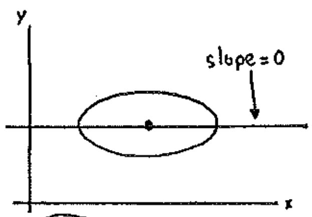
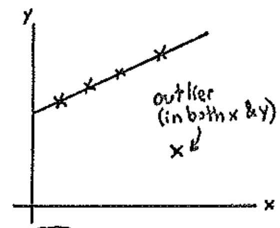
This time: correlation & regression

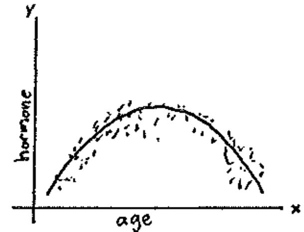
Next time: finish correlation & regression

Read: LN pp. L-245 - 268 This time: LN p.p. L-225 →

facts about r $r = 0$ → 3 different scatterplot slope possibilities





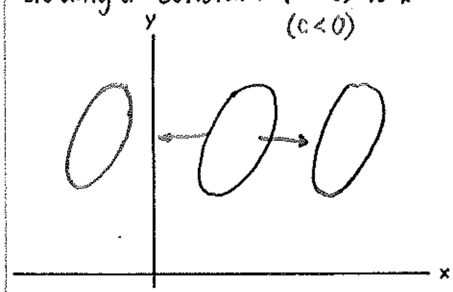


$r \doteq 0$ no linear assoc. so not useful in predicting y

$r \doteq 0$ outliers can distort r, especially with small n

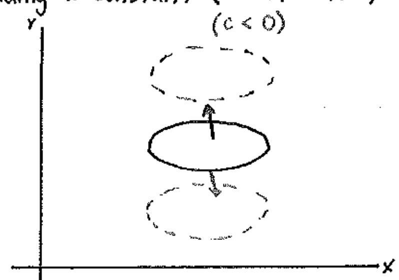
$r \doteq 0$ strong non-linear relationship (this ex. is quadratic)

adding a constant ($c > 0$) to x ($c < 0$)



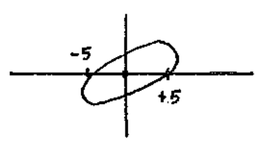
leaves r unchanged!

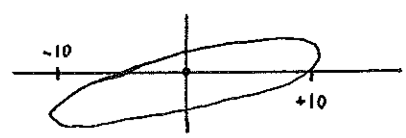
adding a constant ($c > 0$) to y ($c < 0$)



leaves r unchanged!

multiply all x by 2: same tilt!
(mult. by a neg. number gives a reverse tilt)





$r = +0.87$ $\left(\overset{\text{wing}}{\ell}, \overset{\text{tail}}{\ell}\right)$ sparrows example $\hat{y}$ = predicted y

$\boxed{\text{Q:}}$ is this r large in practical terms?

$\boxed{\text{A:}}$ smallest $x = 10 \rightarrow \hat{y} = 7\,\text{cm}$ tail ℓ
largest $x = 11.5 \rightarrow \hat{y} = 8.25\,\text{cm}$

This difference is large in practical terms so r is practsig!

Pg. 230

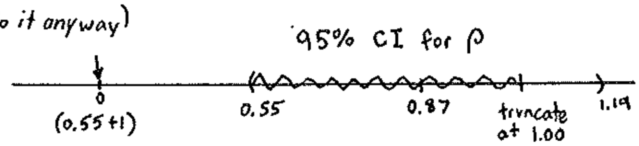
for population model: mean = μ_y, SD = σ_y, corr = ρ (Roh)

math fact: $E_{IID}(r) \doteq \rho$

$$SE_{IID}(r) = \boxed{\sqrt{\frac{1-\rho^2}{n-2}}} \qquad \widehat{SE}_{IID}(r) = \sqrt{\frac{1-r^2}{n-2}}$$

