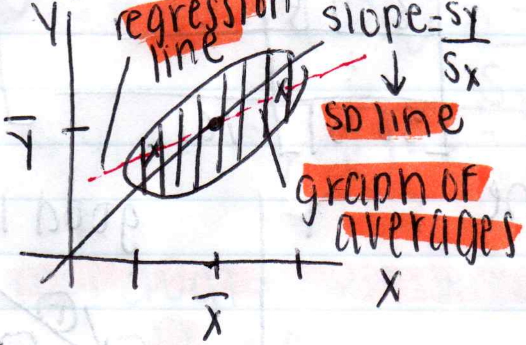
11/27/18
Wave
Moretto

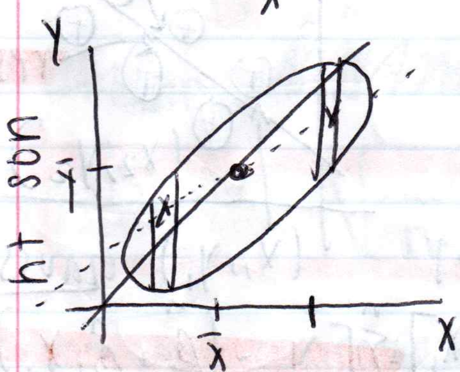
Regression (Lecture #17)

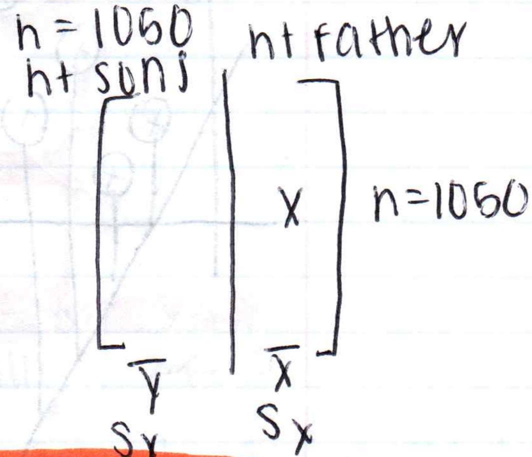
L-216

- y = tail length x = wing length



Best line for predicting y from x?
- Gauss (~1800)
- Francis Galton (1890)



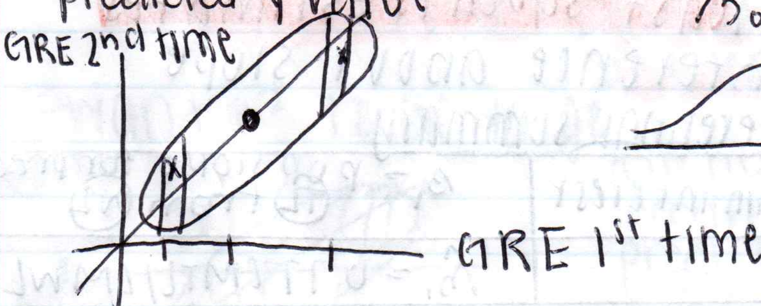


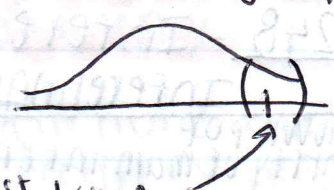
R-25 (Eq. 17, 21)

- $\boxed{\text{Slope of regression line for predicting } y \text{ from } x = \hat{\beta}_1 = r\frac{s_y}{s_x}}$
- Equation of regression line:

