Question: Suppose that 1500 men are surveyed and their level of education and annual income is observed. Of these men, 140 have more than 16 years of education.

Which is higher, the correlation between income and education for the entire set of data, or the correlation between income and education for the 140 most highly educated men?

Answer: *The correlation should be smaller in the smaller set because there will typically be more variation in the incomes for each level of education. With a narrower range of education levels, other variables play a bigger role, so the association will be weaker.*

Observation:

If there is a lot of variation in y-values (a big range) for a relatively small range of x-values, then the correlation coefficient will tend to be smaller.

Lines and linear functions: a refresher

- A straight line is the graph of a linear equation. These equations come in several forms, for example:

  $$(i)\ ax + by = c, \quad (ii)\ y = y_0 + m(x - x_0), \quad (iii)\ y = mx + b.$$

- The ***slope*** of a line is the ratio

  $$\frac{\text{rise}}{\text{run}} = \frac{\text{change in } y}{\text{change in } x} = \frac{y_1 - y_0}{x_1 - x_0}$$

  In equations (ii) and (iii) above, the slope is given by m.

- The slope is the amount by which y is changing for every *one unit* change in x.

  In other words:

  $$\overbrace{y - y_0}^{\text{change in } y} = m \cdot \overbrace{(x - x_0)}^{=1} = m.$$

Another property of r_{xy}*...*

(*) The correlation coefficient gives a measure of how the data is clustered around the **SD-line**.

(*) The SD-line is the line that passes through the *point of averages* $(\overline{x}, \overline{y})$ with *slope* $m = \pm \dfrac{SD_y}{SD_x}$.

$\Rightarrow$ The slope is $\dfrac{SD_y}{SD_x}$ if $r \geq 0$.

$\Rightarrow$ The slope is $-\dfrac{SD_y}{SD_x}$ if $r < 0$.

