this time: regression

↑

next time: ANOVA

read: LN pp. L-269-289

AMS 7
18 Jul 2016

1

today: LN pp. L-247 →

homework 4 due Fri 22 Jul in class

take-home final due Mon 25 Jul noon

extra office hours:

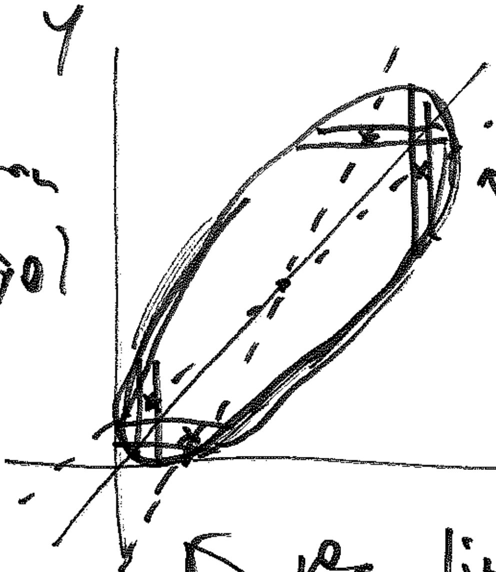
Sat	DD	23 Jul	noon-1pm
Sun	DD	24 Jul	noon-1pm
Sat	GD	23 Jul	2-3p
Sun	GD	24 Jul	2-3p

Mon 25 Jun 9A-10A

take elevator to 3rd floor Baskin: door propped open

Y

Galton (~1890)



X

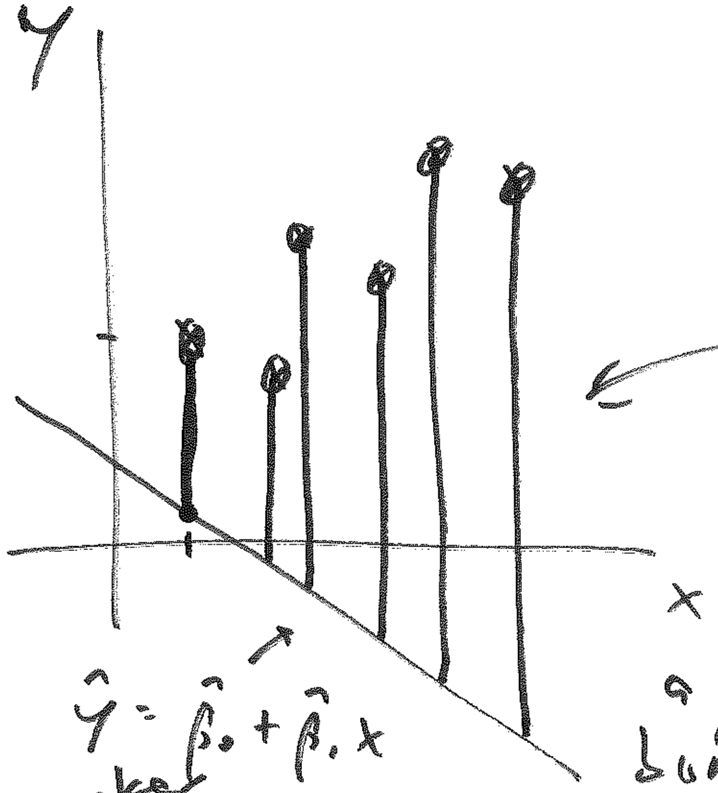
SD line: best for capturing the trend of (x, y) pairs

