Stat 88 lec 36

warm up 7:00-7:10

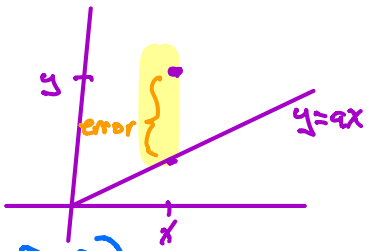
3. Sometimes data scientists want to fit a linear model that has no intercept term. For example, this might be the case when the data are from a scientific experiement in which the attribute X can have values near 0 and there is a physical reason why the response Y must be 0 when $X = 0$.

So let (X, Y) be a random pair and suppose you want to predict Y by an estimator of the form aX for some a. Find the least squares predictor $\hat{Y}$ among all predictors of this form.

$$MSE(a) = E((Y-aX)^2)$$

Solve

$\frac{d}{da} MSE(a) = 0$ for a.



$$E((Y-aX)^2) = E(Y^2 - 2aXY + a^2X^2)$$

$$= E(Y^2) - 2aE(XY) + a^2E(X^2)$$

$\frac{d}{da} MSE(a) = 0 \Rightarrow$

$$0 - 2E(XY) + 2aE(X^2) = 0$$

Solve for a

$$\hat{a} = \frac{E(XY)}{E(X^2)}$$

regression line with x-intercept = 0 is

$$\hat{y} = \frac{E(XY)}{E(X^2)} x$$

last time sec 11.3 least squares regression

let (X,Y) be a random pair

ex X = # cookies Homer eats

Y = length of Homer's naps

$\mu_X = E(X) = 11.5$ $\sigma_X = SD(X) = 2.5$ $r = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y} = .5$ (D_X, D_Y)

$\mu_Y = E(Y) = 30$ $\sigma_Y = SD(Y) = 8$

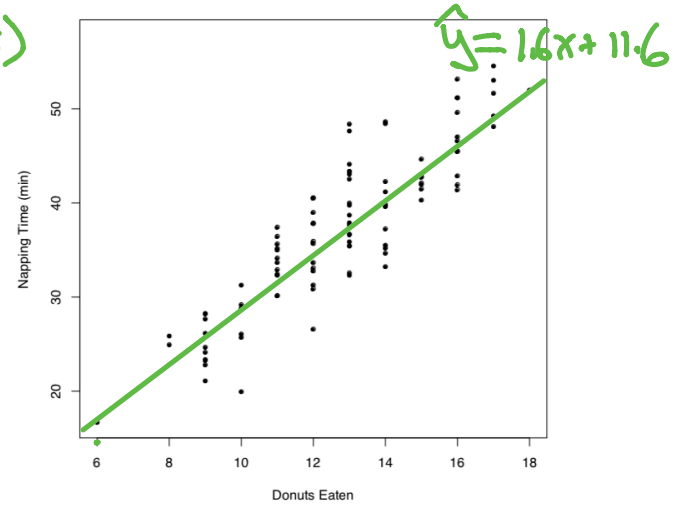
We wish to find the "best" fitting line $\hat{y} = \hat{a}x + \hat{b}$, through the scatter plot of all (X,Y) pairs.

We showed that

$\hat{a} = r\frac{\sigma_y}{\sigma_x} = .5\frac{8}{2.5} = 1.6$

$\hat{b} = \mu_y - \hat{a}\mu_x = 30 - 1.6(11.5) = 11.6$

$\hat{y} = 1.6x + 11.6$



Today

1. sec 11.4 Properties of correlation

① Sec 11.4 Bounds on correlation

$$r = r(X,Y) = \frac{E(D_x D_y)}{\sigma_x \sigma_y} = E\left(\frac{X-\mu_x}{\sigma_x}\,\frac{Y-\mu_y}{\sigma_y}\right)$$

($X-\mu_x$, $Y-\mu_y$)

$$= E(X^* Y^*)$$

is unitless.