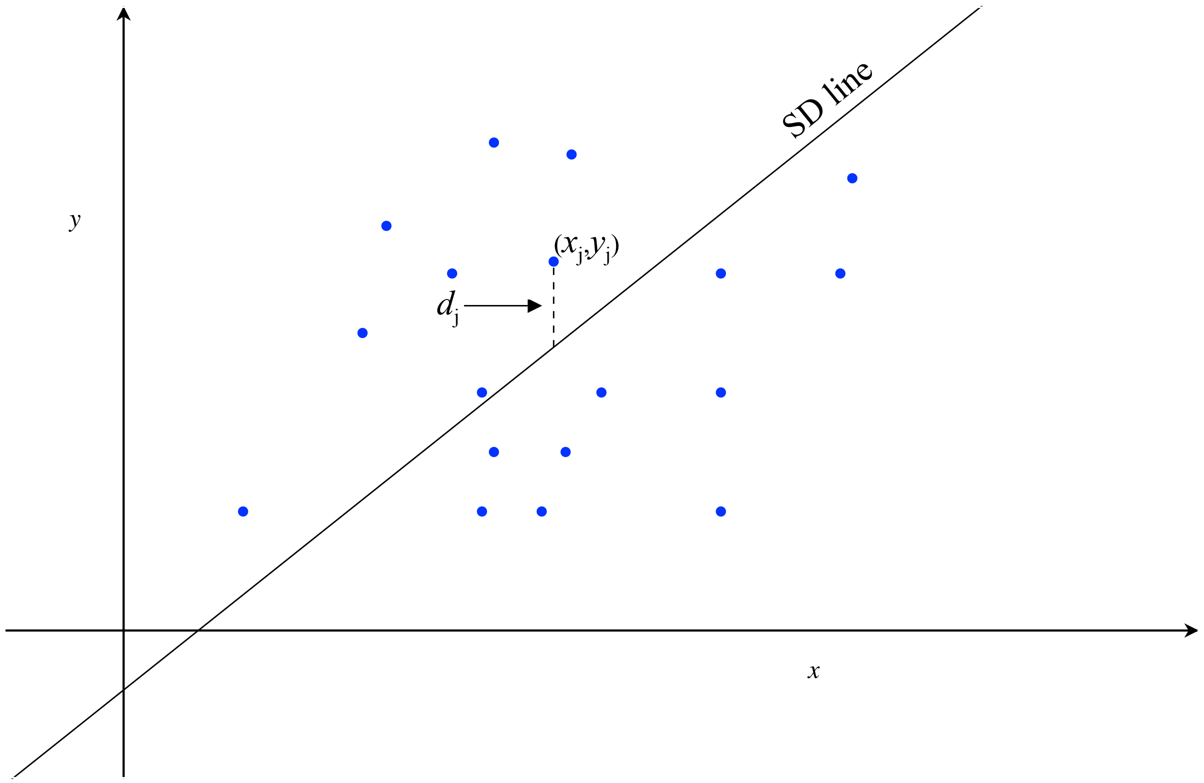
(*) This is the line with equation

$$y - \overline{y} = \pm \frac{SD_y}{SD_x}(x - \overline{x}).$$

On this line, y increases (or decreases) by one SD_y for every one SD_x increase in x.

Correlation and clustering around the SD-line

(*) Suppose that (x_j, y_j) is a point in the scatter plot, and d_j is the vertical distance of this point from the SD-line.



The (vertical) clustering is quantified by $\sqrt{\dfrac{1}{n}\sum_j d_j^2}$.

(*) The R.M.S. of the vertical distances to the SD-line can be computed quickly in terms of the correlation coefficient:

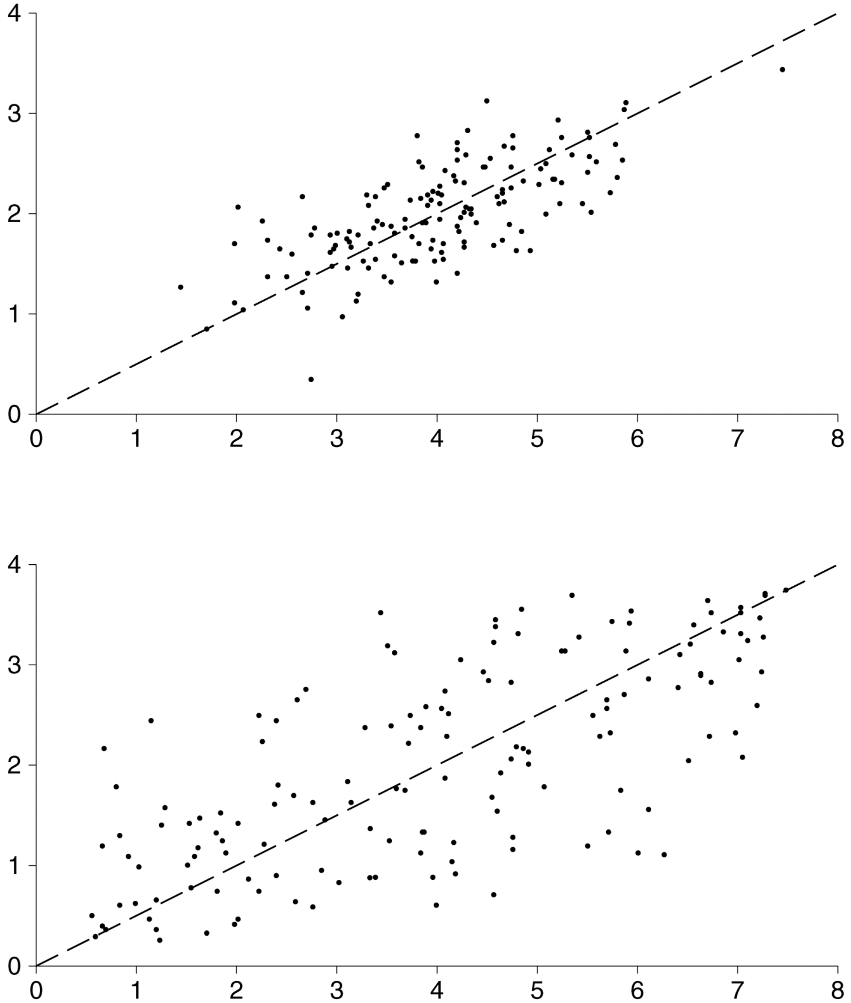
$$\sqrt{\frac{1}{n}\sum_j d_j^2} = \sqrt{2(1-|r_{xy}|)} \times SD_y.$$

The smaller this number is, the more tightly clustered the points are around the SD line.

