this time: Correlation & regression

next time: analysis of variance

read: LN pp. L-214 → L-259

AMS 7 15 Jul 2016

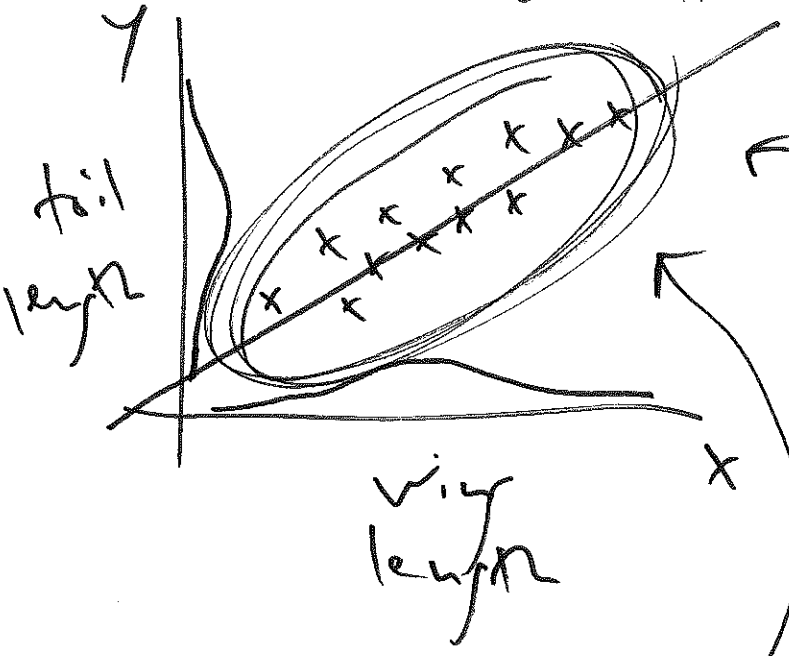
①

today: LN pp. L-214 →

please go to myuw.wisc.edu & fill out the online course evaluation by Fri 22 Jul

Lab 3 due Mon 18 Jul in class

scatter plot (scatter diagram)



← elliptical shape

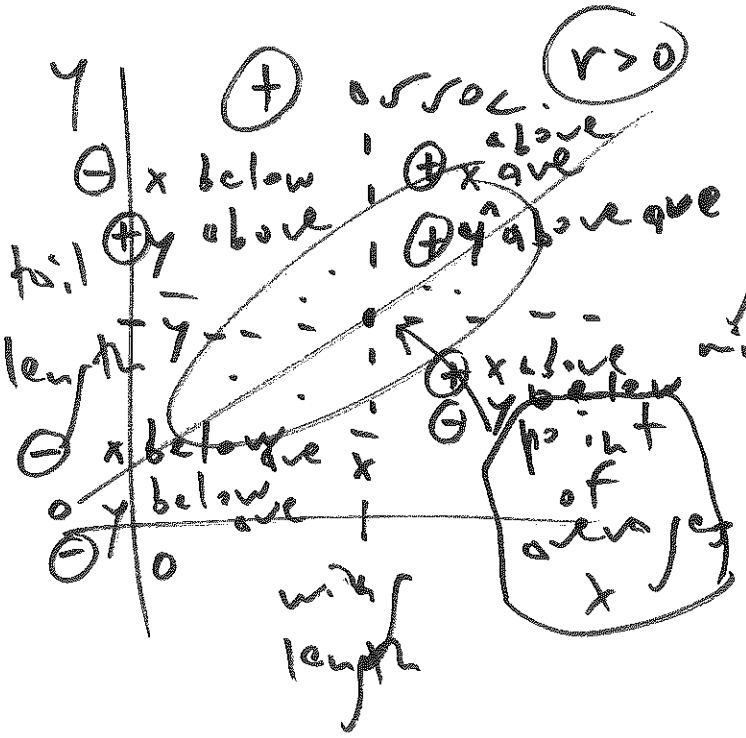
(bivariate normal dist.)

