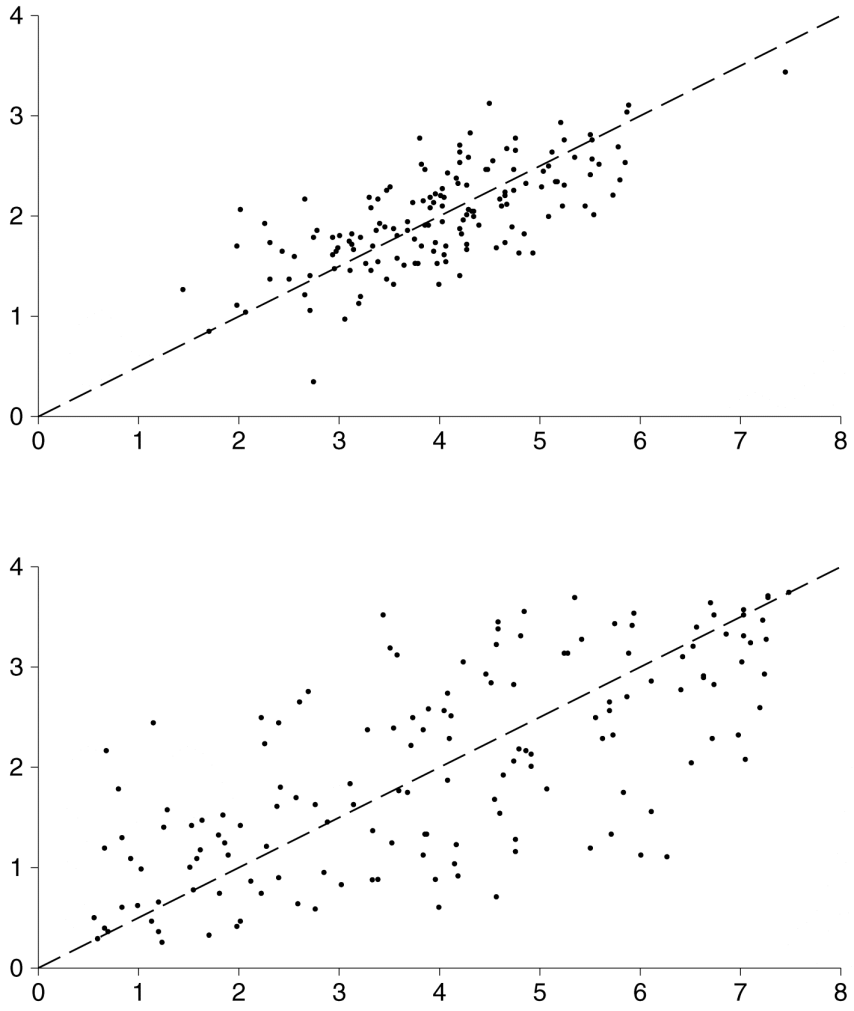
The correlation coefficient.

For a set of paired data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ the correlation coefficient is

$$r_{xy} = \frac{1}{n} \sum z_{x_j} \cdot z_{y_j}$$

- $-1 \leq r_{xy} \leq 1$ for all sets of paired data. If $r_{xy} = \pm 1$ then $y_j = mx_j + b$, for all (x_j, y_j) in the set.
- r_{xy} is not useful for identifying *nonlinear* relationships. E.g., you can have $r_{xy} \approx 0$, even though $y \approx f(x)$ for some nonlinear function f.
- r_{xy} gives an indication of how the data is clustered about a line, but only relative to the sizes of the SDs. I.e., two sets of paired data can have identical correlations, with one of the two scatterplots appearing to be much more tightly clustered around a line than the other.

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.



- Data with significant outliers are not well-described by the correlation coefficient.
- If there is a lot of variation in y-values for the same (or similar) x-values, then the correlation coefficient will tend to be smaller.

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: *We would expect the correlation to 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.*

Question: Suppose that 2600 women are surveyed across the U.S. and their educational levels and annual incomes are observed. Additionally, the average level of education and the average income are computed state by state for the women in the study.

Which is higher, the correlation between income and education for all 2600 observations or the correlation between average income and average education for the 50 states?