(*) The closer $|r_{xy}|$ is to 1, the more tightly clustered the data will be around the SD line. But this measure of clustering around the SD-line depends on both r_{xy} ***and*** SD_y.

⇒ Two sets of data can have the same correlation, even though one of them appears to be more tightly clustered around the SD line than the other, because of changes in scale (smaller standard deviations).

Figure 3. The effect of changing SDs. The two scatter diagrams have the same correlation coefficient of 0.70. The top diagram looks more tightly clustered around the SD line because its SDs are smaller.



Given a set of paired data, we want:

A ***formula*** for predicting the (approximate) y-value of an observation with a given x-value.

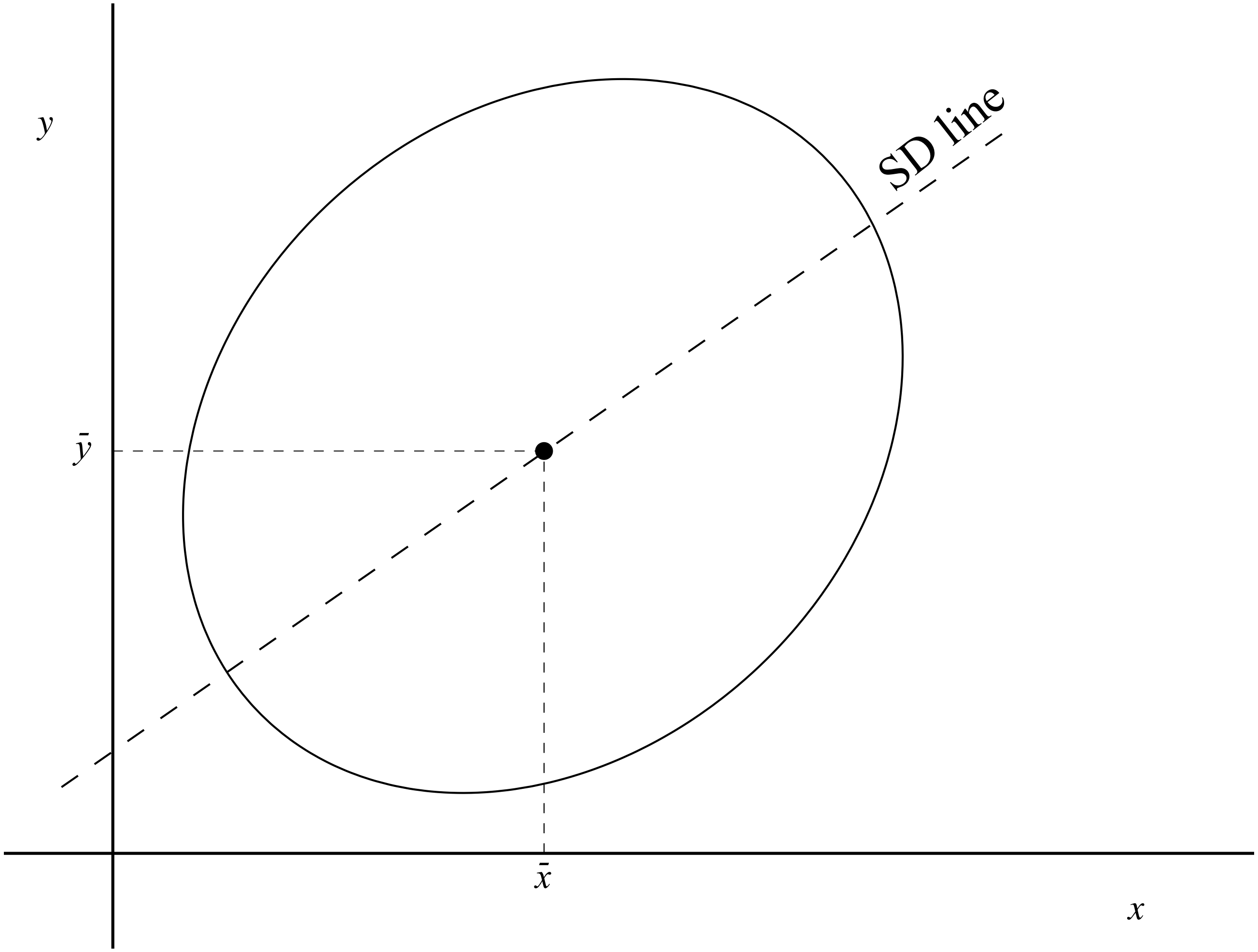
What we can reasonably hope to find:

A formula for predicting the (approximate) ***average*** y-value for all observations having the same x-value.

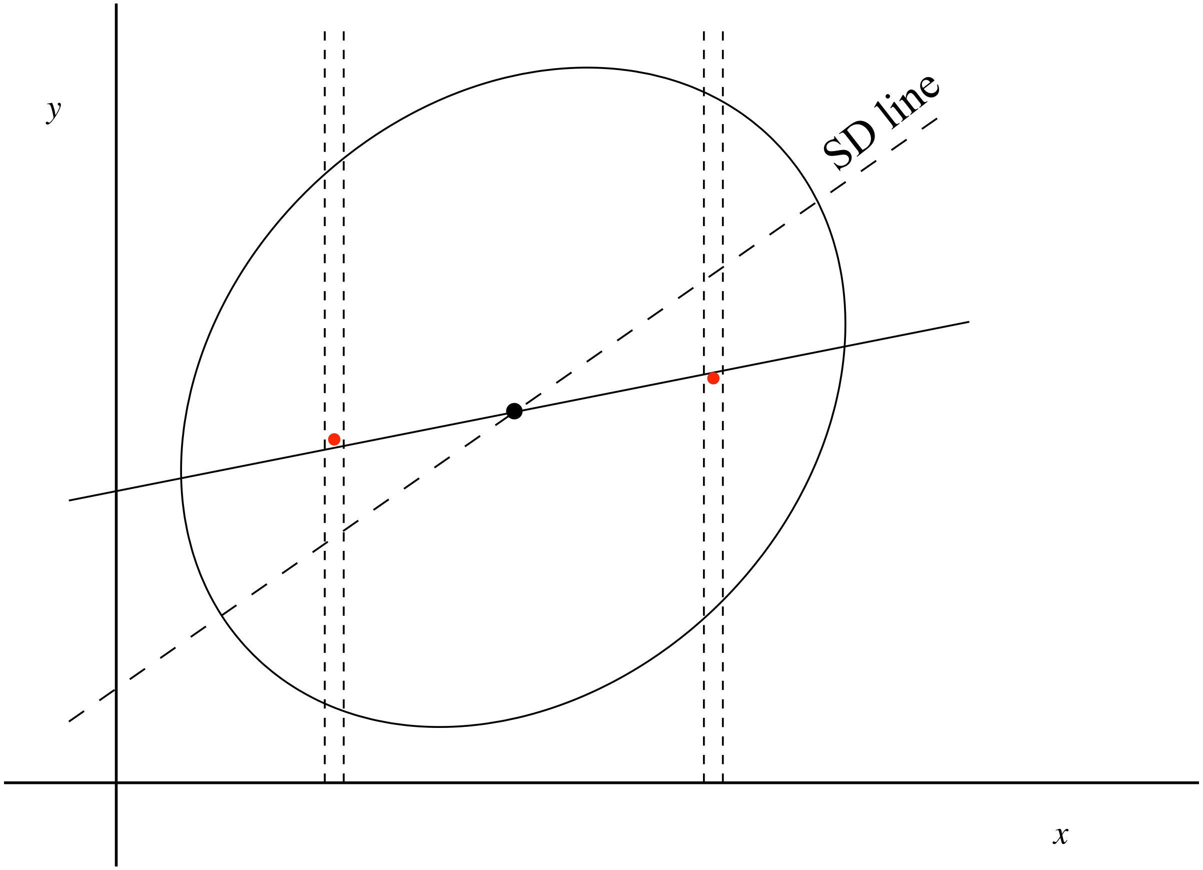
First Guess: **The SD-line.**

(*) The SD line is a natural candidate and predicts the average y-value accurately when $x = \overline{x}$ (it predicts $\overline{y}$), but...

