

Stat 88    lec 3c

Wann v. 2:00 - 2:10

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 experiment 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

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

$$\begin{aligned} E((Y - aX)^2) &= E(Y^2 - 2aXY + a^2X^2) \\ &= E(Y^2) - 2aE(XY) + a^2E(X^2) \end{aligned}$$

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

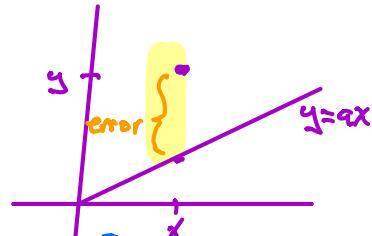
$$0 - 2E(XY) + 2aE(X^2) = 0$$

Solve for  $a$

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

regression line with  $x\text{-intercept} = 0$   $\rightarrow$

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



## Last time sec 11.3 least squares regression

Let  $(x, y)$  be a random pair

let  $x = \# \text{ cookies Homer eats}$

$y = \text{length of Homer's nap}$

$$\mu_x = E(x) = 11.5 \quad \sigma_x = SD(x) = 2.5 \quad r = \frac{E((x - \mu_x)(y - \mu_y))}{\sigma_x \sigma_y} = .5$$

$$\mu_y = E(y) = 30 \quad \sigma_y = SD(y) = 8$$

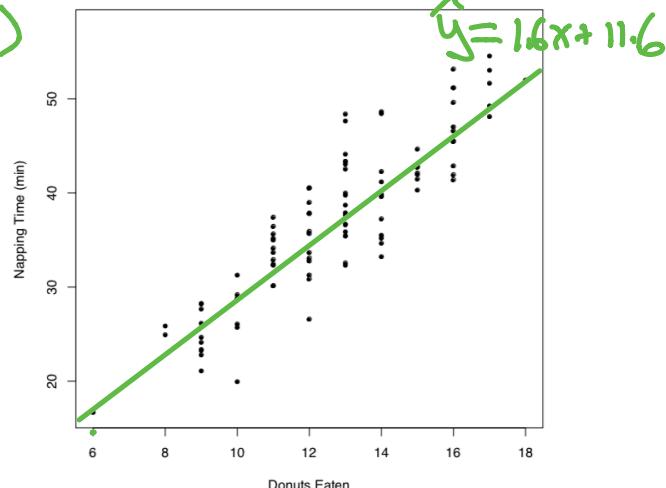
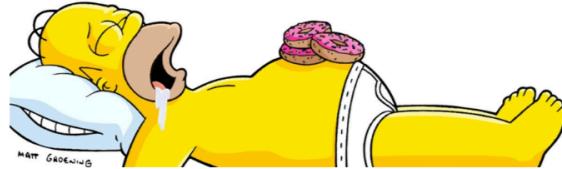
We wish to find the "best" fitting line  $\hat{y} = \hat{a}x + \hat{b}$ , through the scatter plot of all  $(x, y)$  pairs.

We showed that

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

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

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



Today

① sec 11.4 Properties of correlation

① sec 11.4 Bounds on correlations

$$r = r(x, y) = \frac{E(D_x D_y)}{\sigma_x \sigma_y} = E\left(\frac{x - \mu_x}{\sigma_x} \frac{y - \mu_y}{\sigma_y}\right) = E(x^* y^*)$$

is unital,

we show  $-1 \leq r \leq 1$

What is  $E(x^*)$ ?  $\boxed{0}$   
 $\text{var}(x^*) = E(x^{*2}) - (E(x^*))^2$   
 $\text{var}(x^*)?$   $\boxed{1}$

$$\begin{aligned} x^* &= \frac{x - \mu_x}{\sigma_x} \\ E(x^*) &= E\left(\frac{x}{\sigma_x}\right) - \frac{\mu_x}{\sigma_x} \\ &= \frac{E(x) - \mu_x}{\sigma_x} = \boxed{0} \end{aligned}$$

$$\begin{aligned} E(x^{*2}) &= \text{var}(x^*) + (E(x^*))^2 \\ &\stackrel{?}{=} \text{var}(x^*) + \boxed{1} \end{aligned}$$

$$\begin{aligned} \text{var}(x^*) &= \text{var}\left(\frac{x - \mu_x}{\sigma_x}\right) \\ &= \frac{1}{\sigma_x^2} \text{var}(x - \mu_x) = \frac{1}{\sigma_x^2} \cdot \text{var}(x) = 1 \end{aligned}$$