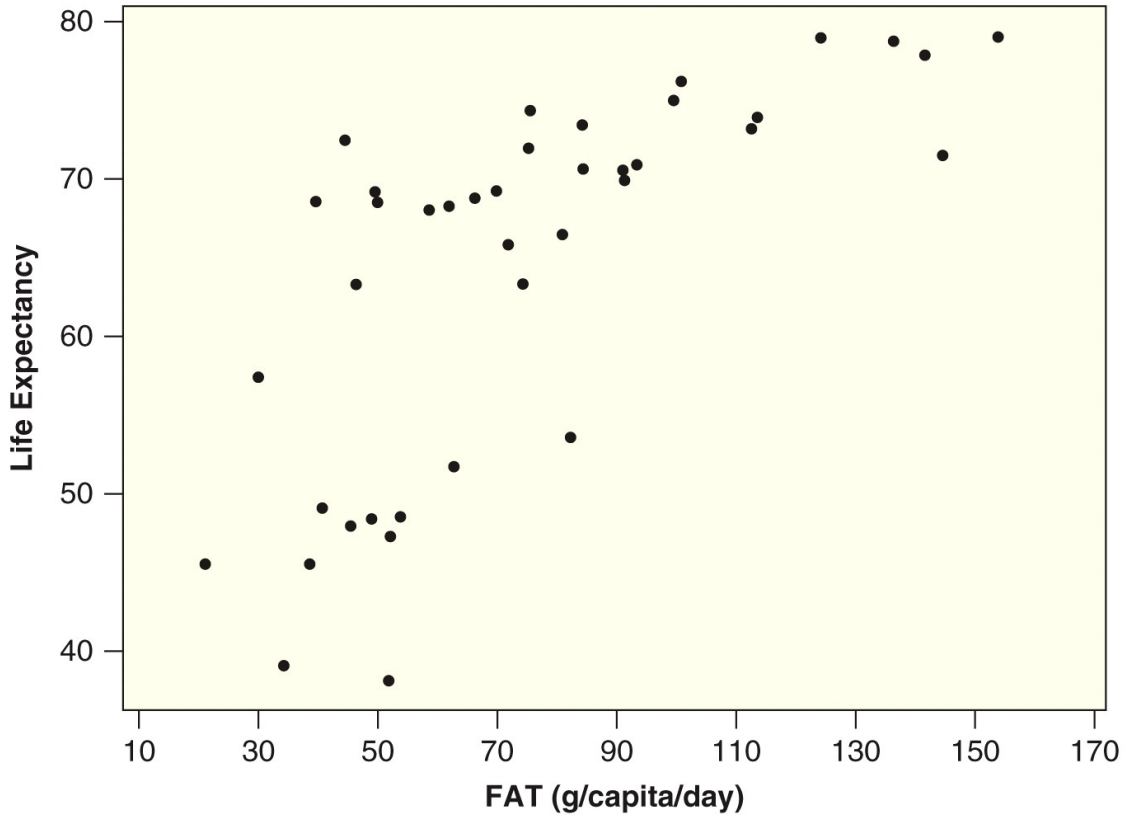
Answer: *The correlation between average education level and average income by state will typically be higher. Taking averages lowers the variation in income level for each education level and this will tend to make the correlation coefficient higher.*

Ecological correlations: An *ecological correlation* is one that measures the correlation between class averages across several classes, instead of measuring the correlation between the variables for individual observations.

Ecological correlations tend to exaggerate the strength of the relationship because class averages *reduce the variability in the data.* Class averages tend to cluster around a line much more nicely than individual data.

The smaller the number of classes, the more pronounced this effect can be.

Fat consumption and life expectancy, *by country*.



The correlation here is $r \approx 0.7$, but this exaggerates the strength of the relationship. If individual data was plotted and summarized instead of national averages, the correlation would be much weaker.

Lines and linear functions: a refresher

- A straight line is the graph of a linear equation. These 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 *unit* change in x.

  In other words:

  $$y - y_0 = m \cdot (x - x_0).$$

Given a set of paired data, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, what (straight) line best describes the relationship seen in the data?

First Guess: The ***SD-line***. This is the line that passes through the *point of averages*, $(\overline{x}, \overline{y})$, with slope SD_y/SD_x if the association is positive, and slope $-SD_y/SD_x$ if the association is negative.

