AMS 7 17 Nov 09

1

this time: regression

next time: ANOVA

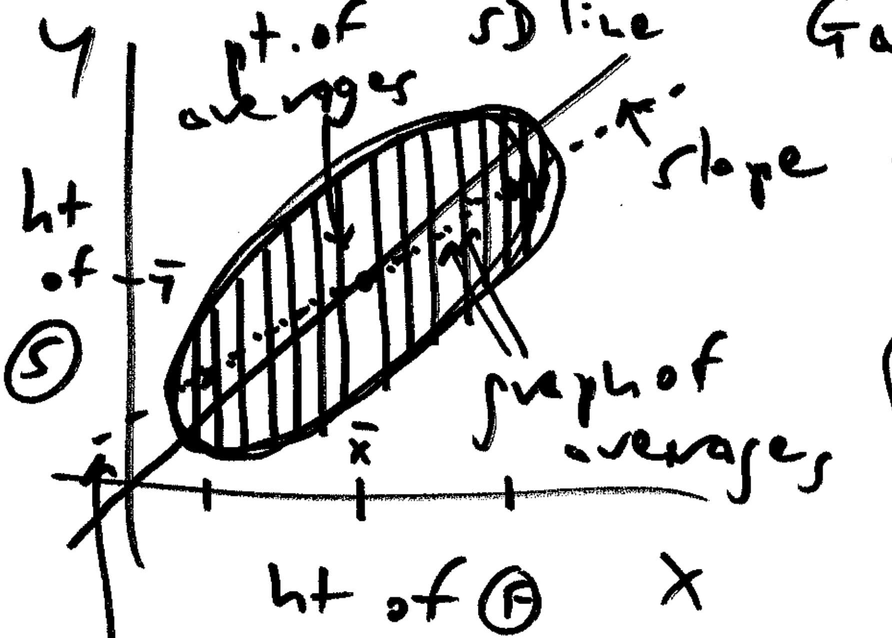
read: lecture notes pp. L-269 → L-289

hwk 4 due Tue 24 Nov

lab 6 due by 5pm Mon 23 Nov

next week: M-W: class as usual on Tue but no disc. sec & no labs M 23 Nov → Wed 25 Nov (Thu - Fri holiday)

Galton (1880s)



slope of line that smooths graph of averages

regression line

predicted y = estimated y-intercept + (estimated slope)·(x)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

math fact: $\boxed{\hat{\beta}_1 = r \cdot \frac{s_y}{s_x}}$

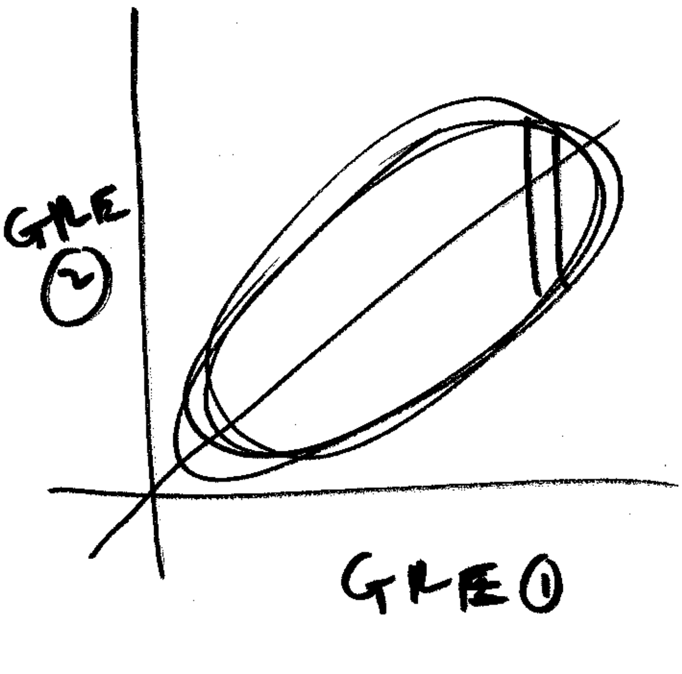
$= (.8704) \dfrac{0.3499 \text{ cm of tail length}}{0.3950 \text{ cm of wing length}}$