We show $-1 \le r \le 1$

What is $E(X^*)$? 0

$X^* = \frac{X-\mu_x}{\sigma_x}$

$E(X^*) = E\left(\frac{X}{\sigma_x}\right) - \frac{\mu_x}{\sigma_x}$

$= \frac{E(X)}{\sigma_x} - \frac{\mu_x}{\sigma_x} = 0$

$\mathrm{var}(X^*) = E(X^{*2}) - (E(X^*))^2$

$\mathrm{var}(X^*)$? 1

$\mathrm{var}(X^*) = \mathrm{var}\left(\frac{X-\mu_x}{\sigma_x}\right)$

$= \frac{1}{\sigma_x^2}\mathrm{var}(X-\mu_x) = \frac{1}{\sigma_x^2}\cdot\mathrm{var}(X) = 1$

$E((X^*)^2)$? $E((X^*)^2) = \mathrm{var}(X^*) + (E(X^*))^2 = 1$ ($\mathrm{var}(X^*) = 1$, $(E(X^*))^2 = 0$)

lower bound

We will show that $r = E(X^* Y^*) \ge -1$

FOIL $E((X^* + Y^*)^2) \ge 0$

$E((X^*)^2 + 2(X^*Y^*) + (Y^*)^2) \ge 0$

$\Rightarrow E((X^*)^2) + 2E(X^*Y^*) + E((Y^*)^2) \ge 0$

$(X^*)^2 = 1$, $(Y^*)^2 = 1$

$\Rightarrow 2 + 2E(X^*Y^*) \ge 0 \Rightarrow E(X^*Y^*) \ge -1$

Upper bound

Similarly,

$$E\left((X^* - Y^*)^2\right) \geq 0 \Rightarrow r = E(X^* Y^*) \leq 1$$

So $\boxed{-1 \leq r \leq 1}$

Other properties of correlation

a) $R(X,Y) = R(Y,X)$

Pf a) $E(X^* Y^*) = E(Y^* X^*)$ ✓

b) $r(aX+b, cY+d) = \begin{cases} r(X,Y) & \text{if } ac > 0 \\ -r(X,Y) & \text{if } ac < 0 \end{cases}$

$a \neq 0$

$c \neq 0$

Pf b)

Put $aX+b$ in std units

$E(aX+b) = a\,E(X) + b$

$SD(aX+b) = |a|\,SD(X)$

$aE(X)+b$

so $(aX+b)^* = \dfrac{aX+b - E(aX+b)}{SD(aX+b)}$ $= |a|SD(X)$

$= \dfrac{a(X-E(X))}{|a|\,SD(X)} = \begin{cases} X^* & \text{if } a>0 \\ -X^* & \text{if } a<0 \end{cases}$ X^*

Similarly $(cY+d)^* = \begin{cases} Y^* & \text{if } c>0 \\ -Y^* & \text{if } c<0 \end{cases}$

so

$r(aX+b, cY+d) = E\left((aX+b)^*(cY+d)^*\right)$

$E(X) = \mu_X$

$E(aX+b) = \mu_{aX+b}$

$= \begin{cases} E(X^*Y^*) & \text{if } ac>0 \\ -E(X^*Y^*) & \text{if } ac<0 \end{cases}$ $r(X,Y)$

$-r(X,Y)$

☺

exercise 11.6.7

7. Let (X, Y) be a random pair and let $r = r(X, Y)$. Let X^* be X in standard units and let Y^* be Y in standard units.

a) Find $r(X^*, Y^*)$.