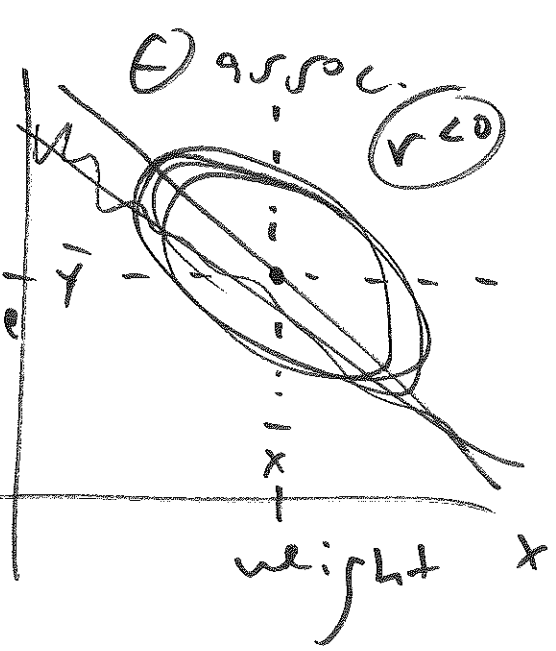
(x is normal, y is normal)

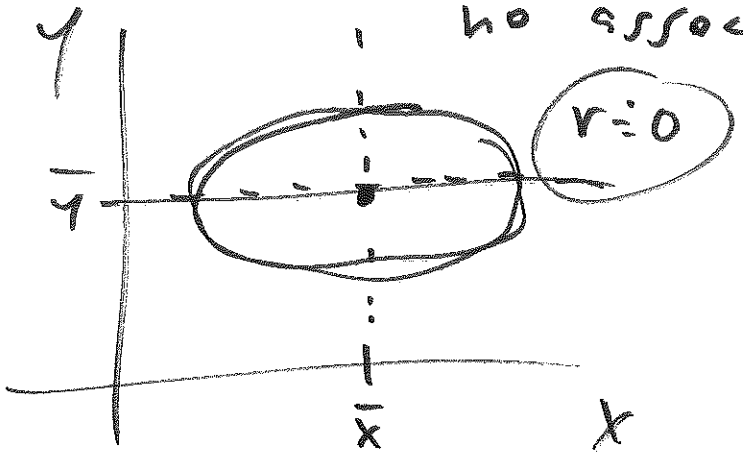
Karl Pearson (1890)

positive association

as ↑ x, y ↑ on ave.







correlation (coefficient)

r = strength of linear association between x & y

$$\begin{bmatrix} y_1 & x_1 \\ y_2 & x_2 \\ \vdots & \\ y_n & x_n \end{bmatrix} \quad \uparrow \; n \; \downarrow$$

mean $\bar{y}$ $\bar{x}$

SD s_y s_x

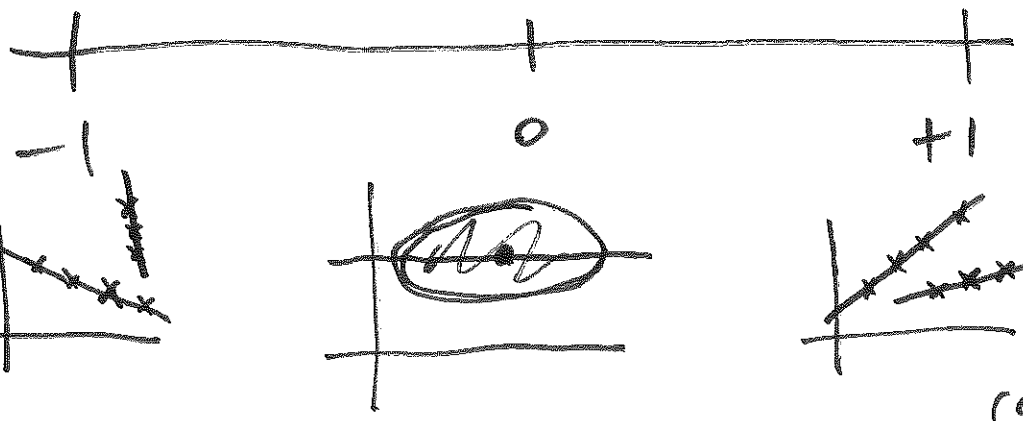
$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x^*}\right)\cdot\left(\frac{y_i - \bar{y}}{s_y^*}\right)$$