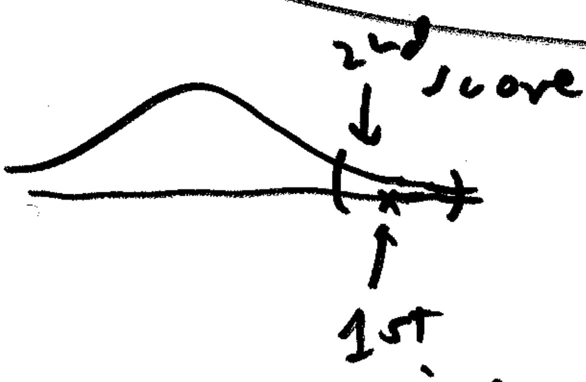
$= .771 \dfrac{\text{cm tail}}{\text{cm wing}}$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

cm tail

$\left(\dfrac{\text{cm tail}}{\cancel{\text{cm wing}}}\right)\left(\cancel{\text{cm wing}}\right)$





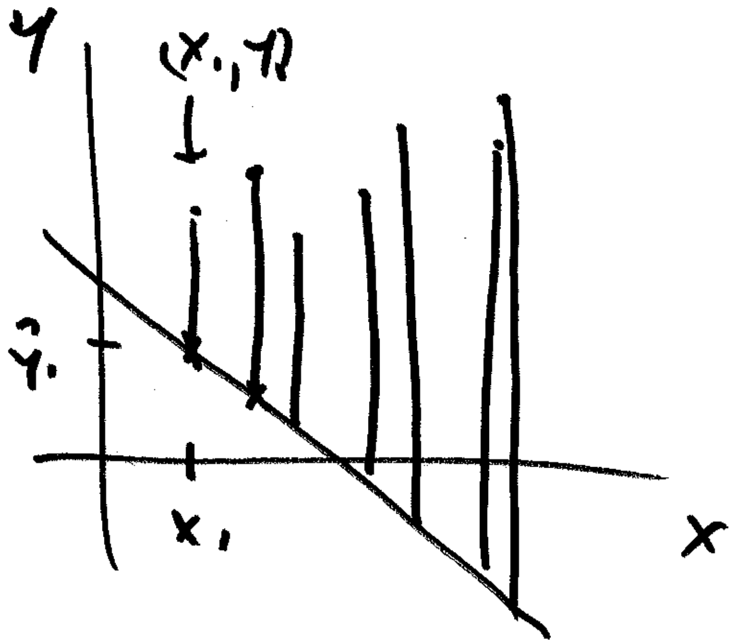
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ | when $x = \bar{x}$, $\hat{y} = \bar{y}$:

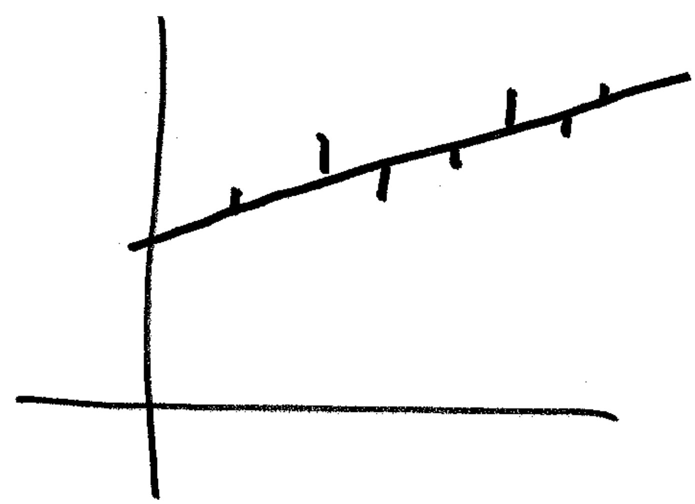
$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

(Laplace $\sum_{i=1}^{n} |y_i - \hat{y}|$)

Gauss (1800)

