read:
LN pp. L-269 → L-301
today:
LN pp. L-248

this time: regression; ANOVA
next time: " ; "

HW 4 due Fri 8 Jun 18 by 11:59 pm

Ashley Tai
AMS 7
29 May 18

Inference in Regression

pop: the relevant birds of this species

sample: the observed birds

imaginary data: all possible values of $\hat{\beta}_1$

TL	WL
y	x

$N = ?$ (big)

actual → like SRS $\doteq$ IID

TL	WL
y	x

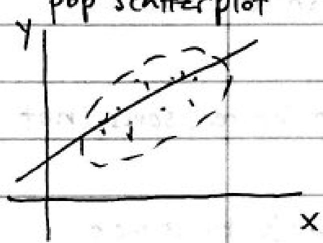
$n = 12$

0.771
0.81
⋮

$M \to \infty$

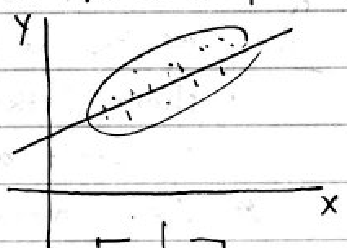
pop slope $\beta_1 = ?$

pop scatterplot



sample slope $\hat{\beta}_1 = 0.771$

sample scatterplot



hyp IID

y	x

ex. 0.81

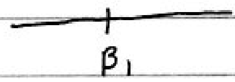
long run mean

$E_{IID}(\hat{\beta}_1) = \beta_1$

est. long run SD

$\hat{SE}_{IID}(\hat{\beta}_1) \doteq 0.14$

long run hist



facts:

① $E_{IID}(\hat{\beta}_1) = \beta_1$

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

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

residual SD = "root mean squared error" (RMSE)

OR

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

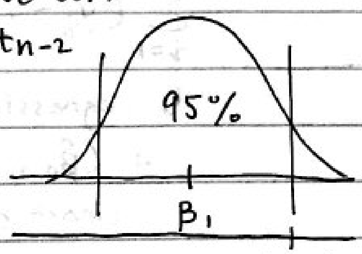
$S_{y|x} = (0.3499)\sqrt{1-(0.8704)^2}\sqrt{\frac{12-1}{12-2}} = 0.1807$ CMTL

$\hat{SE}_{IID}(\hat{\beta}_1) = \frac{0.1807 \text{ CMTL}}{(0.395 \text{ CMTL})\sqrt{12-1}} = 0.1379$ CMTL

