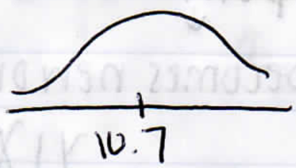
11/20/18
Wave Moretto

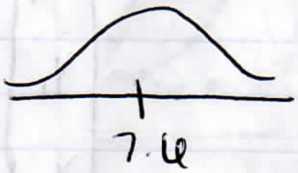
Lecture # 16:
Correlation and Regression

Y = tail length (cm)
X = wing length (cm)

sample



hist
X



hist
y

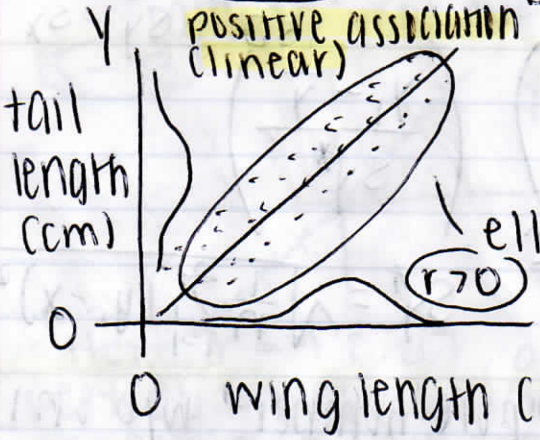
Y	X
7.4	10.4
7.6	10.8
⋮	⋮
8.3	10.4

n = 12

$\bar{y} = 7.6$ cm $\bar{x} = 10.7$ cm

SD $s_y = 0.35$ SD $s_x = 0.43$ cm

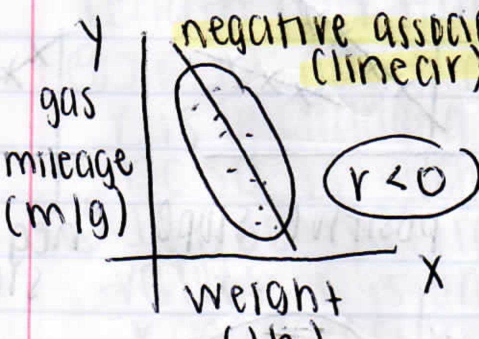
L-216

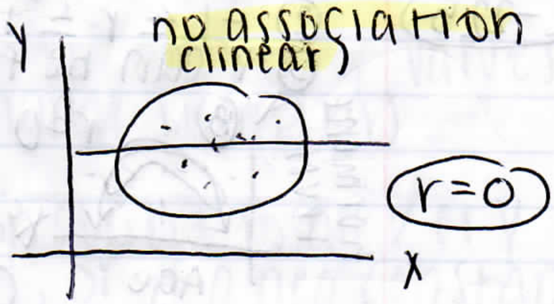


scatterplot (scatterdiagram)
← typical w/ 2 normal curves

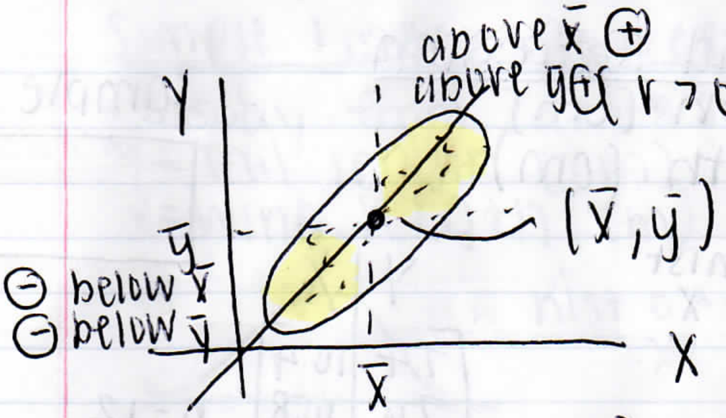
elliptical shape (bivariant normal)

most scatterplots we see in class

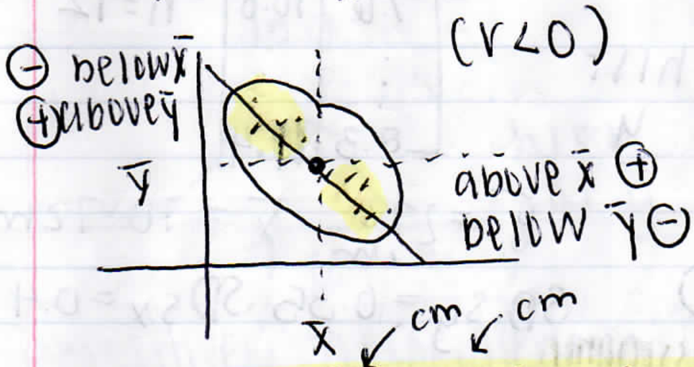




• Karl Pearson & ~~Frank~~ Francis Galton (1890s)



($r < 0$)



$$
\begin{array}{c|c}
y & x \\
\hline
y_1 & x_1 \\
y_i & x_i \\
y_n & x_n \\
\hline
\bar{y} & \bar{x} \\
s_y & s_x
\end{array}
\quad n
$$

mean

SD

cm, cm, 5cm

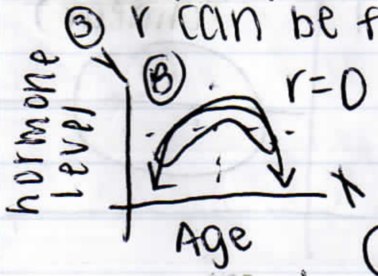
$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x^*} \right) \left(\frac{y_i - \bar{y}}{s_y^*} \right)$$

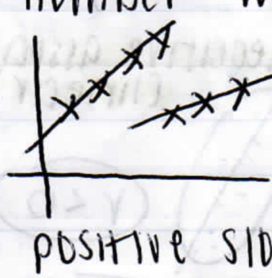
correlation between x and y

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

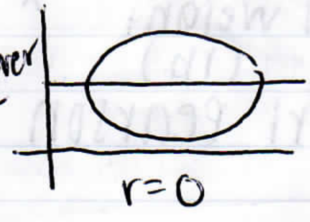
L-225

1. r is always a "pure number" w/o units
2. $-1 \le r \le +1$
3. r can be fooled...