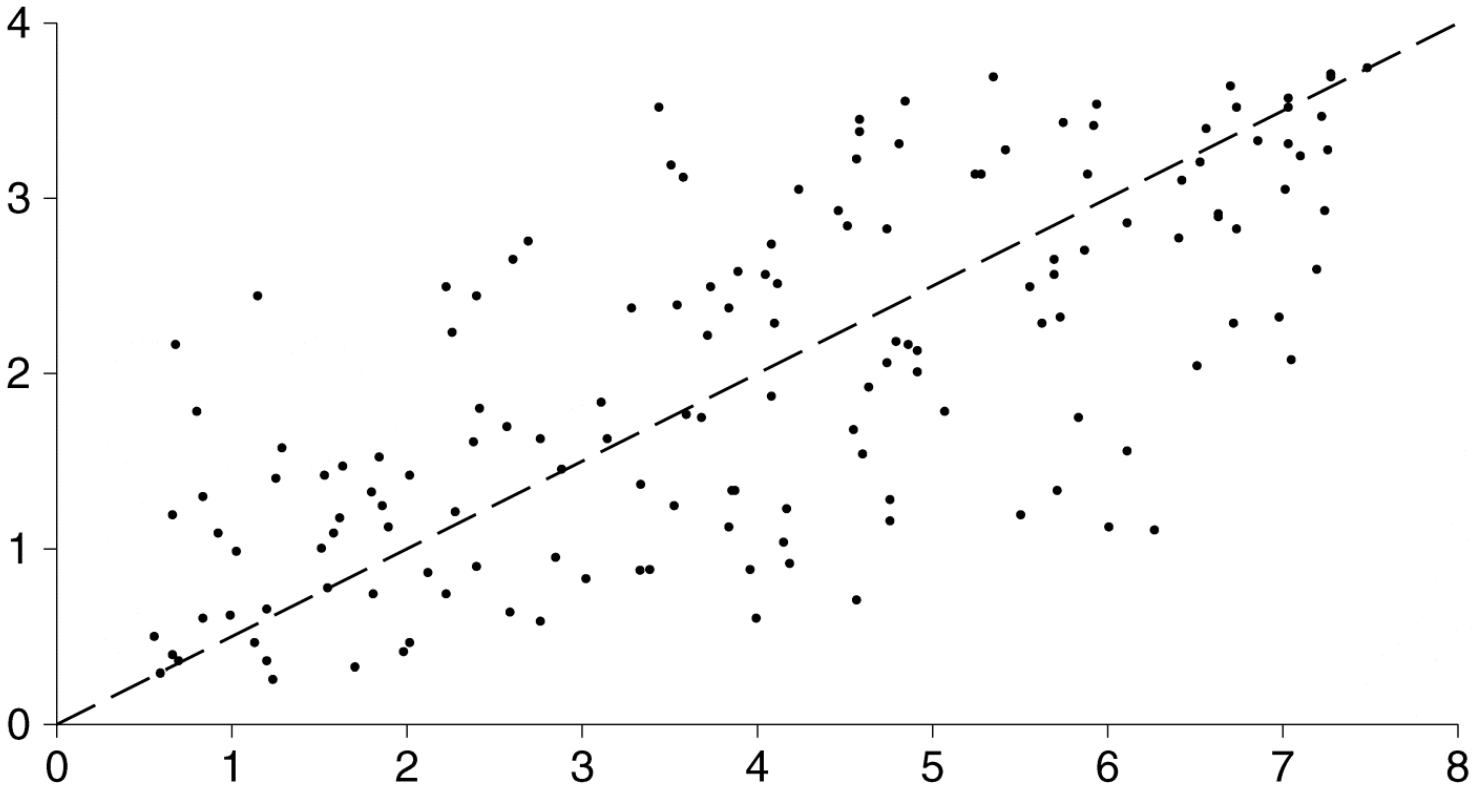
Equations:

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

if the association between x and y is positive, and

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

if the association between x and y is negative.


4
3
2
1
0
0
1
2
3
4
5
6
7
8

The SD-line does...

• ...run through the center of the scatterplot, i.e., the points in the scatterplot are spread more-or-less symmetrically around the SD-line;

• ...describe the trend in the data as a whole. I.e., it gives the correct *general direction* for the 'cloud' of points in the scatterplot.

