AMS7
Lecture 13
7/18/16

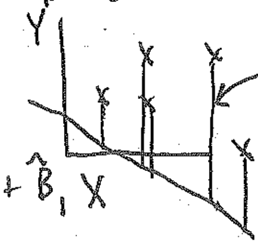
1: reg line for predicting x from y

2: SD line to capture trend

3: reg line for predicting y from x

Galton 1890

How predict y from x?



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

So this is a bad line

A better line →

$$\sum_{i=1}^{n} [y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{\hat{y}_i}]$$

← least squares line

math fact: $\boxed{\text{regression line} = \text{least squares line}}$

ex. tail + wing length example cont. L-248

Inf. Summary

unknown	β_1 = pop slope for predicting TL from WL
estimate	$\hat{\beta}_1 = .77$ cm TL / cm WL

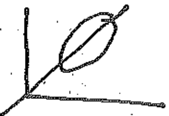
math facts: 1. $E_{IID}(\hat{\beta}_1) = \beta_1$

2. $\widehat{SE}_{IID}(\hat{\beta}_1) = \dfrac{S_{y|x}}{S_x\sqrt{n-1}}$ given, where $S_{y|x} = S_y\sqrt{1-r^2} \cdot \sqrt{\dfrac{n-1}{n-2}}$

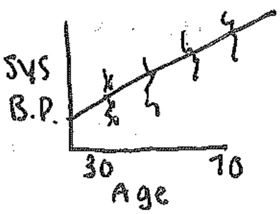
residual SD = "root mean squared error" RMSE

$= \dfrac{S_y\sqrt{1-r^2}}{S_x\sqrt{n-2}}$

To judge slope $\hat{\beta}_1$ is large in practical terms, use same reasoning as w/ sample corr. r

L-253 Biological Meaning of Y-int.  $\hat{r} = -.67, .77$

For WL | TL, it makes sense that as WL → 0 TL → 0.

L-254 

unobservable

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

obs = truth + error

observable

$y_i = (\hat{\beta}_0 + \beta_1 x_i) + \hat{e}_i$

obs y = predicted y + residual

We define $\hat{e}_i$ as $y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ & = obs minus predicted

σ = residual SD

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

~~$\hat{\sigma}_{y|x}$ Represents the typical amt~~

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

Is the reg. practically useful?

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

"associated with" is better

fact: $v(\hat{e}) = (1-r^2) v(y)$

$r^2 = \frac{v(\hat{y})}{v(y)}$ = % of variance in y ~~"explained by"~~ reg. of y on x

r² is called the coefficient of determination

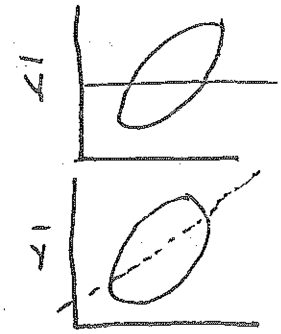
[0 - 1]

For complete formula list + deriving equations: See L notes + the reader! :)

How useful is the regression?

① Predict y ignoring x

$\hat{y}_{no\,x} = \bar{y}$



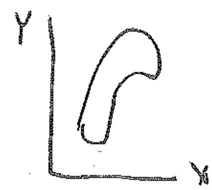
② Predict y using x

$\hat{y}_{no\,x} = \hat{\beta}_0 + \hat{\beta}_1 x$

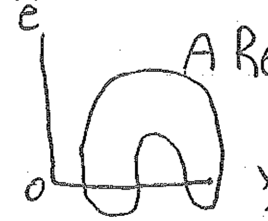
w/ $\hat{SE}(\bar{y}_{use\,x}) = s_y\sqrt{1-r^2}$

this is smaller than $SE(\hat{y}_{no\,x})$

Residual Plots



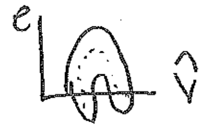
Residuals



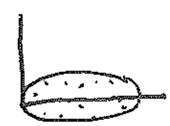
A Residual Plot

$\hat{\beta}_0 + \hat{\beta}_1 X$

Official Res Plot



unhealthy : non linear curvature



healthy: no trend/ pattern

one column: univariate sample

2 or more : multivariate sample

$y_i = \beta_0 + \beta_1 x_i + e_i$ simple linear reg, 1 X

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ ... multiple linear reg

K>1 x variables

Can generalize least squares to get estimate

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots\ldots + \hat{\beta}_k x_{i\cdot k}$

Q: Is the mult. reg useful?
A: R^2 = multiple R^2 = coefficient of det.
↑ want big

Section 7: One-Way Analysis of Variance

L-270 case study: Trees w/ 4 treatment groups
models for each treatment

$$\left.\begin{array}{l} \mu_1, \mu_2 \ldots \mu_4 \\ \sigma \ldots \\ \bar{y} \ldots \\ SDs \ldots \end{array}\right\} 4$$

remember basic model assumes all $\sigma_{1-4} = \sigma$
(I) how many $I = 4$