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

math facts about r

① r is a pure number (no units) ③

② r



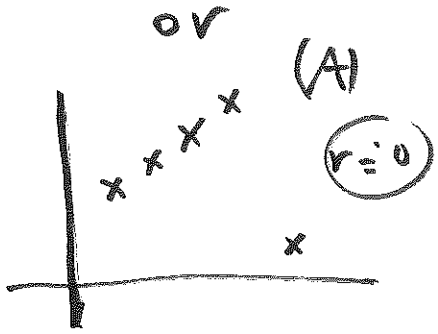
$-1 \leq r \leq +1$

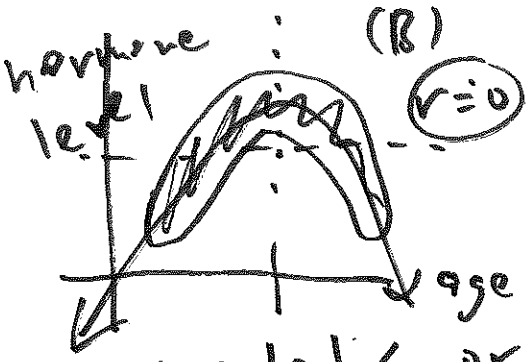
(9.55)

r can be fooled

by (A) outliers

or (B) non-linearity



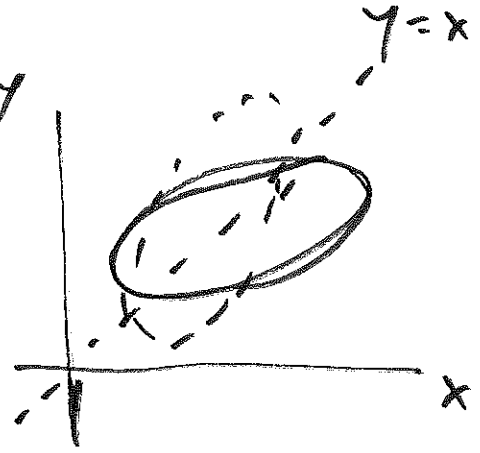


parabolic or quadratic shape (not linear)

④ Corr. of any x with itself is +1

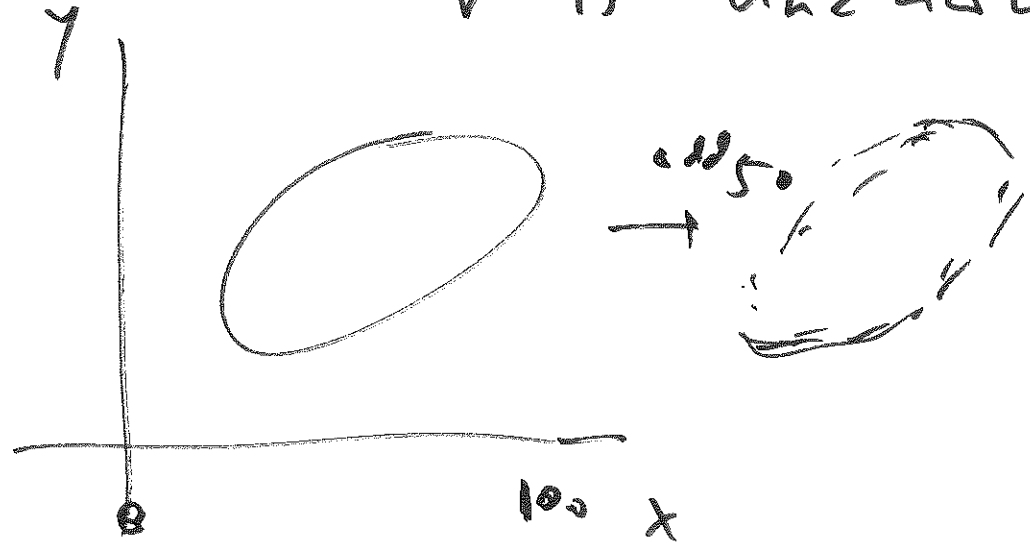
Spearman:

$r = +0.87$

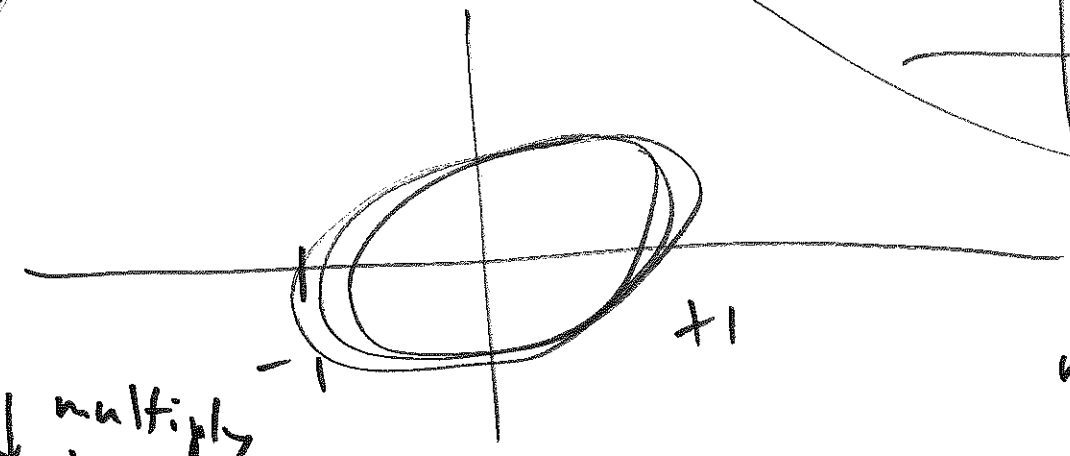


③ r is unchanged when the roles of x & y are reversed

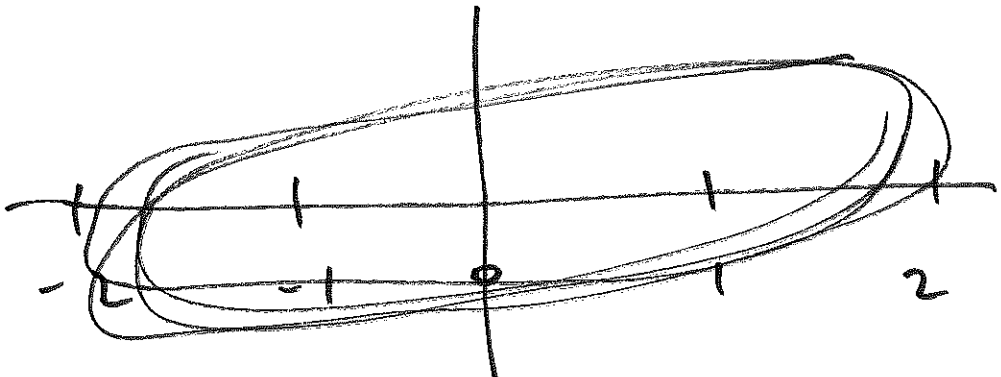
⑤ if you add (or subtract) a ④
constant to (from) all the x values,
r is unchanged; same with y



⑥



↓ multiply by 2



if you multiply x by a positive constant, r stays the same; same with y

using r inferentially | Q1: is a corr. between wing length (x) and tail length (y) of r = +.87 large in practical terms? | A: smallest sparrows have (x = 10 cm, y = 7 cm); largest sparrows have (x = 11.5 cm, y = 8.25 cm); is the difference between 7 cm for y & 8.25 cm large in practical terms? $\frac{8.25\text{ cm} - 7\text{ cm}}{7\text{ cm}} = \frac{1.25}{7} \approx 18\%$

Q2: is this corr. of +.87 large in statistical terms?