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

$\hat{SE}$ 0.14

t_{n-2}



$t_{10}^{0.95}$

95% CI for β_1

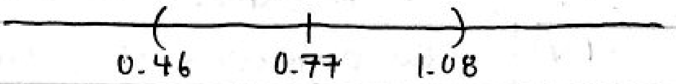
$= 0.7709 \pm 2.228\,(0.1379)$

$= (0.46,\ 1.08)$

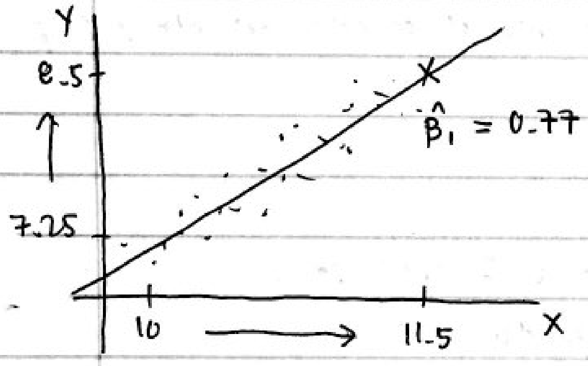
Inferential Summary

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

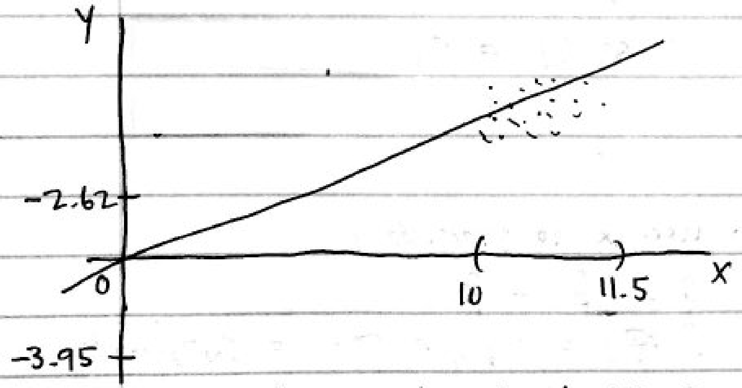
95% CI for β_1



* Is $\hat{\beta}_1$ large in practical terms?



* start from smallest x or y & determine whether difference is large



y-intercept

β_0 = predicted y-value when $x = 0$

* 0 is in the 95% CI for β_0, as it should be, b/c it makes biological sense that as $WL \to 0$, TL will also $\to 0$

Another Way to Think About Regression

$\underbrace{Y_i = (\beta_0 + \beta_1 X_i) + e_i}_{\text{observed = truth + error}}$ unobservable

$\underbrace{Y_i = (\hat{\beta}_0 + \hat{\beta}_1 X_i) + \hat{e}_i}_{\text{observed = predicted + residual}}$ observable

$$\hat{\sigma}_{y|x} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2}$$

(residual SD)

$\hat{e}_i^2$: squared error

$\frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2$: mean squared error

$\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2}$: root mean squared error

★ Is the regression practically useful?

① r^2 : "coefficient of determination"

$V(y) = S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

$V(y) = V(\hat{y} + \hat{e})$ → in regression

<u>fact</u>: $V(\hat{y} + \hat{e}) = V(\hat{y}) + V(\hat{e})$

$V(\hat{y}) = r^2 V(y)$ & $V(\hat{e}) = (1-r^2)V(y)$

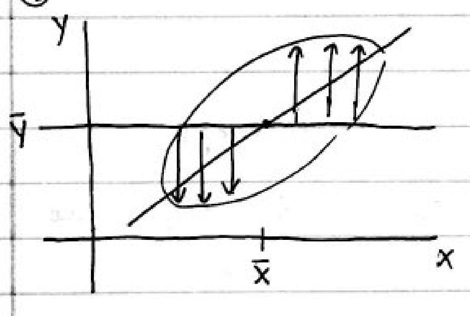
so, $\boxed{r^2 = \frac{V(\hat{y})}{V(y)}}$ = % of variance in y explained by the regression of y on x

- $-1 \le r \le 1$ so $-1 \le r^2 \le 1$
- we want r^2 to be <u>big</u> → strength of correlation b/t x & y

$SD(\hat{e}) = \sqrt{1-r^2} \cdot SD(y)$

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

②

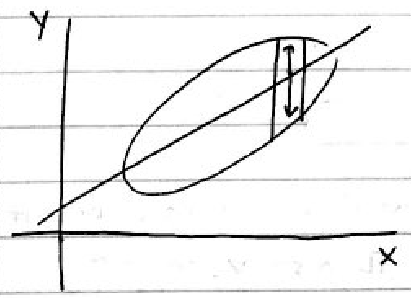


- ignore x & predict y

best prediction:

→ $\hat{y} = \bar{y}$

→ $SD(y) = S_y$

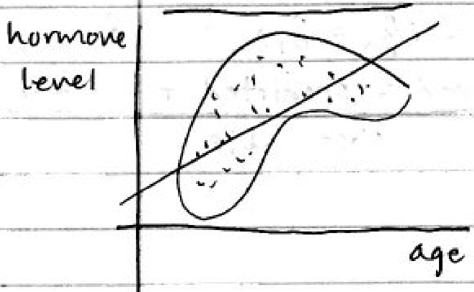


- use x to predict y

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

$\widehat{SE}(\hat{y}) = S_{y|x} = \sigma_{y|x} = S_y\sqrt{1-r^2}$

<u>Model - Checking</u>



* linear regression fails to capture relationship b/t x & y

↳ nonlinearity