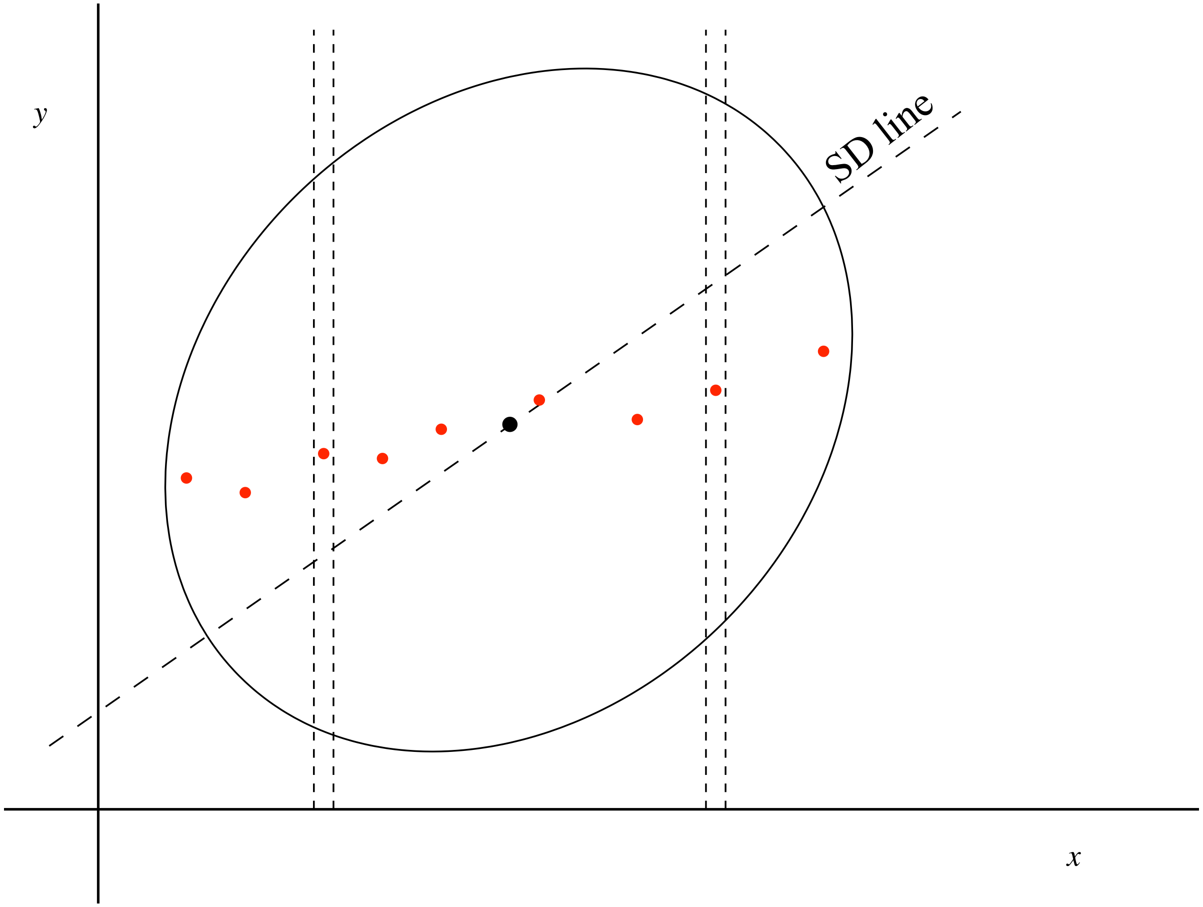
(*) As the observations move away from the *point of averages*, the points on the SD-line tend to lie above or below the average y-value that we are trying to estimate, and the further we move from the point of averages, the bigger the errors become.