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$





find line that minimizes $\sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right]^2$:

this is the least squares (LS) line

math fact: LS line = regression line

math facts:

① $E_{IID}(\hat{\beta}_1) = \beta_1$

② $\hat{SE}_{IID}(\hat{\beta}_1) = \dfrac{s_{Y|X}}{s_X \sqrt{n-1}}$ ← Y given X

where $s_{Y|X} = s_Y \sqrt{1-r^2} \sqrt{\dfrac{n-1}{n-2}}$

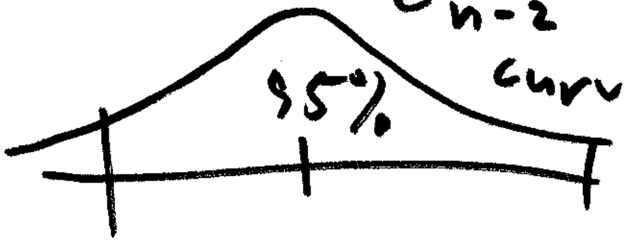
residual SD = "root mean square(d) error" (JMP)

combine to get

$$\hat{SE}_{IID}(\hat{\beta}_1) = \frac{s_Y \sqrt{1-r^2}}{s_X \sqrt{n-2}}$$

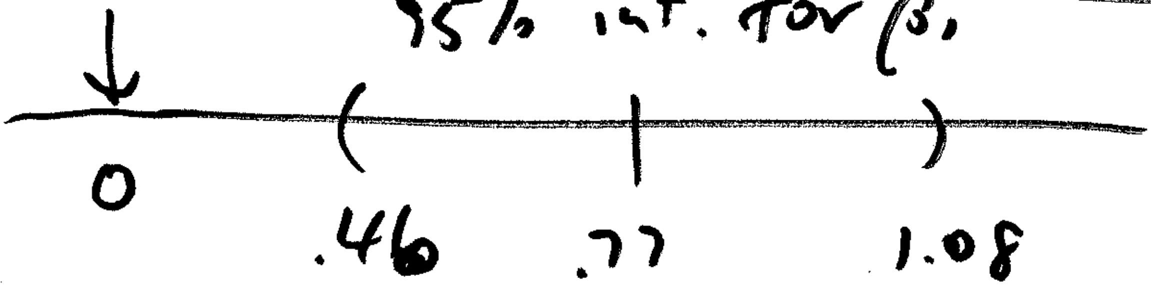
$\hat{SE}$ 0.14

t_{n-2} curve ← t_{10}



long run list of $\hat{\beta}_1$, accounting for all relevant uncertainties

2.228 ← $t_{10}^{0.95}$

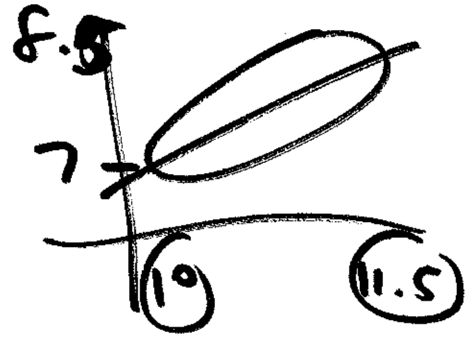


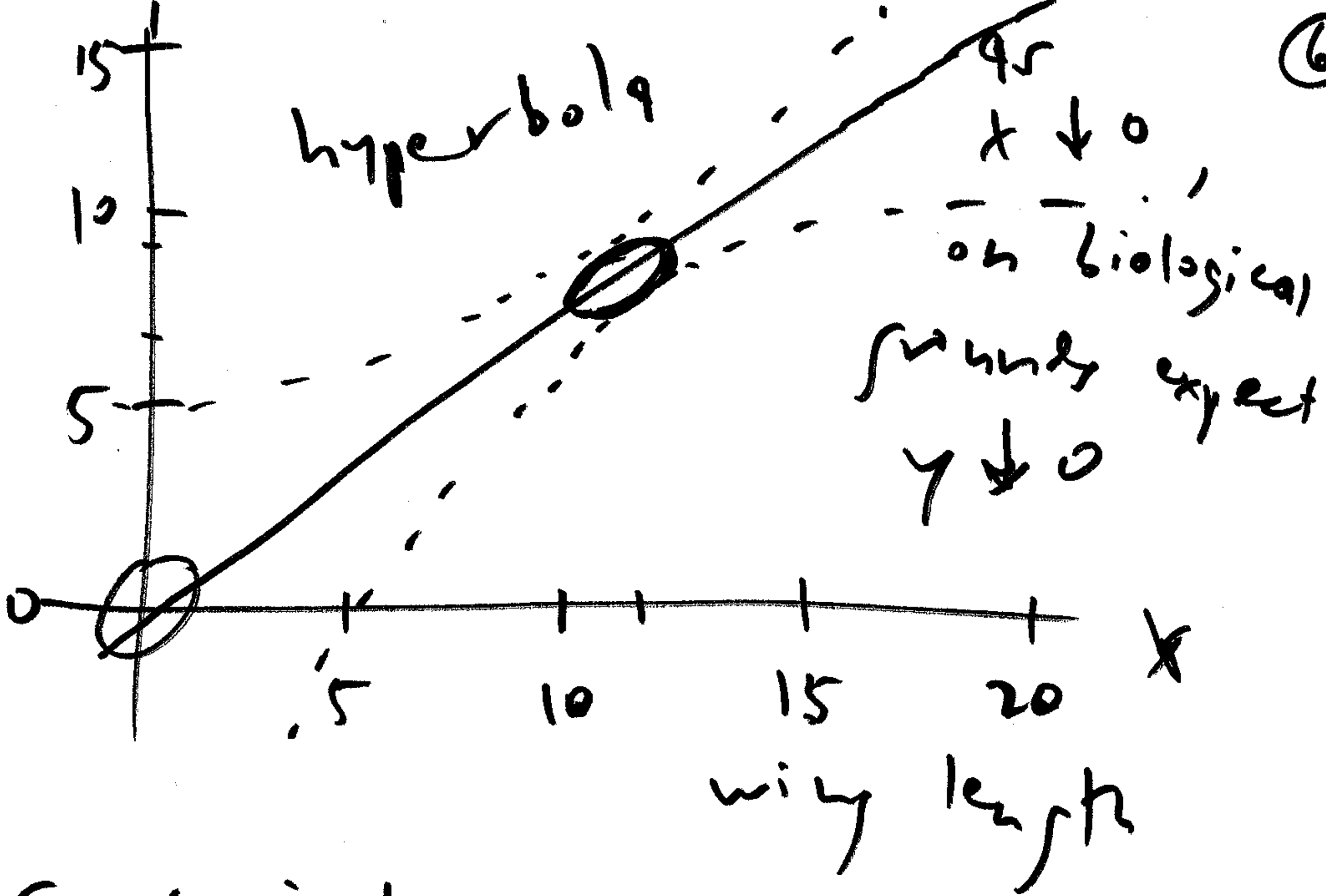
0 is not in 95% int., so this slope is stat. sig.

Q: Is this slope large in practical terms?

A: ~~yes~~

A: same answer as Q: is corr. practic. s?

