


read: FPP ch 10, 11  
(p. 22) VIII.B in reader

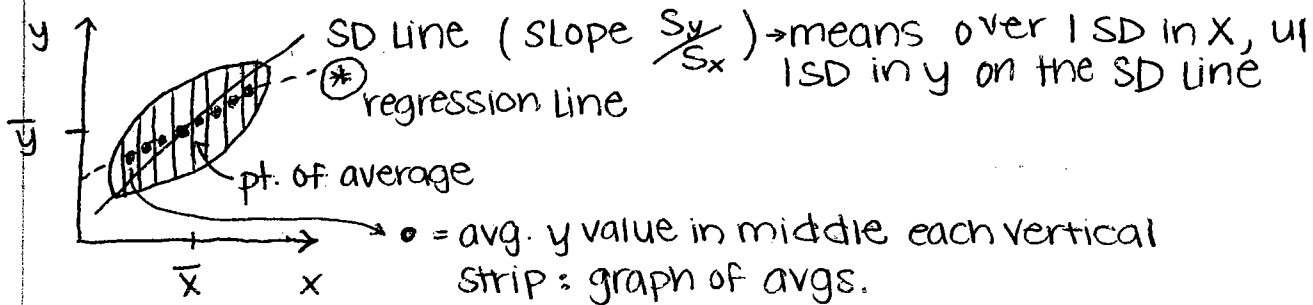
26 May 2005

① (A) (B)

### Case Study 18 cont.

height distribution =  symmetric

weight distribution =  long right tail



Galton: find the line that is a smooth version of the graph of averages

$\Rightarrow$  regression line (\* (----))  $\Rightarrow$  slope =  $r \frac{S_y}{S_x}$

equation of regression line =  $\hat{y} = \hat{a} + \hat{b}x$

$\hat{y}$ : predicted y-value  
 $\hat{a}$ : y-intercept  
 $\hat{b}$ : slope

2 ways to make regression predictions:

- 1) A certain guy is 70 1/2 inches tall  $\rightarrow$  (mean 68 in., SD 2 1/2 in.)  
so in S.H. he's  $\frac{70\frac{1}{2} - 68}{2\frac{1}{2}} = +1 = 1$  SD above avg. in x  
 $\rightarrow$  We predict he will be  $r \cdot 1 = +0.36$  SD's above avg. in y  $\rightarrow$  (mean 158 lbs., SD 25 lbs.) y  $\rightarrow$   $(.36)(25 \text{ lbs.}) = 9 \text{ lbs. above avg.}$   $\rightarrow$  167 lbs

Therefore I predict that a guy that is 70 1/2 in. tall will weigh about 167 lbs.

\* SD Line is always steeper than the regression line

26 May 2005

(2)

2) work out slope and intercept of regression line and plug  $x$  into the equation slope  $\hat{b} = r^{sy}/s_x$   $\hat{a} = ?$

Fact: reg. line goes through point of averages

$$(\bar{x}, \bar{y}) : \hat{y} = \hat{a} + \hat{b}x \text{ (in general)}$$

$$\bar{y} = \hat{a} + \hat{b}\bar{x} \text{ (passing through } (\bar{x}, \bar{y}))$$

$\rightarrow \hat{a} = \bar{y} - \hat{b}\bar{x}$ ; for predicting  $y$  from  $x$ , best slope is  $\hat{b} = r^{sy}/s_x$ , best  $y$ -intercept is  $\hat{a} = \bar{y} - \hat{b}\bar{x}$

equation of reg. line in words:

$$\begin{pmatrix} \text{predicted} \\ \text{y-value} \end{pmatrix} = (\text{intercept}) + (\text{slope})(x\text{-value})$$

So in case study 18,  $x = 70.5$  in

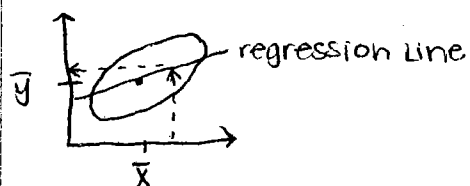
$$\hat{b} = r^{sy}/s_x = 0.36 \left( \frac{25 \text{ lbs}}{2.5 \text{ in}} \right) = 3.6 \text{ lb/in.}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 158 \text{ lb} - (3.6 \text{ lb/in})(68 \text{ in}) = -87 \text{ lbs.}$$

$$\hat{y} = (-87 \text{ lb}) + (3.6 \text{ lb/in})(70.5 \text{ in}) = 167 \text{ lbs.}$$

\* With only 1 regression prediction to make, method 1 is somewhat faster, but with 2 or more, method 2 may be faster.

Back to C.S. 18: So I predict he weighs ~167 lbs, but give-or-take how much?



1) Suppose I ignore  $x$  and try to predict  $y$  anyway: best prediction would be  $\bar{y}$ , give or take the SD of  $y = s_y$ .

26 May 2005

(3)

2) now suppose instead I use  $x$  to predict  $y \rightarrow x = x^*$   
 $\rightarrow \hat{y} = \hat{a} + \hat{b}x^*$  (best prediction)

From the plot, my give or take should be smaller:

$$\hat{SE}(\hat{y}) \doteq S_y \sqrt{1-r^2}$$

( $\doteq$  means is approx. equal to)

$\hookrightarrow$  exact formula is more complicated, but this is good for  $x$ -values inside the football

check: if  $r = 0 \rightarrow \hat{SE}(\hat{y}) = S_y \checkmark$

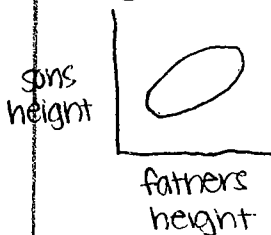
if  $r = +1, \text{ or } -1 \rightarrow \hat{SE}(\hat{y}) = 0 \checkmark$

ex. C.S. 18 -

if I don't know this guy's height I would predict his weight to be  $\bar{y} = 158$  lbs. and my give or take would be  $S_y = 25$  lbs.

if he's 70.5 in tall I predict he weighs 167 lbs. and my give or take for this prediction is  $\hat{SE}(\hat{y}) = S_y \sqrt{1-r^2} = 25 \text{ lb.} \sqrt{1 - .36^2} \doteq 23$  lbs. (not much smaller than the situation where I don't know his height at all, because height and weight are not very strongly correlated).

Why did Galton call it regression?



\* tall fathers tend to have tall sons, but: if father is 2 SD's above average in height, we only predict his son will be  $r \cdot 2$  SD's above avg. =  $(+0.5)(2 \text{ SD's}) = 1 \text{ SD}$  above avg.  $\rightarrow$  so tall fathers then tend to have tall sons, but not as tall as their fathers were. Similarly, short fathers tend to have short sons, but not as short as their fathers were.  $\Rightarrow$  "regression toward mediocrity"

now  $\Rightarrow$  regression effect