The SD-line does ***not...***

• ...do a good job of ***predicting*** y-values from given x-values.

Why not?

• The SD-line doesn't take the *correlation* between the variables into account.

What we want:

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

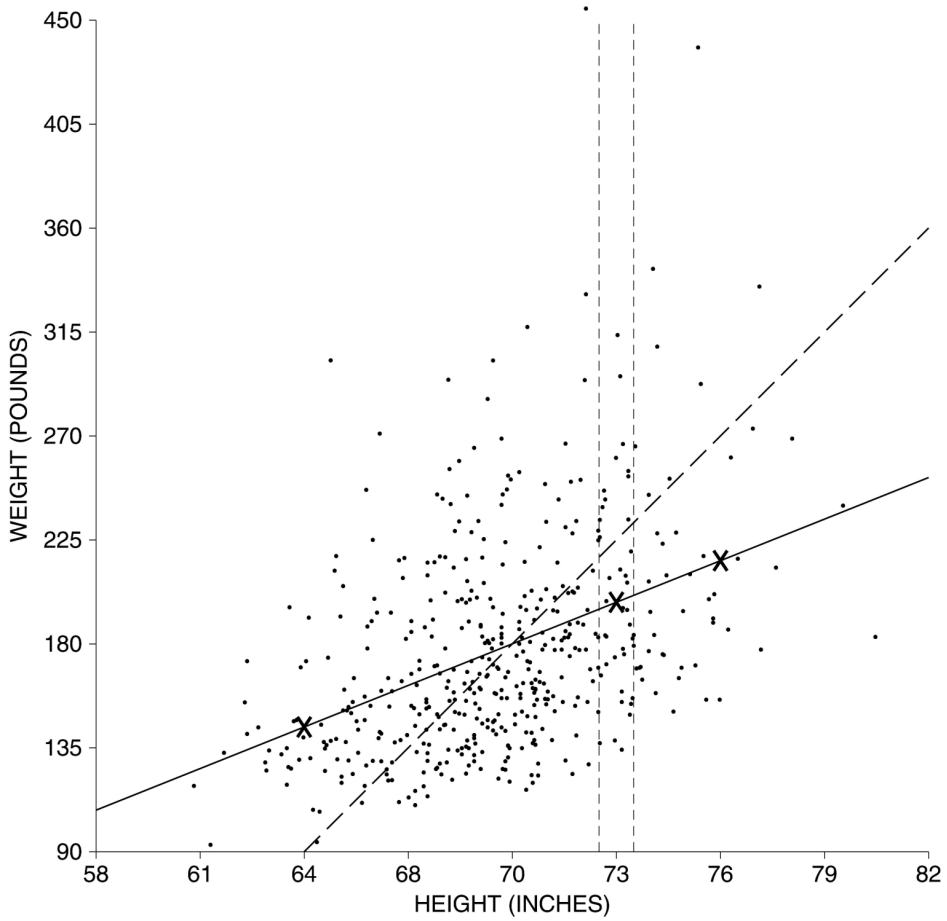
What we can reasonably hope to find:

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

The SD-line prediction:

For every 1 SD_x increase in x, there is a 1 SD_y increase in the *average* value of y... But this is usually *wrong*:

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.



Conclusion: The SD-line goes up (or down) too steeply. I.e., its slope $\pm SD_y/SDx$ is too *big* (in absolute value).... To better predict the average y-value for a given x-value, we use ...

Better Guess: The *regression* line.

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

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

- In practical terms, 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.

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

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

$$r_{hw} \approx 0.32$$

These numbers can be used to answer questions about average weights for women of different heights...

- How much do 5'6"-tall women weigh on average?

  These women are $3/SD_h = 3/2.7 \approx 1.11$ standard deviations above average (height), so, on average they will about $0.32 \times 1.11 \approx 0.355$ standard deviations above the average weight. I.e., the average weight for these women is about

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

- How much does average weight go up when height increases by 1 inch?

  1 inch represents $1/SD_h = 1/2.7 \approx 0.37$ standard deviations for *height*, so each additional inch of height adds about $0.32 \times 0.37 \approx 0.1184$ SD_w to the average *weight*: this is about 2.66 lbs.