This time: regression, ANOVA

Next time: regression, ANOVA

Read: LN pp. L-269 → L-282 Today: LN p. L-248 →

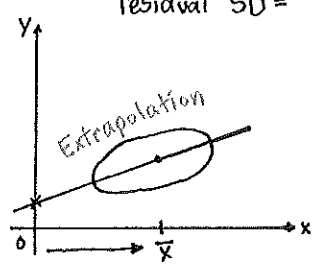
Math fact: ① $E_{IID}(\hat{\beta}_1) = \beta_1$

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

note: $y|x$ ← "given" - how variabled y is after you take x into account

where: $S_{y|x} = \underbrace{S_y \sqrt{1-r^2}}_{\text{interesting part}} \cdot \sqrt{\frac{n-1}{n-2}}$ ← boring (close to 1)

residual SD = "root mean squared error" rmse (JMP uses this name)



- need to extrapolate to create line to find x at zero
- the further away from middle of data, the less secure

so large extrapolation = large uncertainty

Warning: risky to extrapolate regression predictions outside the observed range of x

$y_i = \underbrace{(\beta_0 + \beta_1 x_i)}_{\text{"truth"}} + e_i$ ← from normal curve w/mean 0 & SD $\sigma_{y|x}$

"error" (vary from the mean)

(hoping the values fit a strait line)

y_i (observed y) $= (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ (predicted) $+ \hat{e}_2$ (residual)

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

$\sigma_{y|x}$ represents the typical amount by which you expect y & $\hat{y}$ to differ

2 ways to tell if regression is useful:

$v(y) = s^2_y = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$

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

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

$v(\hat{y}) \doteq r^2 v(y)$

$v(\hat{e}) \doteq (1-r^2) v(y)$

So $r^2 = \frac{v(\hat{y})}{v(y)}$ . .

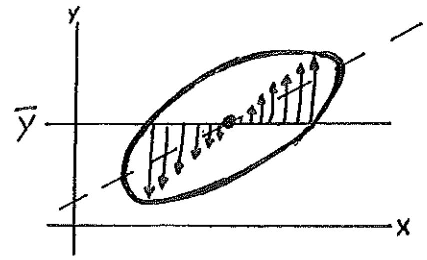
% of variance in y is "associated with" the regression of y on x, called the coefficient of determination.

we want r^2 to be big

case 1: ignoring x (or if you don't have x)

$\hat{y} = \bar{y}$

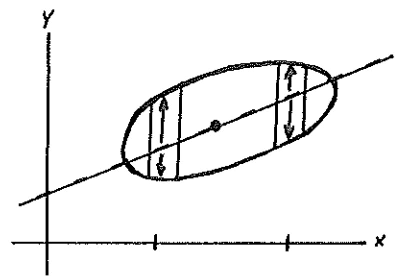
& $SE(\hat{y}) = S_y$



case 2: use x to predict y:

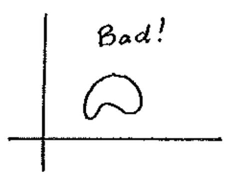
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

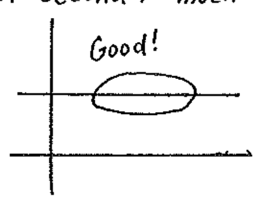
$$\widehat{SE}(\hat{y}) = \underbrace{S_y \cdot \sqrt{1-r^2}}_{\text{residual}}$$



first case looks impressive (can be 76%), but second is much better

Residual plot:





don't want trends

multiple linear regression model: more than 1 variable

Unit 7: One-way Analysis of Variance

For the tree study: weight diff. <u>is</u> practsig!

4 independent variables, need 4 model diagrams instead of 2

the assumption of all equal σ's is bad