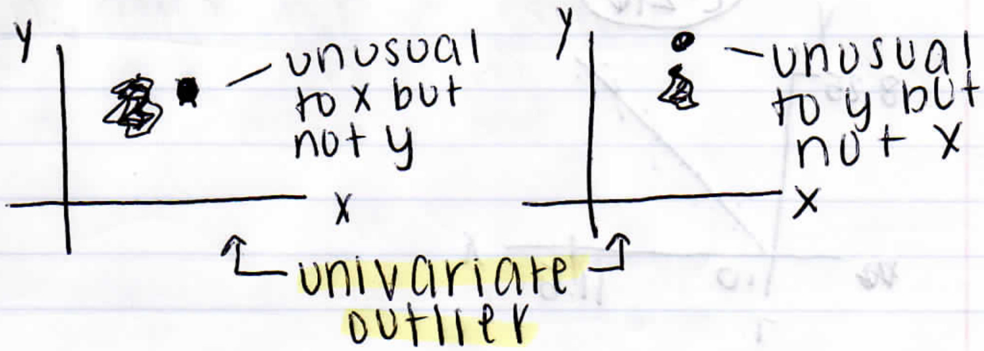


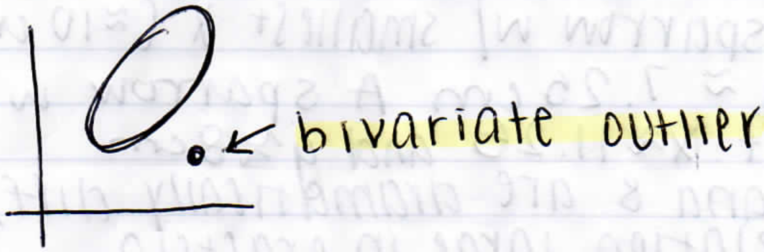


however



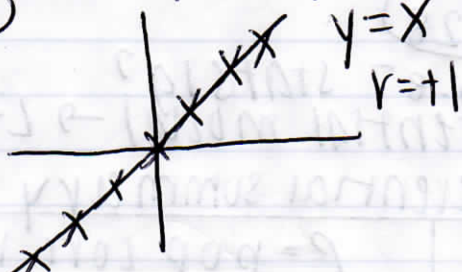
- fooled by outliers Ⓐ or nonlinearity Ⓑ





(R-73) - page of scatterplots
- training your eye to read correlation values

(R-73)



$$\left[\begin{array}{c|c} y & x \\ 3 & 3 \\ 9 & 9 \\ -4 & -4 \end{array}\right]$$

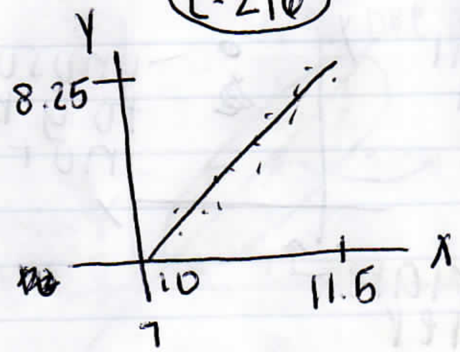
• Here $r = +0.87$, a strong but not perfect linear attraction between wingth length and tail length

④ If you add a constant to all x or y values, r is unchanged (just change location) bc SD is unchanged

⑤ If you multiply a constant to all x or y values, r is unchanged, or if a neg constant, r changes sign

Q: Is an R of +0.87 large in practical terms? (is it pract sig?)

(L-216)



A: A sparrow w/ smallest x ($\approx$10 cm) has y $\approx$ 7.25 cm. A sparrow with largest x$\approx$11.25 and y$\approx$8 cm
- 7.5 and 8 are dramatically diff, so correlation large in practsig (sharply different from 0)

(L-228)

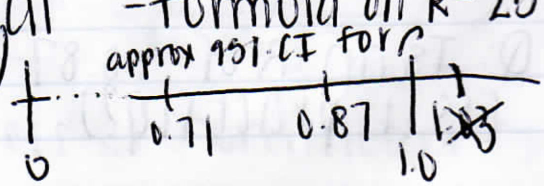
Q: Is r = +0.87 statsig?
A: Buid inferential model → L-229

inferential summary

pop	unknown pop q of interest	ρ = pop corr between wing and tail length in this species
sampu	estimate of ρ	r = +0.87
imag.	Give or take for ~~ra~~ r as est. of ρ	$\hat{SE}$ (r) = 0.081
data	95% CI for ρ	approx (0.71, 1.0) exact (0.59, 0.96)

*L-231 → L-244 : extra credit material
↳ takehome final -formula on R-25

$+0.87 \pm \underbrace{(2)(0.081)}_{0.16}$



A: Diff is statsig!