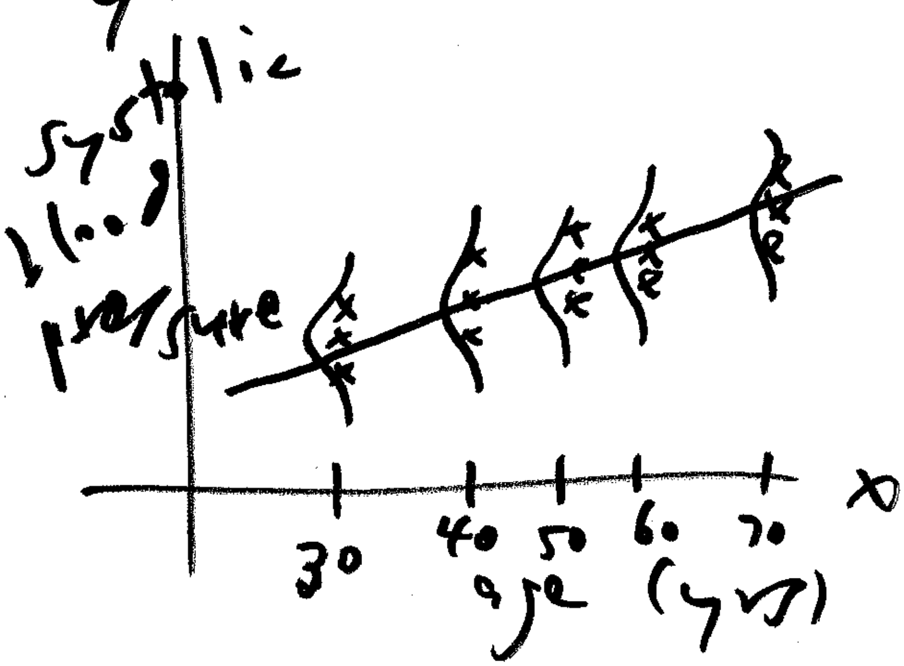




so y-intercept should have not been stat sig different from 0 (& it wasn't)

---

extrapolation far away from bulk of data in x is risky in regression



$$\begin{array}{ccc} \text{predicted b.p.} & \text{obs. b.p.} & \text{age} \\ \begin{bmatrix} \hat{y}_i \end{bmatrix} & \begin{bmatrix} y_i \end{bmatrix} & \begin{bmatrix} x_i \end{bmatrix} \end{array} \quad n$$

another way to think about regression

(meas. error)

obs = true + ~~bias~~ + random "error"

$$y_i = \theta + \cancel{b} + e_i$$

model

obs = truth + random "error"

$$y_i = (\beta_0 + \beta_1 x_i) + e_i$$

obs

obs

$i = 1, \ldots, n$

unobs

IID normal with mean 0 & SD $\sigma_{Y|X}$

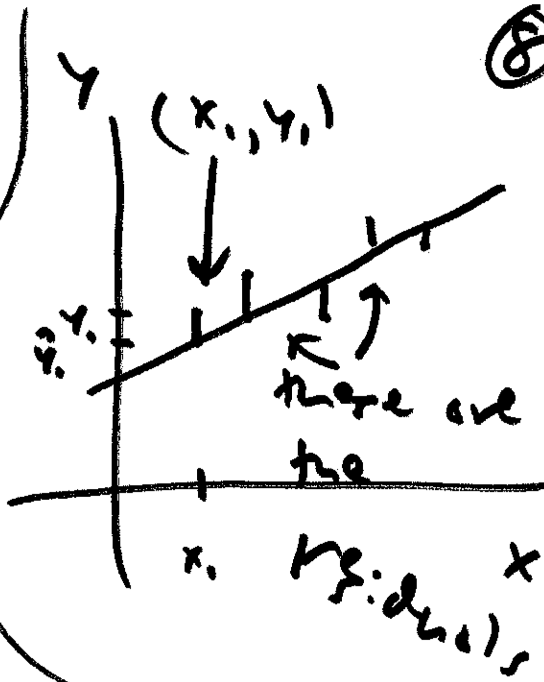
data

obs

$$y_i = \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{\text{predicted } (\hat{y}_i)} + \hat{e}_i$$

where $\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ = obs. − predicted

$\hat{e}_i$ = the residuals



$Y_i = \beta_0 + \beta_1 x_i + e_i$

IID normal mean 0 SD $\sigma_{Y|X}$

this is the residual SD : a good est. of it would be

$$\hat{\sigma}_{Y|X} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} \left(\hat{e}_i - \cancel{\bar{\hat{e}}}\right)^2}$$

$$\hat{\sigma}_{Y|X} = \sqrt{\underbrace{\frac{1}{n-2} \sum_{i=1}^{n}}_{\text{mean}} \hat{e}_i^{\,2}}$$

root mean squared error

RMSE ← JMP