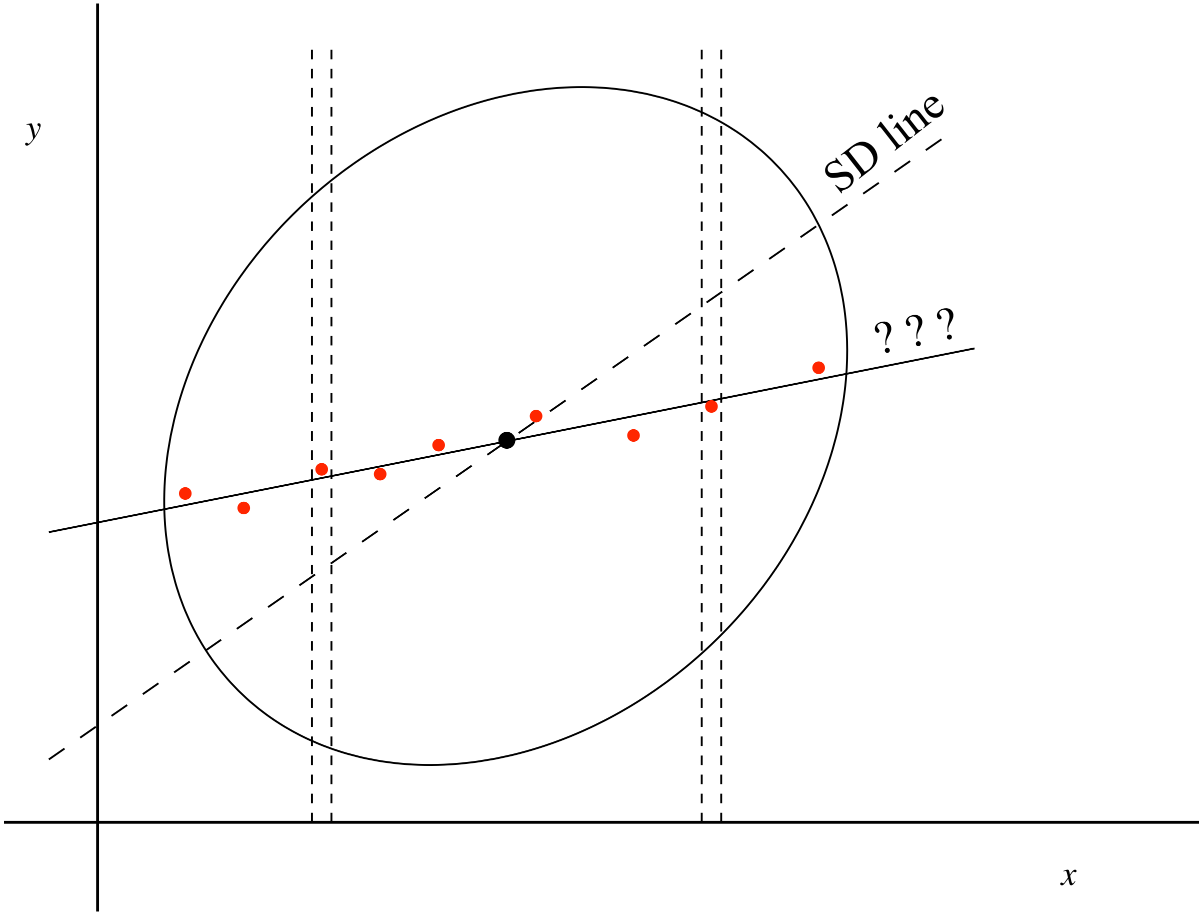
(*) Hypothetical (cloud of) data with point of averages and the SD line.



(*) We want to estimate the average y-value of the points in each vertical strip (the red dots). But The SD line is underestimating the average in the strip to the left of the point of averages and overestimating the average of the strip to the right of the point of averages.



(*) Hypothetical (cloud of) data with *graph of averages* (red dots) and the SD line. The further the vertical strip is from the point of averages, the worse the SD line approximates the average height of the data in that strip.



