this time: regression

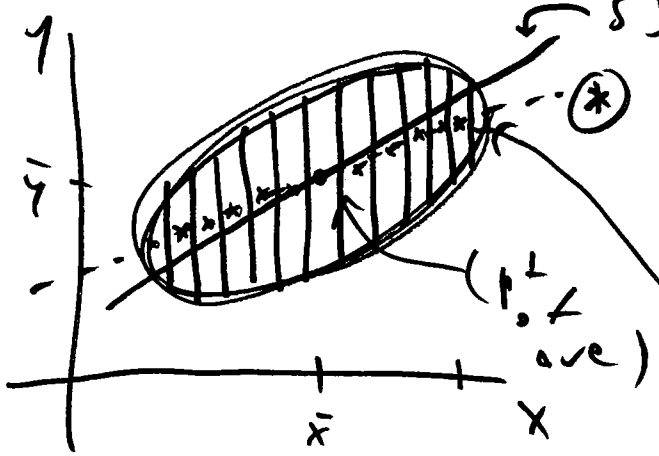
next time:

read: FPP ch. 10, 11 | 26 ① May

big picture: p. (22) VIII B | AMSS

no discussion sections or office hours on Mon (holiday); if Mon is your usual disc. sec. please go to another one on Thu (today) or Fri or Tue

no disc. sec. (next week) Wed - Fri either;
last section for qtr. will be sec. C next Tue 31 May, 7.30 - 8.40 pm



SD line (slope $\frac{s_y}{s_x}$) over 1 SD in x corresponds to up r SDs in y on the SD line

ave y value in middle of each vertical strip: graph of averages

Galton: find line that is a smooth version of graph of ave.

: regression line (⊛ - - - -) → slope of regression line is $\boxed{r\,\frac{s_y}{s_x}} = \hat{b}$

AMS 5

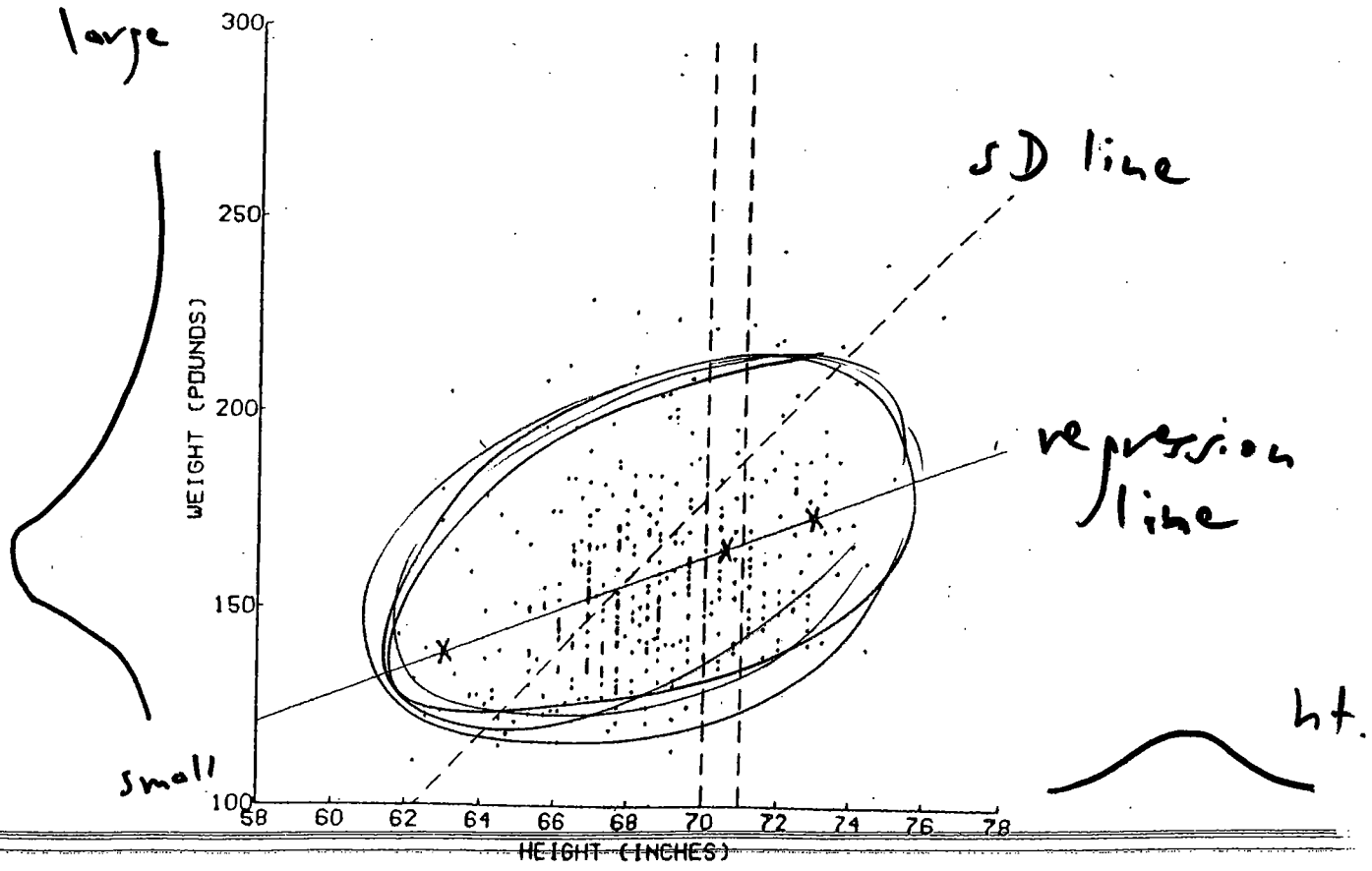
Case Study 18 (physiology): Height and Weight