null hovering value: 0 / modeling

math facts

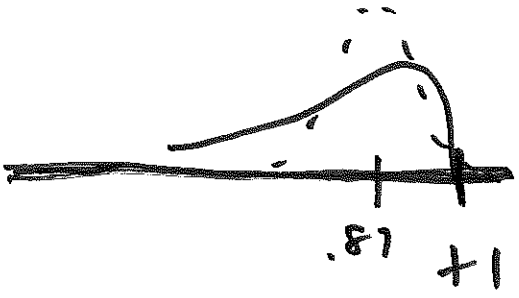
① $E_{IID}(r) \doteq \rho$

② $SE_{IID}(r) = \sqrt{\dfrac{1-\rho^2}{n-2}}$ don't know ρ

so use $\widehat{SE}_{IID}(r) = \sqrt{\dfrac{1-r^2}{n-2}}$

here $\widehat{SE}(r) = \sqrt{\dfrac{1-(+.87)^2}{12-2}} \doteq 0.156$

$\doteq 0.16$



long run hist of r ⊛

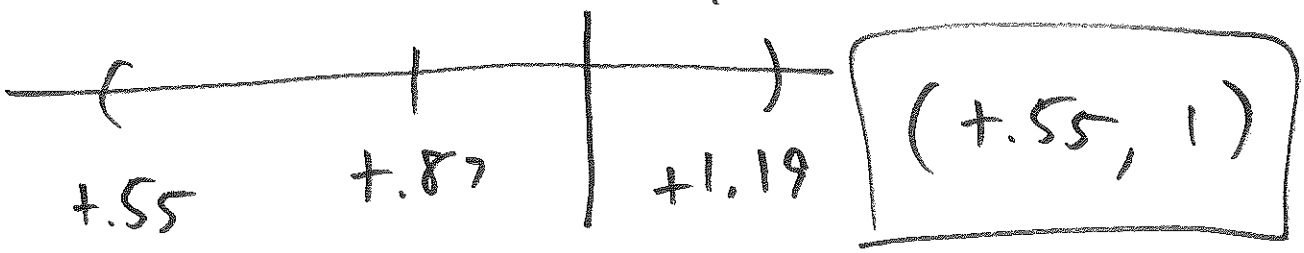
let's try a large-sample (large n)

normal approximation: approx. 95% CI

for $\rho \rightarrow r \pm 1.96\,\widehat{SE}(r)$

here this becomes $(+.87) \pm 1.96(0.16)$ (7)

95% int. for ρ 0.32



+1 truncate at +1

approx (+.55, +1)

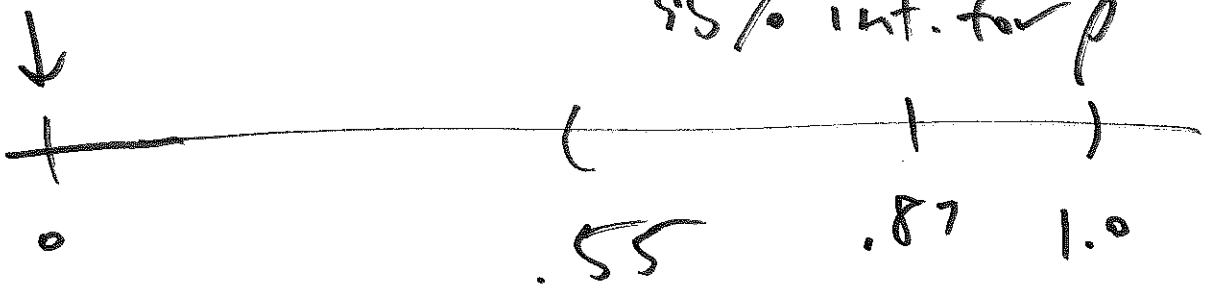
exact (+.59, +.96)

conservative (wider)