We want to find the line that

(i) Passes through the point of averages.

(ii) Approximates the graph of averages as best as possible.

Question: *What information is missing from the SD line?*

Answer: The ***correlation*** between the variables!

(*) Taking correlation into account leads to the ***regression*** line.

- The regression line passes through the point of averages.
- The ***slope*** of the regression line (for y on x) is given by

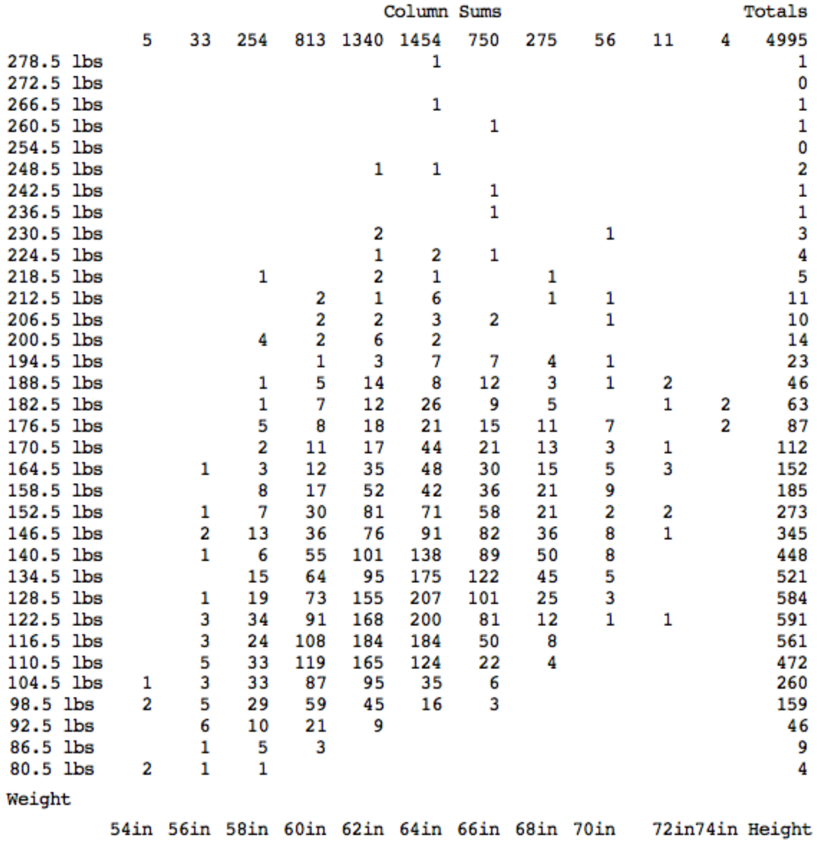
$$r_{xy} \cdot \frac{SD_y}{SD_x}.$$

- The regression line predicts that for every SD_x change in x, there is an approximate $r_{xy} \cdot SD_y$ change in the ***average value*** of the corresponding ys.

Paired data and the relationship between the two variables (x and y) is summarized by the five statistics:

$$\overline{x}, \quad SD_x, \quad \overline{y}, \quad SD_y \text{ and } r_{xy}.$$

Example: Regression of weight on height for women in Great Britain in 1951.