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

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



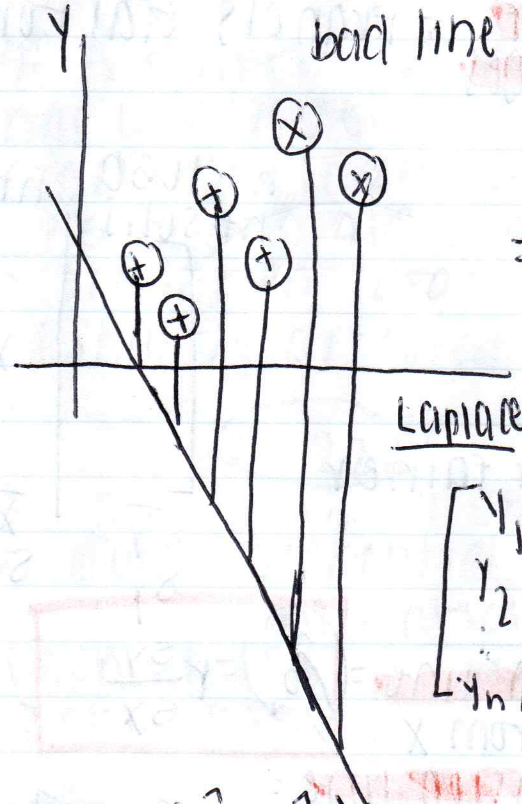


L-217 - jump example

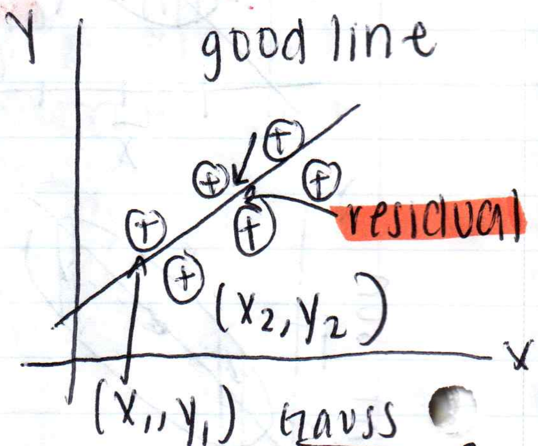
5050

$$\left[\begin{array}{cc|c} 1 & 100 & 101 \\ 2 & 99 & 101 \\ 3 & 98 & 101 \\ & \vdots & \\ 50 & 51 & 101 \end{array}\right] \; 50$$

Gauss



$\Rightarrow$



Laplace 1·1

Gauss

$$\begin{bmatrix} y_1 & x_1 \\ y_2 & x_2 \\ \vdots & \vdots \\ y_n & x_n \end{bmatrix} \; \frac{1}{n}\sum_{i=1}^{n}\left[y_2 - (\hat{\beta}_0 + \hat{\beta}_1 x_1)\right]^2$$

- Find $(\hat{\beta}_0, \hat{\beta}_1)$ to minimize $\frac{1}{n}\sum_{i=1}^{n}\left[y_2 - (\hat{\beta}_0 + \hat{\beta}_1 x_1)\right]^2$
- Result is least squares estimates

L-248 - Inference about slope

Inferential summary

unknown pop quantity of main interest	β_1 = pop slope for predicting TL from WL
estimate	$\hat{\beta}_1 = 0.77$ CMTL/CMWL
Give or take	$\hat{SE}(\hat{\beta}_1) = 0.14$ CMTL/CMWL
95% CI	$\hat{\beta}_1 \pm t_{n-2}^{0.95}\hat{SE}(\hat{\beta}_1) = (0.46, 1.08)$

Facts

① $E_{IID}(\hat{\beta}_1) = \beta_1$

② $\hat{SE}_{IID}(\hat{\beta}_1) = \dfrac{S_{y|x}}{S_x \sqrt{n-1}}$ ← "given" where

$$S_{y|x} = S_y \sqrt{1-r^2} \cdot \sqrt{\frac{n-1}{n-2}}$$

residual SD = root mean squared error

$$\boxed{\hat{\sigma}_{y|x} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} \underbrace{\hat{e}_i^{\,2}}_{\text{squared error}}}}$$

Is the regression practically useful?

2 ways to answer:

① r^2, recall variance $v(y) = S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

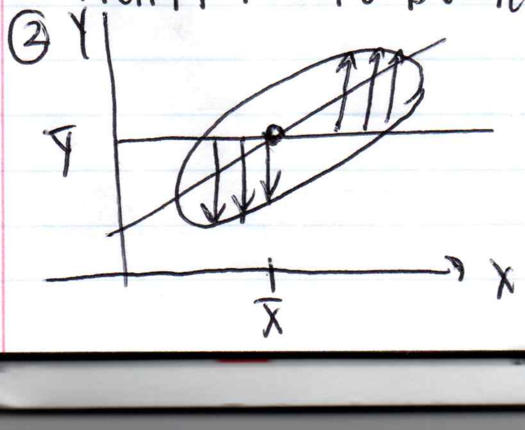
in regression $y = \hat{y} + \hat{e}$ so $v(y) = v(\hat{y} + \hat{e})$

$v(\hat{y} + \hat{e}) = v(\hat{y}) + v(\hat{e})$

so $\boxed{r^2 = \dfrac{v(\hat{y})}{v(y)}}$ = % of variance in y associated w/ regression of y on x

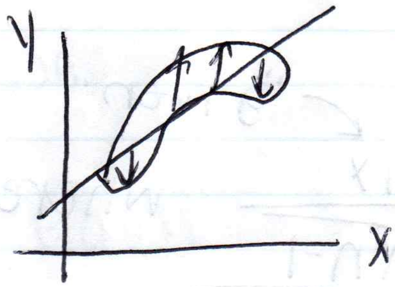
-want r^2 to be large

②



a) ignore x or don't know x predict y anyway - best prediction is $\hat{y} = \bar{y}$ give or take for s_y

b) use x to predict y, best estimate is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$ give or take for this $\hat{y}$ is smaller



If scatterplot looked like this, simple linear regression not good...

→ this is a residual plot