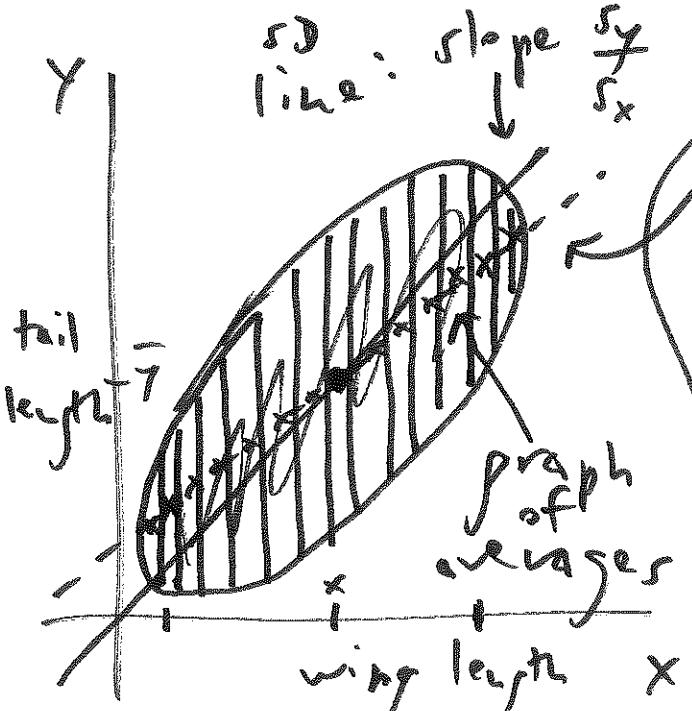
LN (L-231) - (-244)

pr. optimal

approx 95% int. for ρ



0 (may) not in int; r=.87 is highly statsig different from 0 (12.53)



regression line

Q: what's the equation of the line relating $\bar{x}$ & $\bar{y}$?

A. equation of best line for predicting $\bar{y}$ = tail length from $\bar{x}$ = wing length

A₁: (goal) use line to predict $\hat{y}$ from x

A₂: (goal) use line to predict x from y

A₃: (goal) capture trend of x-y relationship

Francis Galton (1905)

here $\beta_1 = r \frac{s_y}{s_x} = \frac{+.8704}{(\text{illegible})} \frac{0.3499 \text{ tail length}}{0.3950 \text{ wing length}}$

$= 0.771 \frac{\text{cm of tail length}}{\text{cm of wing length}}$

slope of regression line

$$\hat{\beta}_1 = r \cdot \frac{s_y}{s_x}$$

math fact

beta

(y)-intercept of reg. line

reg line has to go through $(\bar{x}, \bar{y})$:

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$ (so) $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$

so $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$\hat{y}$: predicted y-value

