

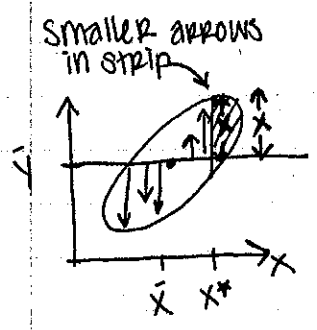
AMS 7
REGRESSION

11/19/09

Lab 6 due 23 Nov, Mon
HW#4 due Tue, 24 Nov

There will be lecture Tue, 24 Nov

Solutions to HW#3, Lab 2, & quizzes 3-6 now in glass case (BE 125)



How accurate are regression predictions?
if $X = X^*$: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X^*$

First guess: $SE(\hat{Y}) = S_Y$, but picture shows this guess is too big

MATH FACT

$S_{Y|X}$ = "root mean square error" (RSE)

$$SE(\hat{Y}) = S_Y \sqrt{1-r^2} \sqrt{\frac{n-1}{n-2}} \quad \left. \vphantom{SE(\hat{Y})} \right\} \text{the larger } n \rightarrow \text{closer to 1}$$

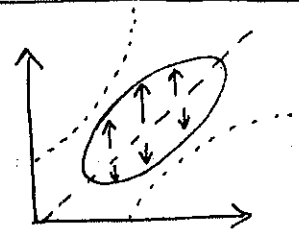
as $r \uparrow 1$ or $\downarrow -1$, $SE(\hat{Y}) \rightarrow 0$
 \Rightarrow the stronger the correlation, the more accurate SE

INFERENCEAL SUMMARY \rightarrow PREDICTIVE SUMMARY

UNKNOWN	Y at $X = X^*$
estimate	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X^*$
give or take	$SE(\hat{Y})$

L-218

- approx $SE(\hat{Y}) = S_{Y|X}$; exact $SE(\hat{Y})$ would follow a hyperbola
- measuring variability in arrows



- tail length / wing length Ex: $S_{y|x} = 0.18 \text{ cm}$, so predictions of tail length from wing length should have a give or take of about 0.18 cm.

(L-256) Q: IS THE REGRESSION MEANINGFUL?

1. r^2 ("RSquare": JMP)

- Recall variance of list of #'s $y = (y_1, y_2, \dots, y_n)$ is square of SD:

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

- in regression $y = \hat{y} + \hat{e}$, so $V(y) = V(\hat{y} + \hat{e})$

MATH FACT

(L-257)

$$\begin{aligned} V(\hat{y}) &= r^2 V(y) \\ \text{AND } V(\hat{e}) &= (1-r^2) V(y) \end{aligned}$$

$$\text{SO } r^2 = \frac{V(\hat{y})}{V(y)} = \begin{array}{l} \% \text{ OF VARIANCE IN } y \\ \text{"EXPLAINED BY"} \\ \text{OR "ASSOCIATED WITH"} \end{array}$$

- $-1 \leq r \leq +1$: $0\% \leq r^2 \leq 100\%$
- want r^2 to be big (close to 1)
- Ex: Bird size $r^2 = 0.76 = 76\%$

2 (a) ignore x (or don't know it): predict y anyway \rightarrow best prediction is $\hat{y} = \bar{y}$; give or take SD of $y = S_y$

(b) use x to predict y ; best est: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \pm \hat{\sigma}_{y|x} = S_{y|x}$

(L-259)

Value of regression is seen by comparing S_y w/ $S_{y|x}$

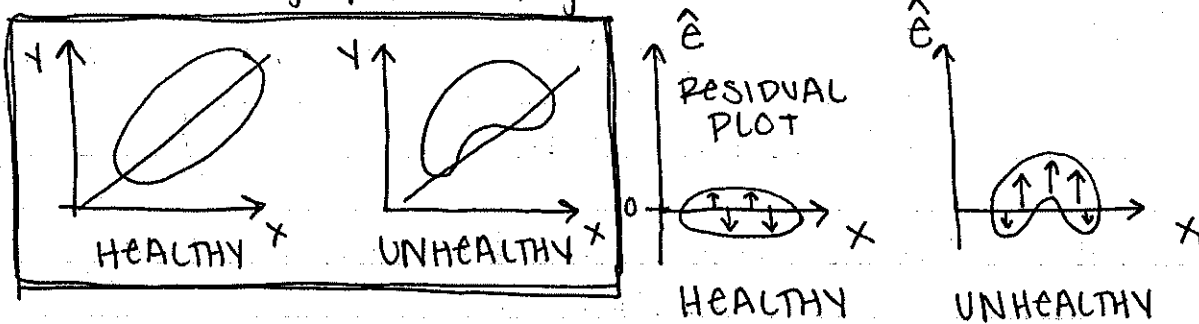
EX: BIRD SIZE: Ignore wing length,
 predicted tail length = $\hat{y} = 7.6\text{cm}$ $\hat{SE} = S_{y|x} = 0.35\text{cm}$

$x = 11.4\text{cm}$ - predict $\hat{y} = 8.1\text{cm}$

$\hat{SE}(\hat{y}) = S_{y|x} = \text{RMSE} = 0.18\text{cm} \Rightarrow$ HALF AS BIG