regression line for predicting y from x

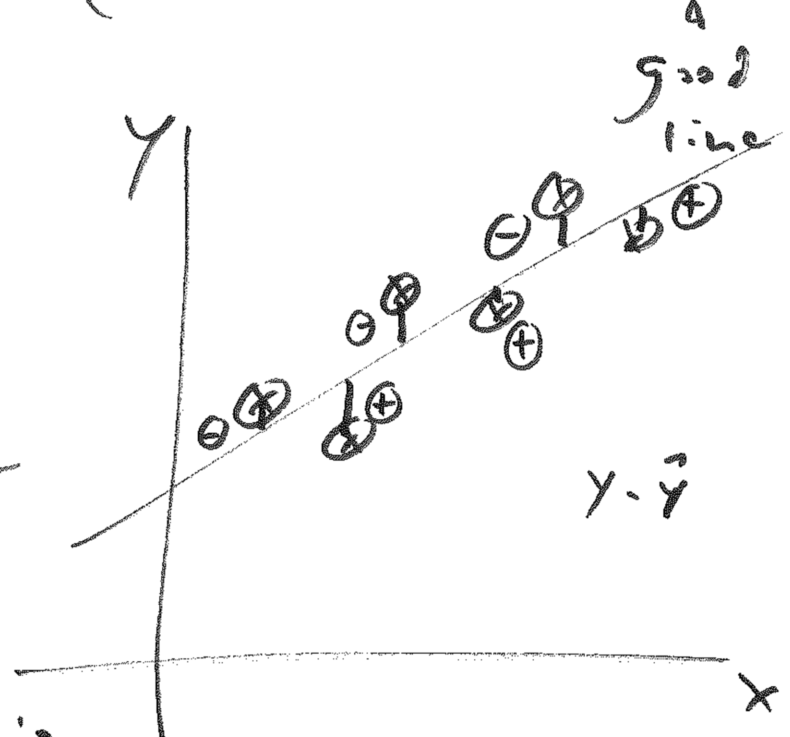
← reg. line for pred. x from y

Goal: how predict y from x?

(residuals)



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



find $(\hat{\beta}_0, \hat{\beta}_1)$ to minimize

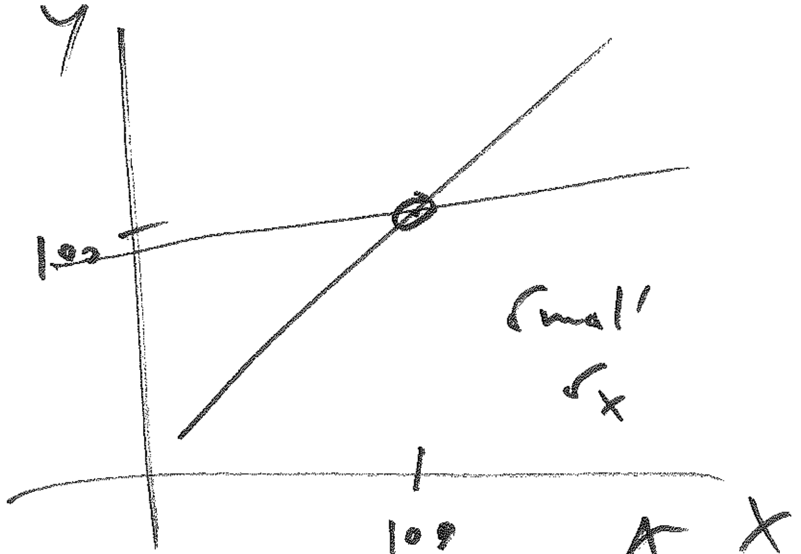
$$\sum_{i=1}^{n} \Big[y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{\hat{y}_i} \Big]^2$$

