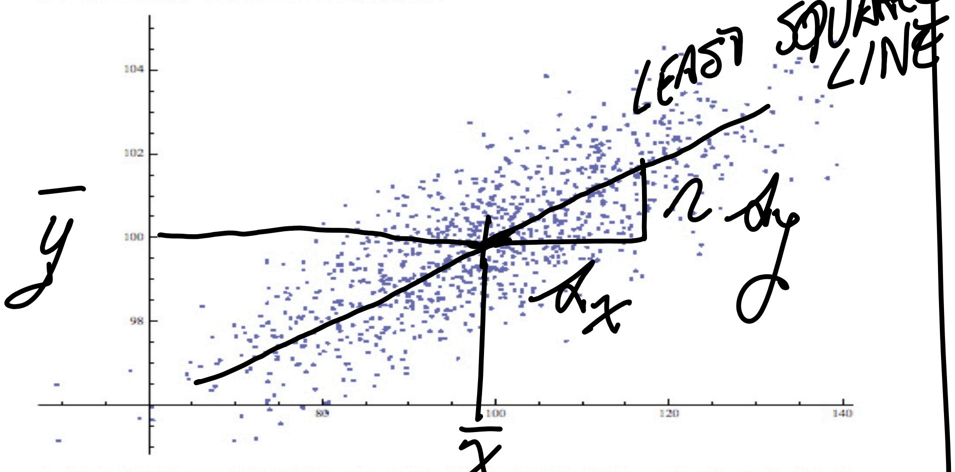

$r(x,y) = r(y,x)$ SYMMETRIC

BY EYE $\bar{x}, \bar{y}$

REGR y on x DIFFERS FROM

x on y.

4. Here is a plot of data having a **jointly normal** (i.e. 2-dimensional normal) distribu
Note the roughly elliptical form of the plot.




a. For a jointly normal plot the regression line is just the line passing through the mi points at each x. You can easily draw it by hand. Just go through the middle of the p line "regression of y on x." Your line should pass through the point of avera approximately (100, 100) for this particular data. If you were to calculate the slope would get this very line you can fit by eye for jointly normal (x, y) data.

$x$ = LAST DIGIT STUDENT #

$y$ = LAST DIGIT + SECOND LAST DIGIT

eg ~ 32

$x = 2$ $y = 5$