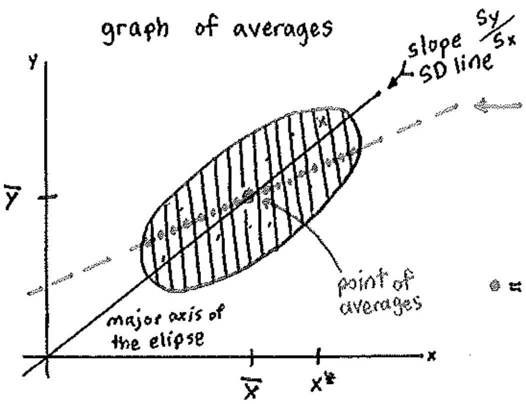
$r = +0.87$ $\quad r \pm 1.96\ \widehat{SE}(r)$ $\quad$ large-n approx. 95% CI

$\widehat{SE}(r) = 0.16$ $\quad$ (12 sparrows is not large n, do it anyway)



Predicting y value from x

graph of averages



← regression line: best line for predicting y from x

slope $r \cdot \frac{s_y}{s_x}$

discovered by Galton

● = y averages for the strip

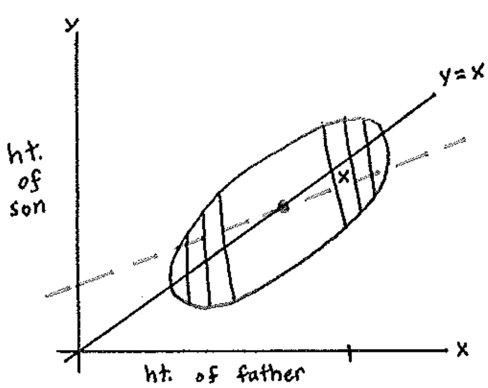
predict y value from x^*

beta

$\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x}$ $\quad$ regression line equation $\quad$ $\boxed{\hat{\beta}_1 = r \frac{s_y}{s_x}}$

y-intercept $\quad$ slope

$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \rightarrow \boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$



$n \doteq 1000$ families with at least 1 son

a random son + ht

father + ht

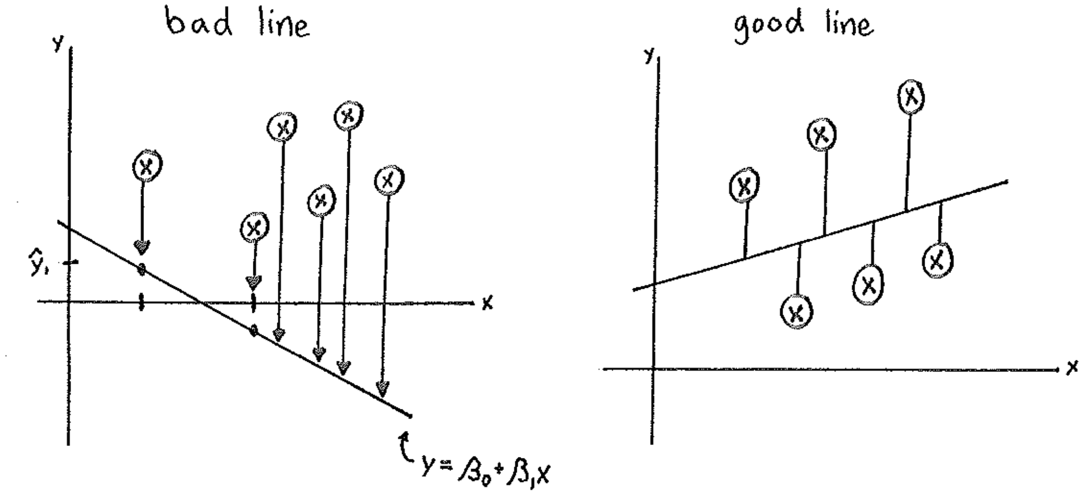
"regressing" toward the mean:

tall fathers have tall sons but not as tall as they were

short fathers had short sons, but not as short as them

note: cm of tail length ≠ cm of wing length → units don't cancel!

Another way to get the best line for pred. y from x



$$\frac{1}{n} \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i) \right]^2$$

find $\beta_0 \, \beta_1$ to minimize

result: least squares line (Gauss 1800)

math fact: regression line = least squares line