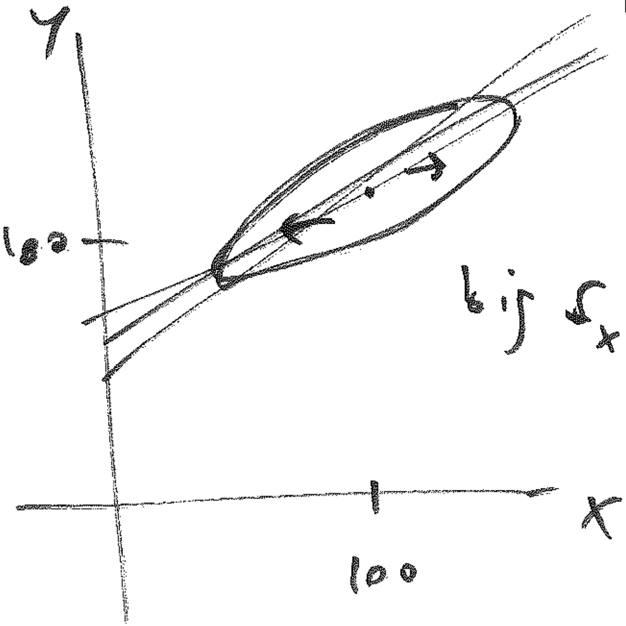
$$\left(\begin{array}{c|c} y_1 & x_1 \\ \vdots & \\ y_n & x_n \end{array} \right)$$

least squares line

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

math fact: regression line = least squares line





harder to estimate β_1 here

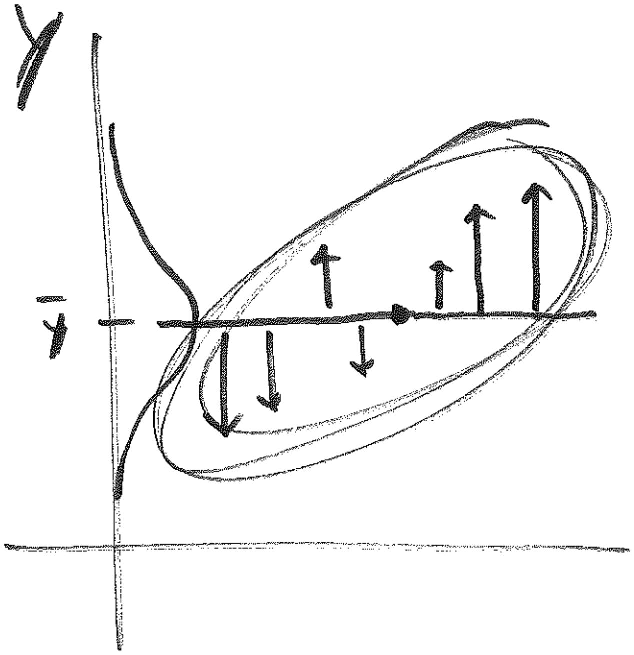
---

to judge whether a sample slope $\hat{\beta}_1 = r \frac{s_y}{s_x}$ is large in practical terms, use exactly the same line of reasoning as with the sample correlation r

(10.00)

(10.27)

how useful is the regression?



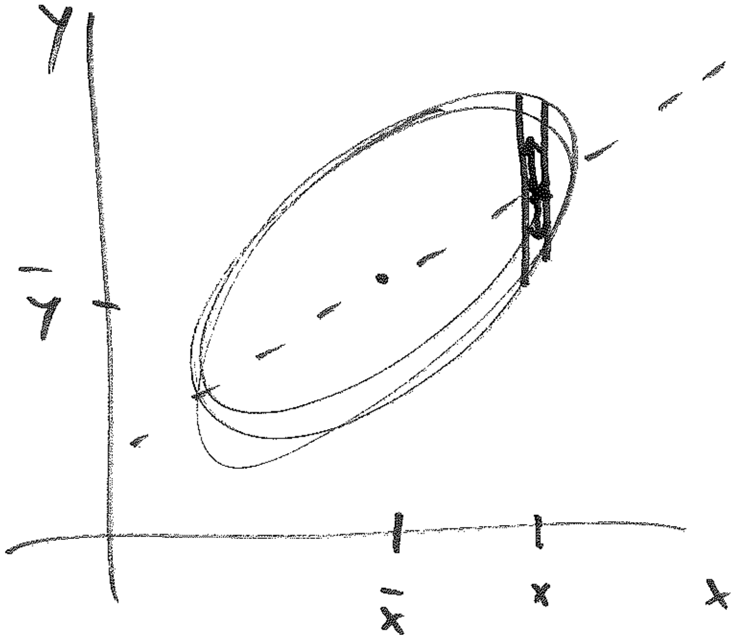
$$\left(\begin{array}{c|c} y_1 & x_1 \\ \vdots & \vdots \\ y_n & x_n \end{array}\right)$$

predictive task

1. ignore x or don't measure it; predict y:

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

with $\widehat{SE}(\hat{y}_{no\,x}) = s_y$



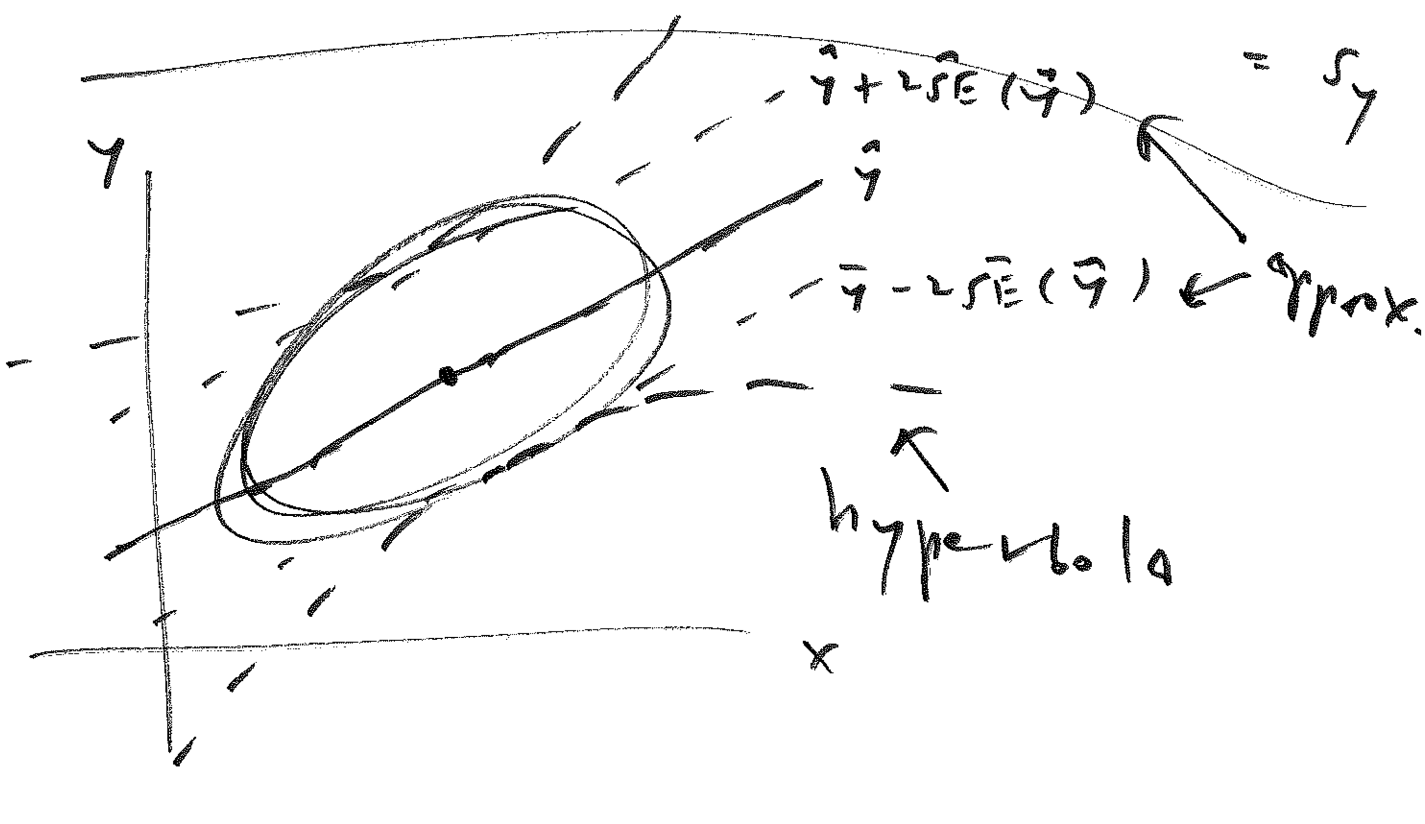
predictive task ②:

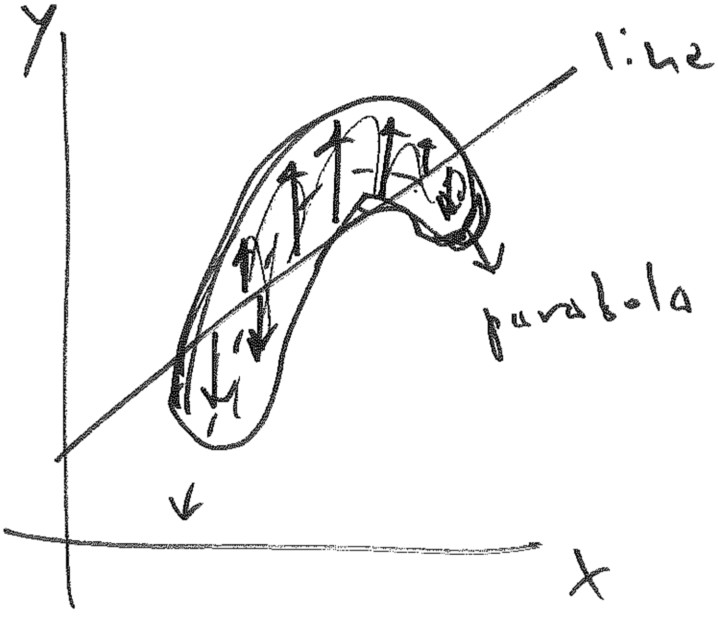
use x to predict y:

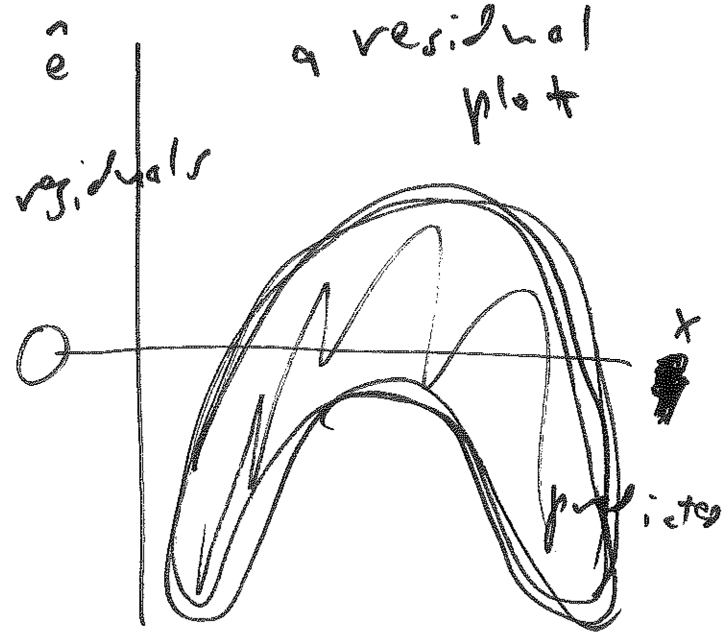
$$\boxed{\hat{y}_{\text{use } x} = \hat{\beta}_0 + \hat{\beta}_1 x}$$