$\hat{\beta}_0$: est. intercept

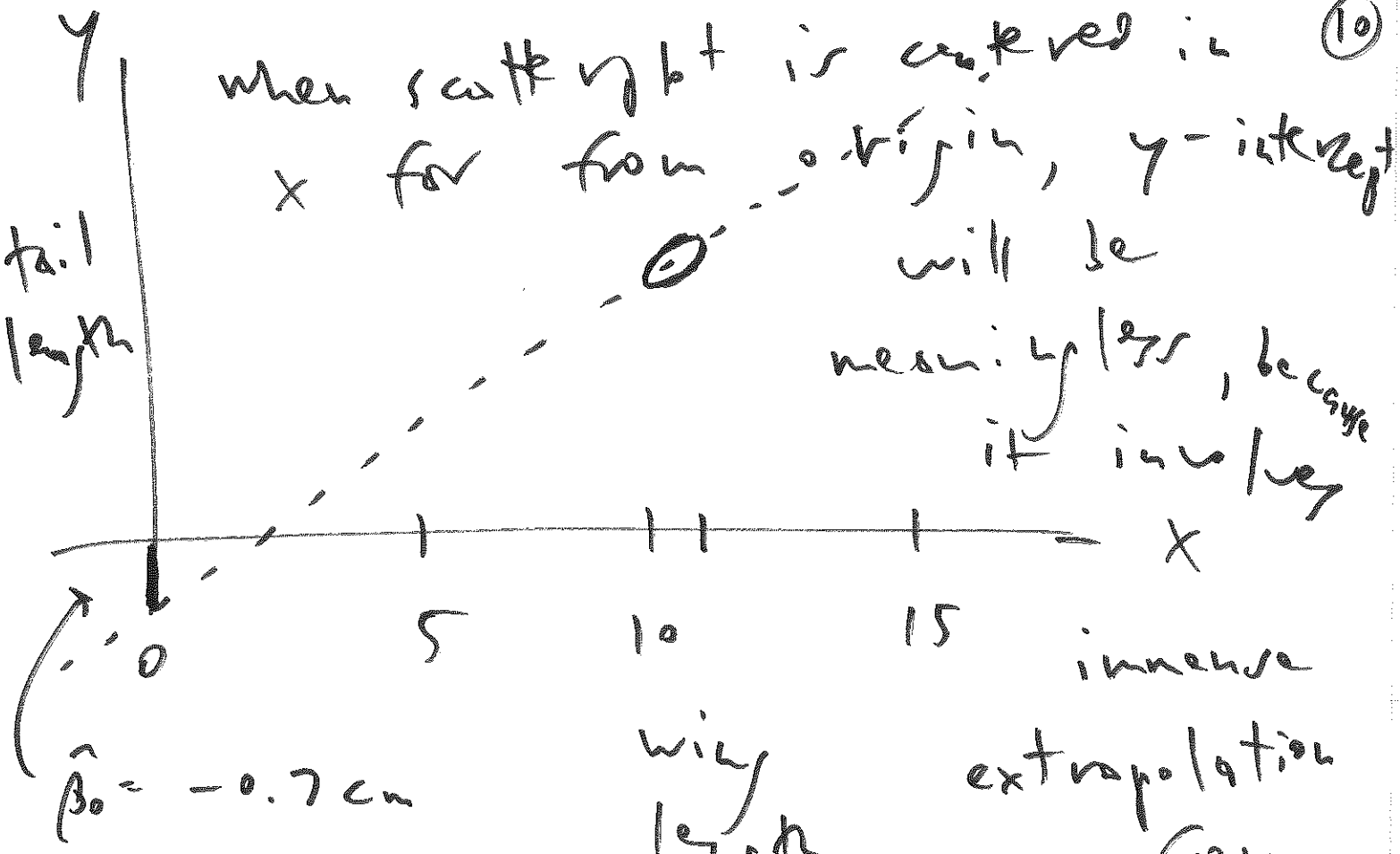
$\hat{\beta}_1$: est. slope

here y-int. is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -0.669 \text{ cm of t.l.}$$

$$= 7.567 \left(\frac{\text{cm of tail length}}{\text{(t.l.)}}\right) - \left(0.771 \frac{\text{cm of t.l.}}{\text{cm of w.t.}}\right)\left(10.6\text{F cm of w.t.}\right)$$

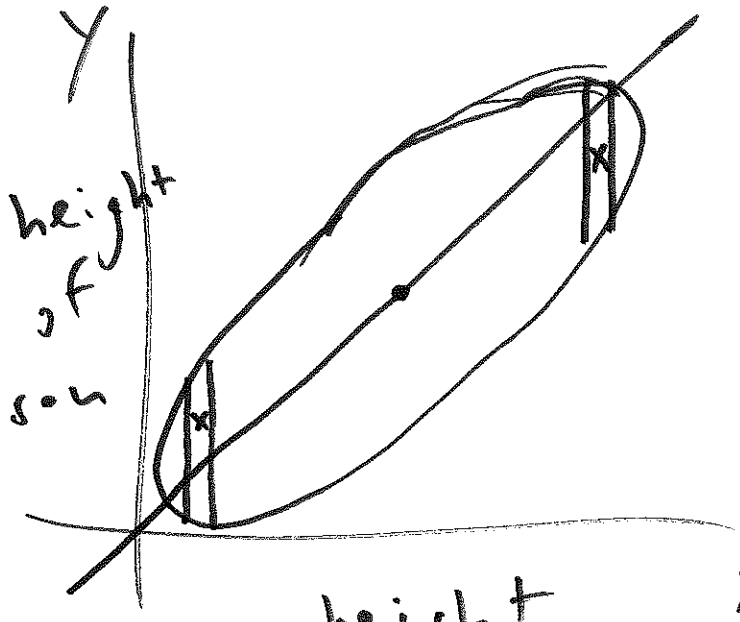
w.t. = wing



when scatterplot is centered in x far from origin, y-intercept will be meaningless, because it involves immense extrapolation away from the data

why length

why regression?



tall (short) father tend to have tall (short) sons, but not as tall (short) as the father