lower bound

we will show that  $r = E(x^* y^*) \geq -1$

FOIL  $E((x^* + y^*)^2) \geq 0$

$$E(x^{*2}) + 2(x^* y^*) + (y^{*2}) \geq 0$$

$$\begin{aligned} \Rightarrow E(x^{*2}) + 2E(x^* y^*) + E(y^{*2}) &\geq 0 \\ \stackrel{?}{=} 1 &\quad \stackrel{?}{=} 1 \end{aligned}$$

$$\Rightarrow 1 + 2E(x^* y^*) \geq 0 \Rightarrow E(x^* y^*) \geq -1$$

## Upper bound

Similarly,

$$E((x^* - y^*)^2) \geq 0 \Rightarrow r = E(x^*y^*) \leq 1$$

so  $-1 \leq r \leq 1$

## Other properties of correlation

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

Pf a)  $E(x^*y^*) = E(y^*x^*) \quad \checkmark$

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

$$E(ax+b) = a E(x) + b$$

$$SD(ax+b) = |a| SD(x)$$

$$\text{so } (ax+b)^* = \frac{ax+b - E(ax+b)}{SD(ax+b)} \stackrel{\text{"}}{=} \frac{ax^* + b - E(ax^*)}{|a| SD(x^*)} \stackrel{\text{"}}{=} \begin{cases} x^* & \text{if } a > 0 \\ -x^* & \text{if } a < 0 \end{cases}$$

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

$$r(ax+b, cy+d) = E((ax+b)^*(cy+d)^*)$$

$$E(x) = \mu_x$$

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

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

"r(x,y)"

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

$$= \boxed{r(x,y)}$$

Since  $ac > 0$

$$x^* = \frac{x - E(x)}{SD(x)} = \frac{1}{SD(x)} x - \frac{E(x)}{SD(x)}$$

$$y^* = \frac{y - E(y)}{SD(y)} = \frac{1}{SD(y)} y - \frac{E(y)}{SD(y)}$$

## 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

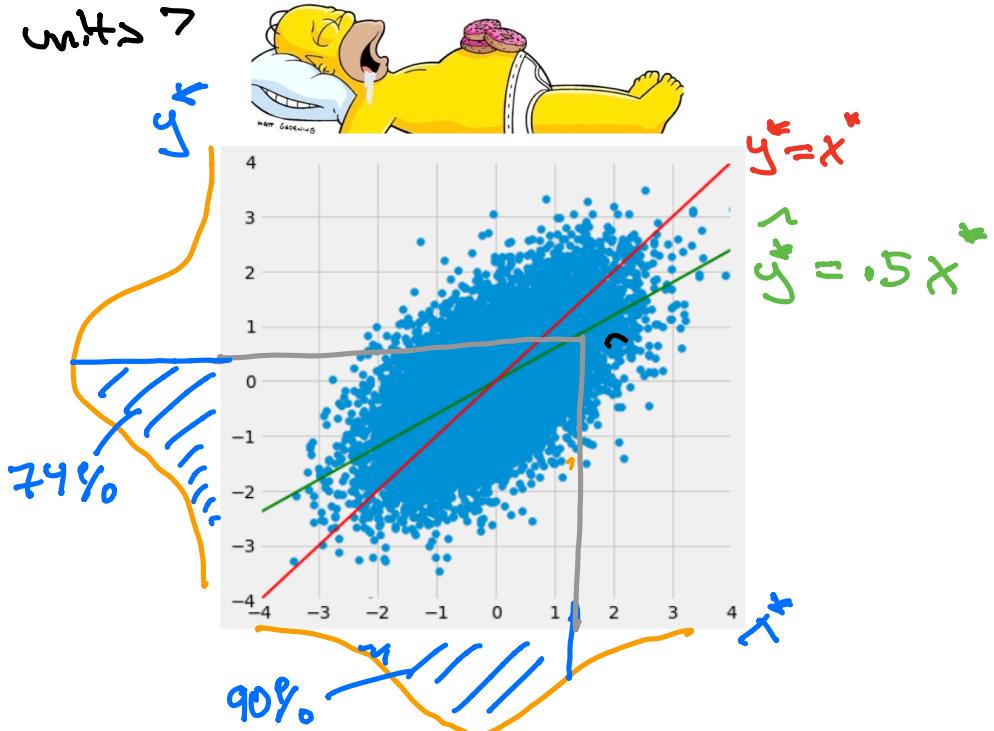
$$\hat{y} = \underbrace{r \frac{\sigma_y}{\sigma_x} x + \underbrace{M_y - r \frac{\sigma_y}{\sigma_x} M_x}_{\hat{b}}}_{\hat{a}}$$

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

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

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

what is the regression line for Homer in std units?