In 1960-1962, the U.S. Public Health Service examined a representative cross section of 6672 Americans aged 18 to 79 in what was called Cycle I of the *Health Examination Survey* (HES). The objective was to take a snapshot of the country's health status, by getting baseline data about physiological variables like height, weight, and blood pressure, psychological variables like anxiety, demographic variables like age, education, and income, and prevalence of diseases, especially various kinds of heart disease. The scatter diagram below gives heights and weights for the 411 men aged 18 to 24 in this survey. How well can you predict a man's weight from his height?

Here are some numerical summaries of the data:

variable	mean	SD	
(x) height	68 inches	2.5 inches	n = 411
(y) weight	158 pounds	25 pounds	r = +0.36

One of the lines in the scatterplot below is the SD line, and the other is the regression line. Which is which? Which one does a better job of estimating the average weight of people in a given vertical strip? Explain briefly. Predict the weight of a randomly chosen man in the vertical strip in the plot below -- these men were 70.5 inches tall (one SD above average) -- and attach an approximate give-or-take to your prediction. Would you say from this that a person's weight can be quite accurately forecast from his height, or is the relationship not that strong? Explain briefly.



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 ok, this guy is $70\frac{1}{2}$ in tall → (mean 68 in, SD 2.5 in)

so in SD's $\frac{70\frac{1}{2} - 68}{2.5} = +1 = 1$ SD above ave in x

→ we predict he will be $+0.36$ SDs above ave in y (mean 158 lb, SD 25 lb) →

$(0.36)(25\text{ lb}) = 9$ lb above ave → 167 lb.

So I predict that a guy who's $70\frac{1}{2}$ in tall will weigh about 167 lb.

② work out slope & intercept of regression line & plug it into the equation

slope $\hat{b} = r\frac{s_y}{s_x}$ | $\hat{a} = ?$

Fact reg. line goes through point of averages $(\bar{x}, \bar{y})$:

$\hat{y} = \hat{a} + \hat{b}x$ (in general)

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