					Column Sums							Totals
	5	33	254	813	1340	1454	750	275	56	11	4	4995
278.5 lbs						1						1
272.5 lbs												0
266.5 lbs						1						1
260.5 lbs							1					1
254.5 lbs												0
248.5 lbs					1	1						2
242.5 lbs							1					1
236.5 lbs							1					1
230.5 lbs					2				1			3
224.5 lbs					1	2	1					4
218.5 lbs			1		2	1		1				5
212.5 lbs				2	1	6		1	1			11
206.5 lbs				2	2	3	2		1			10
200.5 lbs			4	2	6	2						14
194.5 lbs				1	3	7	7	4	1			23
188.5 lbs			1	5	14	8	12	3	1	2		46
182.5 lbs			1	7	12	26	9	5		1	2	63
176.5 lbs			5	8	18	21	15	11	7		2	87
170.5 lbs			2	11	17	44	21	13	3	1		112
164.5 lbs		1	3	12	35	48	30	15	5	3		152
158.5 lbs			8	17	52	42	36	21	9			185
152.5 lbs		1	7	30	81	71	58	21	2	2		273
146.5 lbs		2	13	36	76	91	82	36	8	1		345
140.5 lbs		1	6	55	101	138	89	50	8			448
134.5 lbs			15	64	95	175	122	45	5			521
128.5 lbs		1	19	73	155	207	101	25	3			584
122.5 lbs		3	34	91	168	200	81	12	1	1		591
116.5 lbs		3	24	108	184	184	50	8				561
110.5 lbs		5	33	119	165	124	22	4				472
104.5 lbs	1	3	33	87	95	35	6					260
98.5 lbs	2	5	29	59	45	16	3					159
92.5 lbs		6	10	21	9							46
86.5 lbs		1	5	3								9
80.5 lbs	2	1	1									4
Weight												
	54in	56in	58in	60in	62in	64in	66in	68in	70in	72in	74in	Height

Reproduced from Kendall and Stuart, *op. cit.*, p. 300.

Summary statistics:

$$\bar{h} \approx 63 \text{ inches}, \quad s_h \approx 2.7 \text{ inches},$$

$$\bar{w} \approx 132 \text{ lbs}, \quad s_w \approx 22.5 \text{ lbs}.$$

$$r_{hw} \approx 0.32$$

The summary statistics can be used to estimate the weights of women given information about their height.

Question: How much did 5’6”-tall British women weigh in 1951 on average?

Answer: These women were 3 inches above average height. This is

$$\frac{3}{2.7} \approx 1.11\, SD_h \text{ above average height.}$$

The regression line predicts that on average, they would have weighed about

$$0.32 \times 1.11 \approx 0.355\, SD_w \text{ above the average weight.}$$

So, the average weight for these women would have been about

$$132 + 0.355 \times 22.5 \approx 140 \text{ lbs.}$$

Question: By about how much did average weight increase for every 1 inch increase in height?

Answer: 1 inch represents

$$\frac{1}{2.7} \approx 0.37\, SD_h,$$

so each additional inch of height would have added about

$$0.32 \times 0.37 \approx 0.1184\, SD_w = 0.1184 \times 22.5\text{lbs} \approx 2.66\text{lbs}$$

to the average *weight*.

Example. A large (hypothetical) study of the effect of smoking on the cardiac health of men, involved 2709 men aged 25 - 45, and obtained the following statistics,

$$\overline{x} = 17,\ SD_x = 8,\ \overline{y} = 129,\ SD_y = 7,\ r_{xy} = 0.64,$$

where

(*) y_j = systolic blood pressure measured in mmHg of the j^{th} subject

(*) x_j = number of cigarettes smoked per day by j^{th} subject.

Question: What is the predicted average blood pressure of men in this age group who smoke 20 cigarettes per day?

Answer: 20 cigarettes is 3 cigarettes *above average*, which is $3/8 \cdot SD_x$ above average. The regression line predicts that the average blood pressure of men who smoke 20 cigarettes/day will be

$$r_{x,y} \times \left(\frac{3}{8} \cdot SD_y\right) = 0.64 \times \left(\frac{3}{8} \cdot 7\right) \approx 1.68$$

mmHg *above average* — about 130.68 mmHg.

Question: John is a 31-year old man who smokes 30 cigarettes a day. What is John's predicted blood pressure.

Answer: Our best guess for John is the average blood pressure of men who smoke 30 cigarettes a day. Since 30 is $13 = 13/8 \times SD_x$ above $\overline{x}$, the regression line predicts that John's blood pressure will be about

$$r_{x,y} \times \left(\frac{13}{8} \cdot SD_y\right) = 0.64 \times \left(\frac{13}{8} \cdot 7\right) \approx 7.28$$

mmHg *above average* — about 136.28 mmHg.

(*) How accurate is this estimate likely to be?

⇒ What is the ***spread*** around $\hat{y} = 136.28$ in the vertical strip corresponding to $x = 30$?