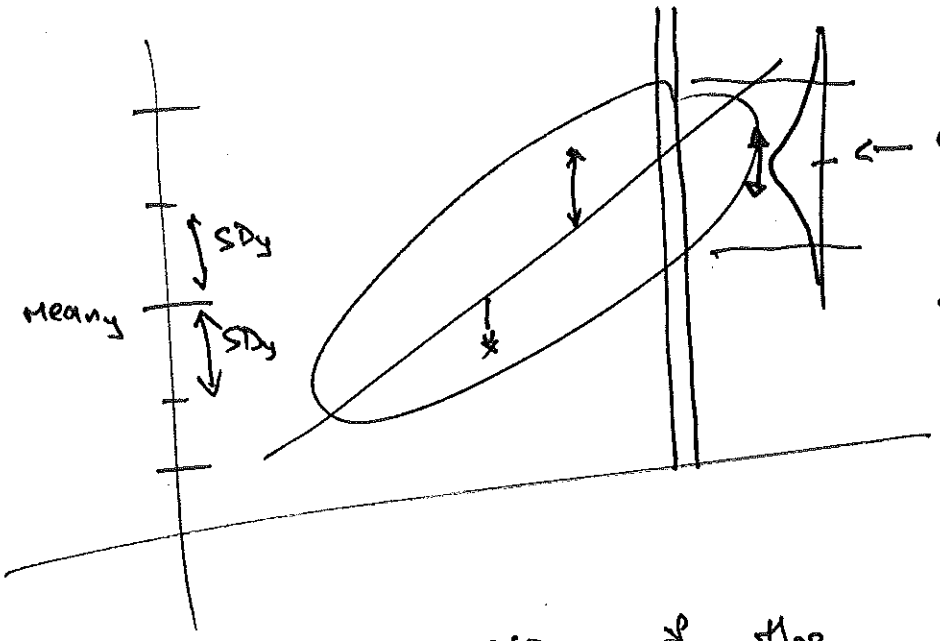
RMS.

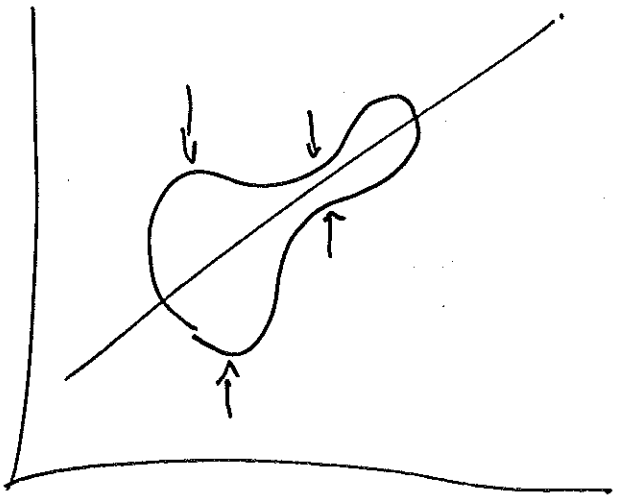


RMS error is average size of the prediction errors.

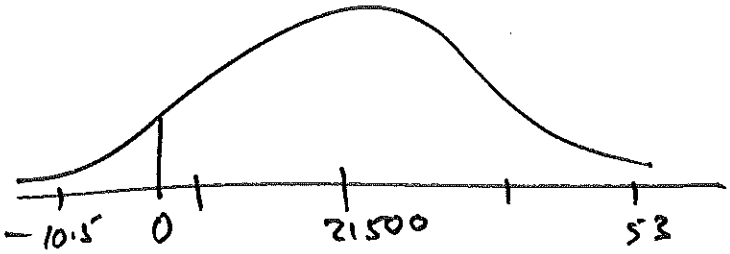
e_1
e_2
e_3
$\vdots$
e_n

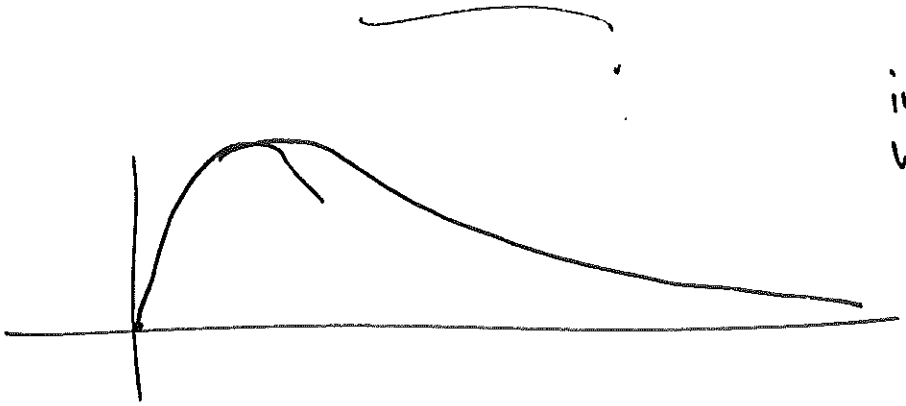
$$\text{RMS error} = \sqrt{e_1^2 + e_2^2 + \dots e_n^2}/n$$

$$= \sqrt{1 - r^2}\ SD_y.$$



RMS error either under / over estimates if the distribution is not "football" shaped