\Rightarrow this is a highly useful regression

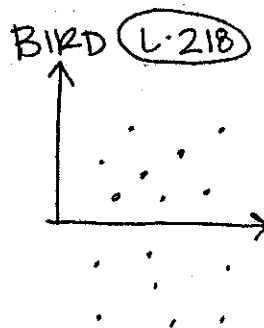
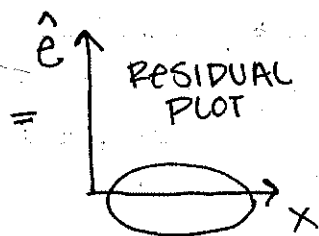
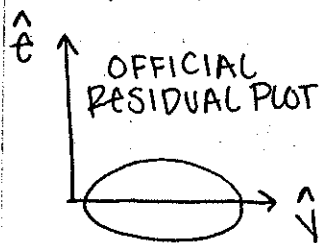
(L-260)



\rightarrow no trend

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

FINAL



= HEALTHY

MULTIPLE LINEAR REGRESSION

\Rightarrow will usually have >1 predictor variable available

$$y_i = [(B_0 + B_1 x_{i1}) + (B_2 x_{i2}) + \dots + (B_k x_{ik})] + e$$

Multiple Linear Regression Model \rightarrow use least squares to estimate $\hat{B}_0, \hat{B}_1, \dots, \hat{B}_k, \hat{\sigma}$

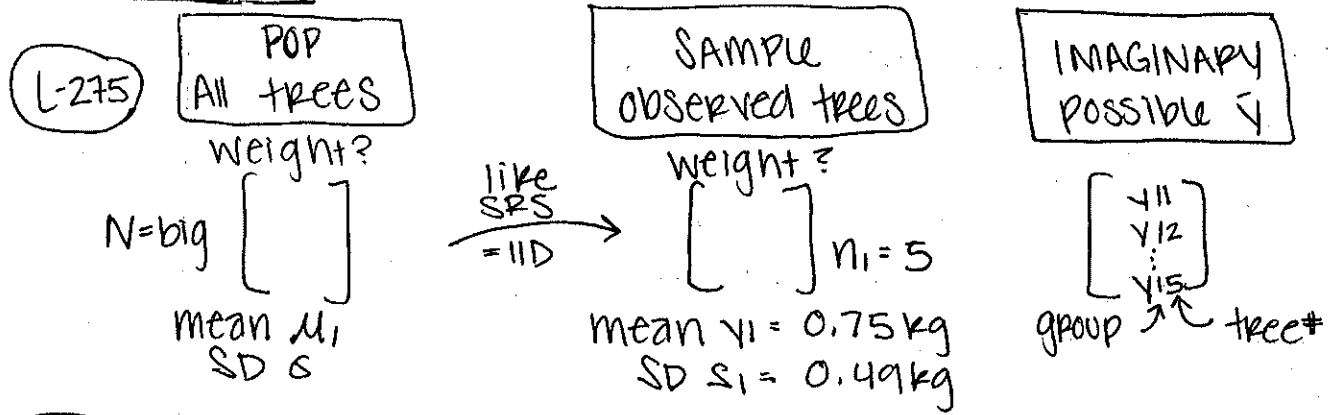
quality of correlation: $R = \text{CORRELATION}(y, \hat{y})$

$R^2 = \text{COEF. OF DETERMINATION}$ ($k=1, r^2 = R^2$)

CHAPTER 7: ONE-WAY ANALYSIS OF VARIANCE

CASE STUDY: GROWTH OF POPLAR TREES

NO TREATMENT



L-271 JMP: look at all parallel conclusions

4 SAMPLE PROBLEM: CONTROL, FERTILIZER, IRRIGATION, BOTH \rightarrow 4 models, compare differences

Group i : $i = 1, \dots, I = 4$ (total # groups) \rightarrow assuming $\text{spl } \sigma = \text{pop } \sigma$

pop mean μ_i ; pop SD σ_i ; spl mean \bar{y}_i ; spl SD s_i ; size n

H_0 : $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma$ all pop σ are the same

H_A : H_0 false: test to see if H_0 true

L-276 Q: Which \bar{y} differ from which others (on avg. by pract sig amount?)?

A1: USE 2-IND. SPL MACHINERY ON ALL PAIRS

- MULTIPLE COMPARISONS PROBLEM \Rightarrow if you make > 1 95% CI, your overall chance of making a mistake $\gg 5\%$

A2: 1-WAY ANALYSIS OF VARIANCE (ANOVA)

H_0 : $\mu_1 = \mu_2 = \dots = \mu_I = \mu$

H_A : H_0 false

Measure difference between (Actual data) vs (Ho Predicted)

WEIGHTED MEAN \Rightarrow

(L-27)

if H_0 TRUE: $\bar{y} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2 + \dots + n_I \bar{y}_I}{n_1 + \dots + n_I}$

GRAND MEAN \Rightarrow mean of all data

$$= \frac{1}{n} \sum_{i=1}^I n_i \bar{y}_i$$

◦ gives more influence to larger samples

Here $n=20$: $= \sum_{i=1}^I n_i \Rightarrow \bar{y} = 0.9115 \text{ kg}$

Natural Measure of the Extent to which \bar{y}_i are NOT close to $\bar{y} =$

$$\sum_{i=1}^I n_i (\bar{y}_i - \bar{y})^2$$

SSB = Sum of Squares between groups

$$= 3.3416455 \text{ kg}^2$$

JMP: "the SS for TREATMENT"

◦ IDEA: FAVOR H_A if SSB IS BIG