

May 24, 2005 Lecture Notes (VIII. Correlation, Regression)

Case Study 17

heights of fathers (f) + sons (s)

- We've talked about how to describe + do inference on 1 variable at a time; what about 2 at a time?

(son) y	(father) x
+ 70 in	72 in
- 67 in	66 in
⋮	⋮
⋮	⋮

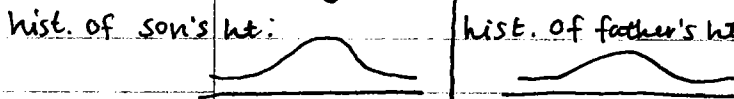
← 1 row for each family; 2 columns (for fathers + sons)  
n = 1078

height of f + s: both quantitative continuous variables

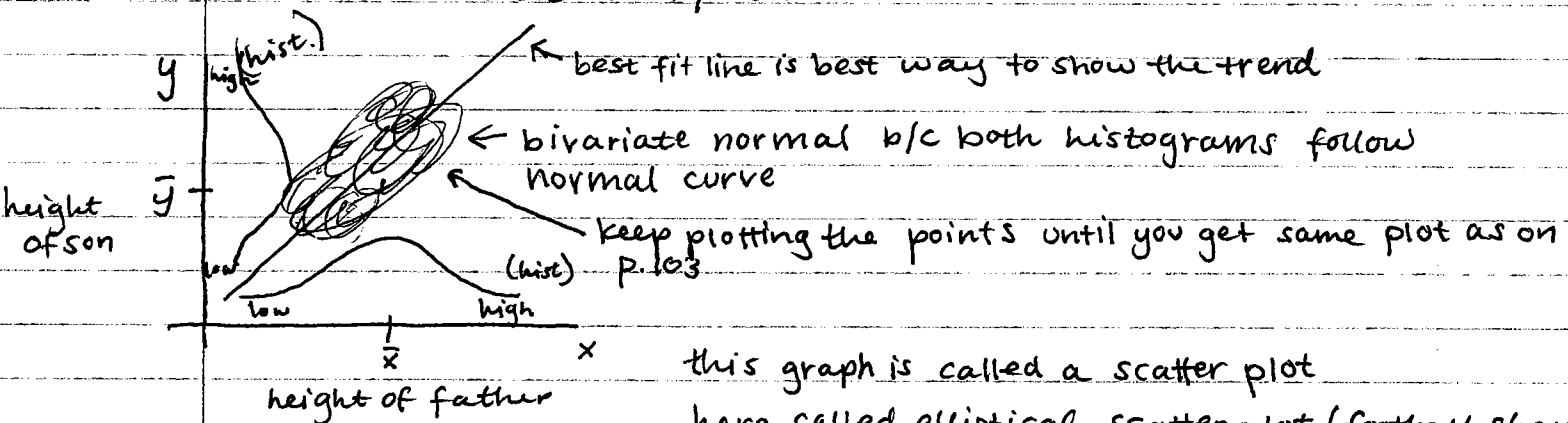
mean: mean  $\bar{y}$  = 69 in    mean  $\bar{x}$  = 68 in

SD:  $S_y$  = 2.7 in     $S_x$  = 2.7 in

Secular trend in height: sons taller than fathers by about 1 in. because of better nutrition



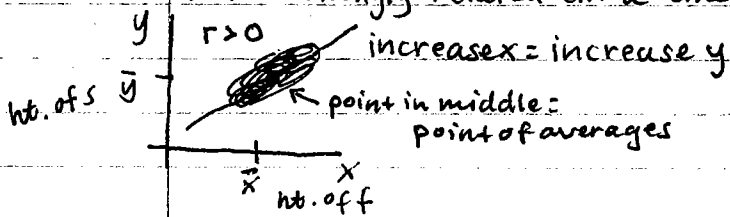
both normal curve b/c CLT



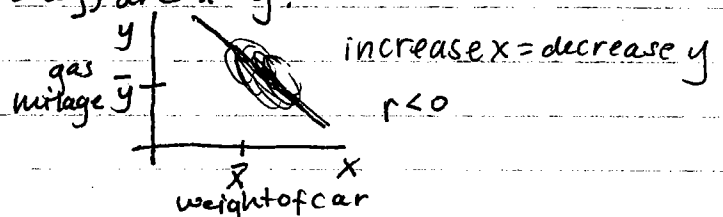
this graph is called a scatter plot here, called elliptical scatter plot (football shape) (what you get when both variables follow the normal curve)

x: independent variable (predictor)  
y: dependent variable (outcome)