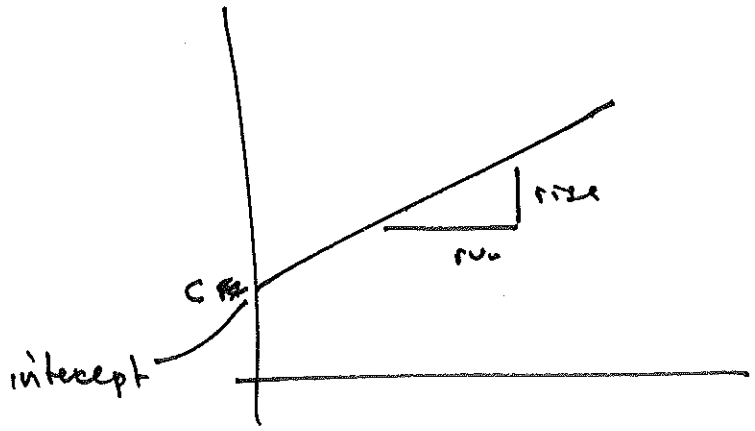




income distributions have a long right tail (+ a cut-off at 300).

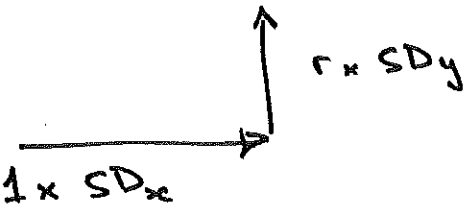
Regression line.

$y = mx + c$



$$m = \frac{\text{rise}}{\text{run}}$$

slope of regression line.



$$\text{slope of regression line} = \frac{r \times SD_y}{SD_x}$$

Example.

555 California men age 25-29 in 1993 were surveyed.

years of education — mean 12.5 — SD 4

income — mean \$21,500 — SD \$16,000

$r = 0.35$

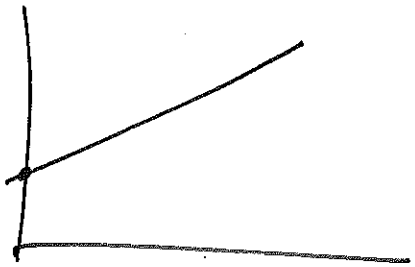
if increase education by 4 years (1 SD)

income increases by $0.35 \times 16{,}000 = 5{,}600$

=> 1 additional year of education

gives $\frac{5600}{4} = 1400$ increase in income.

slope is 1400.



Intercept

is y-value when $x = 0$.

0 years of education

which is 12.5 years below the mean.

each year is worth \$1400

12.5 years below mean corresponds to

$21,500 - 12.5 \times 1400$

$21,500 - 17500$

$= 4000$

$$\text{income} = 1400 \times \#\text{years of education} + 4000$$

$$y = 1400x + 4000$$

$$\text{slope} = \frac{r \times SD_y}{SD_x}$$

$$\text{intercept} = \text{mean}_y - \text{slope} \times \text{mean}_x$$

Regression line is useful when we have to make many predictions.

What is the predicted income of a man with 15 years of education?

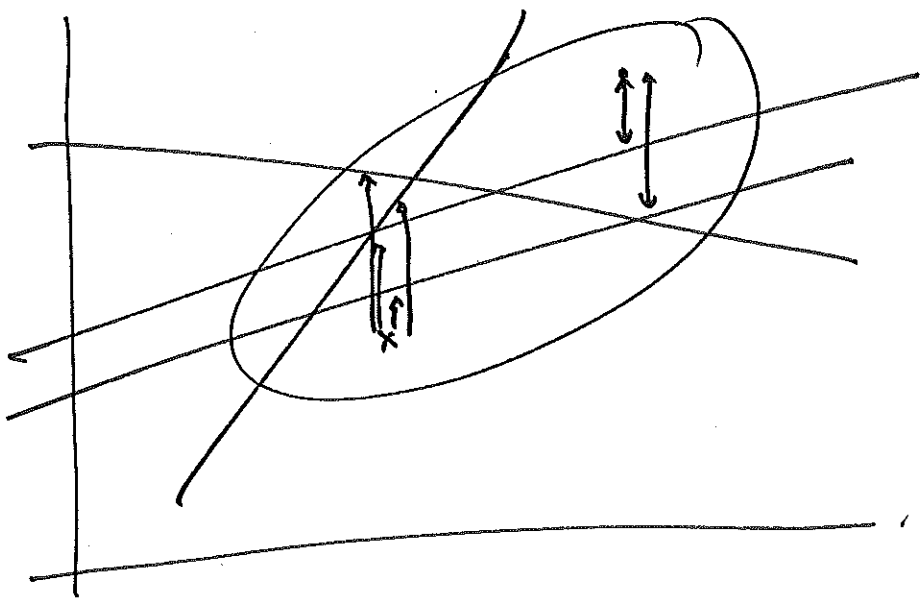
$$y = 1400x + 4000$$
$$= 1400 \times 15 + 4000$$
$$= 25{,}000$$

Q: what happens if we make an intervention?

- maybe

Yes - if the data come from a controlled experiment

No - if the data are from an observational study. - confounding factors.



we can draw many lines through this cloud of points.

Why should we prefer one over another?

We can compute the RMS error for each possible line.

Which line has the smallest RMS error?

→ the regression line.

also known as the least squares line.

Example.

Mid term scores	mean	16	SD	2.8
final scores	mean	40	SD	1

$r = 0.1$

slope of regression line $\dfrac{0.1 \times 1}{2.8} = 0.04$

intercept $= 40 - 0.04 \times 16 = 39.41$

final score = 0.04 × midterm score + 39.41.