Assume  $X$  and  $Y$  are approximately normal.

If  $X$  is 90<sup>th</sup> percentile, estimate the percentile rank of  $Y$ .

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

$$\hat{y}^* = (.5)(1.28) = .64 \leftarrow \begin{array}{l} \text{percentile of } Y \\ \text{in std units} \end{array}$$

$$\text{Percentile rank} = \Phi(.64) = 74\% \leftarrow \begin{array}{l} \text{regression to the} \\ \text{50 percentile} \\ \text{(mean).} \end{array}$$

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's the regression line?

$$x^* = r(y^*, x^*) y^* = r(x^*, y^*) y^* \leftarrow \begin{array}{l} \text{dependent} \\ \text{independent var} \end{array}$$

ex

Suppose  $Y$  is 74<sup>th</sup> percentile for  $Y$ . Estimate the percentile rank of  $X$ ,

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

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

$$\text{Percentile rank} \approx \Phi(.32) = 63\% \leftarrow \begin{array}{l} \text{regression} \\ \text{to mean.} \end{array}$$

② Sec 11.5 The error in regression

The error in the regression estimate is called the **residual** and is defined as

$$D = y - \hat{y}$$

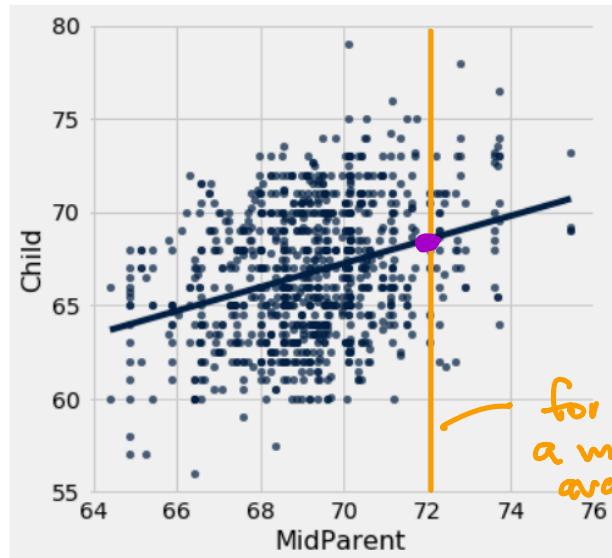
It is useful to write this in terms of the deviations  $D_x = x - \mu_x$  and  $D_y = y - \mu_y$

$$\begin{aligned}\hat{y} &= \hat{\alpha}x + \hat{b} = \hat{\alpha}x + \mu_y - \hat{\alpha}\mu_x \\ &= \hat{\alpha}(x - \mu_x) + \mu_y\end{aligned}$$

$$\begin{aligned}D &= y - \hat{y} = y - [\hat{\alpha}(x - \mu_x) + \mu_y] \\ &= y - \mu_y - \hat{\alpha}(x - \mu_x) \\ &= D_y - \hat{\alpha} D_x\end{aligned}$$

What is  $E(D)$ ?

This says for a fixed  $x = x$ , the average deviation of  $y$  from  $\hat{y}$  is 0.



for those parents with a midparent ht = 72, the avg child ht is the pt on the regression line at  $x=72$ ,

what is  $E(D_x^2)$ ?

$E(D_X D_Y)$ ?

Let's find  $\text{var}(D)$

$$\begin{aligned}
 \text{Var}(D) &= E(D^2) \\
 &= E(D_Y^2) - 2\hat{a}E(D_X D_Y) + \hat{a}^2 E(D_X^2) \\
 &= \sigma_Y^2 - 2r \frac{\sigma_Y}{\sigma_X} r \sigma_X \sigma_Y + r^2 \frac{\sigma_Y^2}{\sigma_X^2} \sigma_X^2 \\
 &= \sigma_Y^2 - 2r^2 \sigma_Y^2 + r^2 \sigma_Y^2 \\
 &= \sigma_Y^2 - r^2 \sigma_Y^2 \\
 &= (1 - r^2) \sigma_Y^2
 \end{aligned}$$

$$\Rightarrow \text{SD}(D) = \sqrt{1-r^2} S_Y$$

The residuals are normally distributed

so

$$D \sim N(0, (1-r^2)\sigma_y^2)$$

Picture