$= r(X,Y)$

since $ac>0$

$X^* = \dfrac{X-E(X)}{SD(X)} = \dfrac{1}{SD(X)}X - \dfrac{E(X)}{SD(X)}$ $a>0$, b

$Y^* = \dfrac{1}{SD(Y)}Y - \dfrac{E(Y)}{SD(Y)}$ $c>0$, d

The regression line for X and Y in std units

we know $r(x,y) = r(x^*, y^*)$

so if a random pair (X,Y) has regression line

$$\hat{y} = \underbrace{r\frac{\sigma_y}{\sigma_x}}_{\hat{a}} x + \underbrace{\mu_y - r\frac{\sigma_y}{\sigma_x}\mu_x}_{\hat{b}}$$

then the random pair (x^*, y^*) has

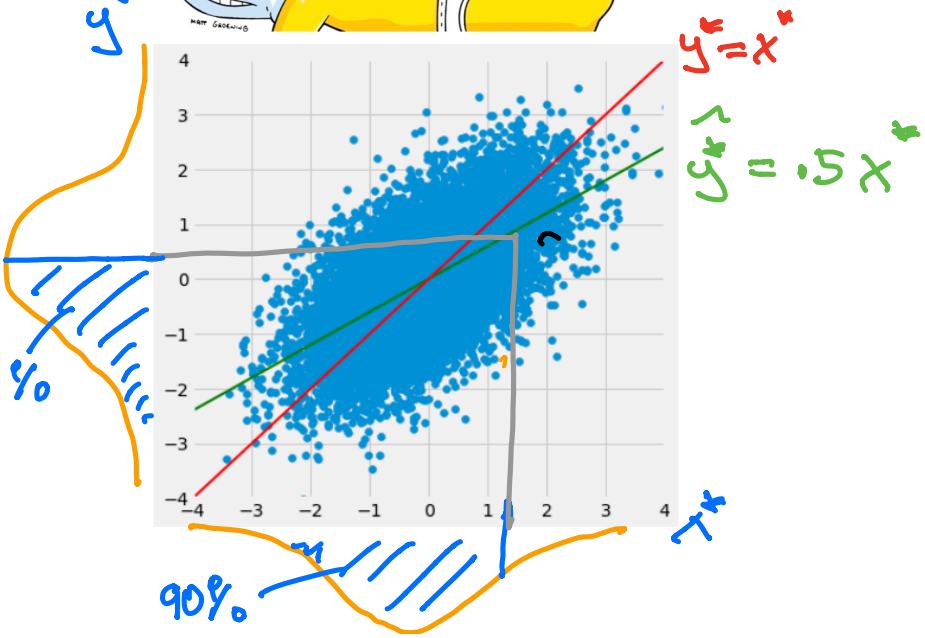
regression line? $\hat{y}^* = r(x^*,y^*)\frac{\sigma_{y^*}}{\sigma_{x^*}} x^* + \mu_{y^*} - r(x^*,y^*)\frac{\sigma_{y^*}}{\sigma_{x^*}}\mu_{x^*}$

$E(y^*)$

$r = r(x^*,y^*)$, $\frac{\sigma_{y^*}}{\sigma_{x^*}} = 1$, $\mu_{y^*} = 0$, $\mu_{x^*} = 0$

$$\hat{y}^* = r x^*$$

what is the regression line for Homer in std units?



Assume X and Y are approximately normal.

If X is 90th percentile, estimate the percentile rank of Y.

Poll 4

$x^* = \Phi^{-1}(.90) = 1.28$

$\hat{y}^* = (.5)(1.28) = .64$ ← percentile of y in std units

Percentile rank $= \Phi(.64) = 74\%$ ← regression to the 50 percentile (mean).

If you wish to predict y from x the regression line is $\hat{y}^* = r(x^*, y^*)\, x^* = r\, x^*$

If you wish to predict x^* from y^* what is the regression line?

$\hat{x}^* = r(y^*, x^*)\, y^* = r(x^*, y^*)\, y^*$

↑ dependent, independent var

ex

Suppose y is 74th percentile for Y. Estimate the percentile rank of X.

$\Phi^{-1}(.74) = .64$

$x^* = .5(.64) = .32$

Percentile rank $= \Phi(.32) = 63\%$

↖ regression to mean.