with $\hat{SE}(\hat{y}_{\text{use } x}) \doteq s_y \sqrt{1 - r^2}$

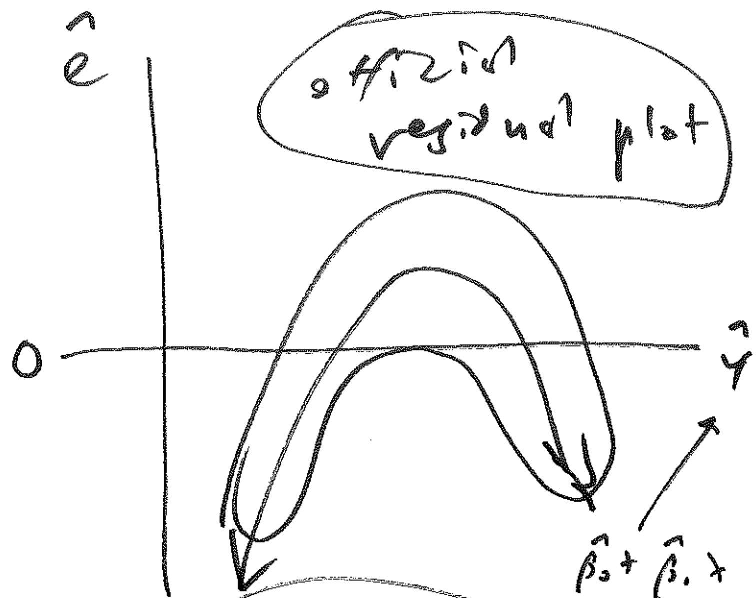
This is smaller than $\hat{SE}(\hat{y}_{\text{no } x}) = s_y$



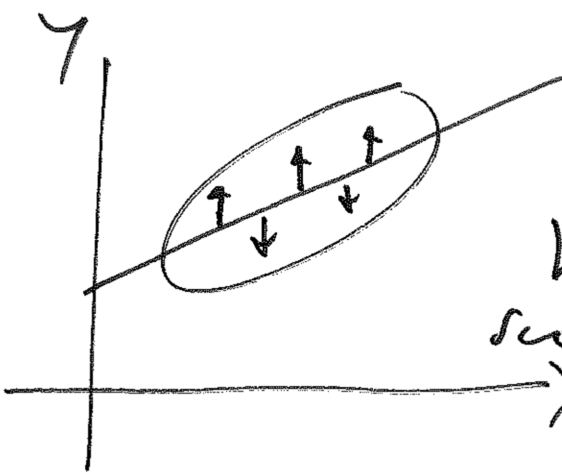




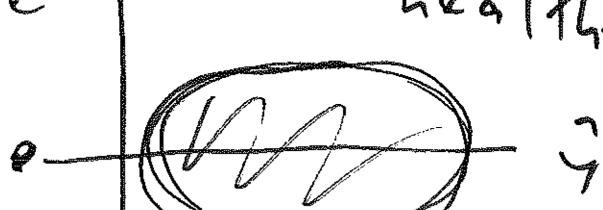
$$\begin{bmatrix} y & x & \hat{y} & \hat{e} \end{bmatrix}$$



unhealthy: nonlinearity (curvature)



healthy scatterplot



healthy residual plot: no trend or pattern

(7)

y = height

x = trunk diameter

only 1 x

$Y_i = \beta_0 + \beta_1 x_i + e_i$

simple linear regression

---

y = height

x_1 = trunk diameter

$\vdots$

$x_4 = x_k$ = volume

multiple linear regression

$k > 1$ x variables

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + e_i$$

---

can generalize least squares to get estimates:

$$\boxed{\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik}}$$

$$\left(\begin{array}{c|cccc} Y_1 & X_{11} & X_{12} & \cdots & X_{1k} \\ Y_2 & X_{21} & X_{22} & \cdots & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ Y_n & X_{n1} & X_{n2} & & X_{nk} \end{array}\right)\left(\begin{array}{c} \hat{Y}_1 \\ \vdots \\ \vdots \\ \hat{Y}_n \end{array}\right)$$

if reg good, Y & $\hat{Y}$ should be close

$\boxed{9}$ is the multiple regression useful?

$\boxed{A_1}$ could compute correlation between Y & $\hat{Y}$ = R ; square to get

R^2 = multiple R^2 = coefficient of determination

↑ want big