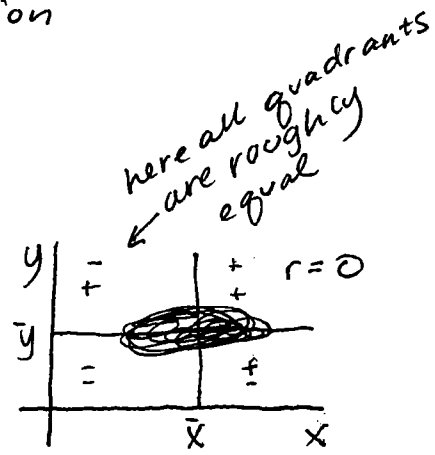
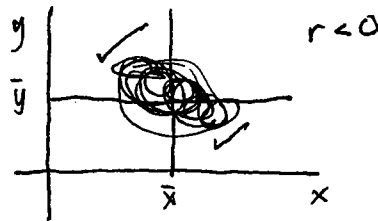
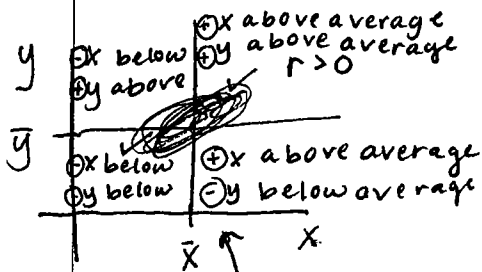
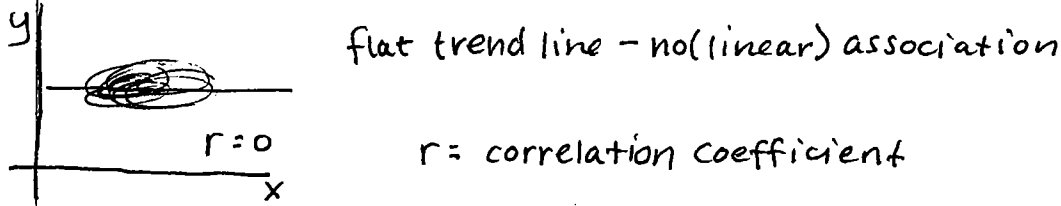
Q: How strongly related (in a linear way) are x + y?



slope positive; x + y positively associated



slope negative; x + y negatively assoc.



✓ = most of data in plot

divided into the 4 quadrants of scatter plot

mentally converting  $x + y$  to standard units:

$r =$  average of the products of  $x + y$  variables in standard units

official formula:

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

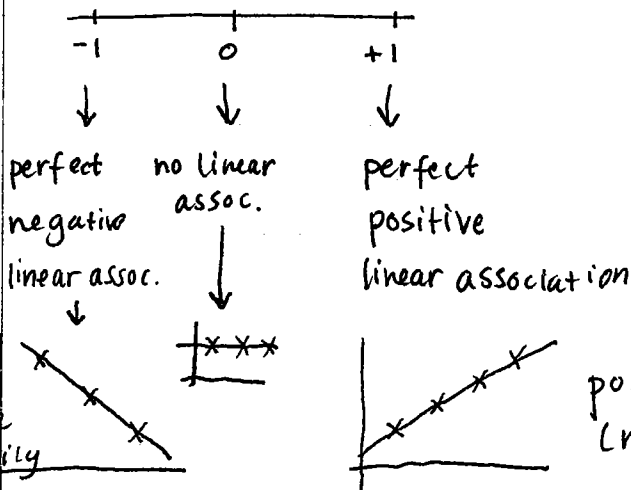
$\left\{ \begin{array}{l} s_x^* : \text{SD but dividing by } n \text{ instead of } (n-1) \\ s_x^* = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2} \end{array} \right.$   
 + same for  $s_y^*$

What's the correlation between heights of fathers + sons?

$r = +$

facts about  $r$ :

①  $r$  always comes out between  $-1 + +1$



see p. 104 for chart of different  $r$  values on scatter plots

pos. slope (not necessarily +1)

neg slope (not necessarily -1)

