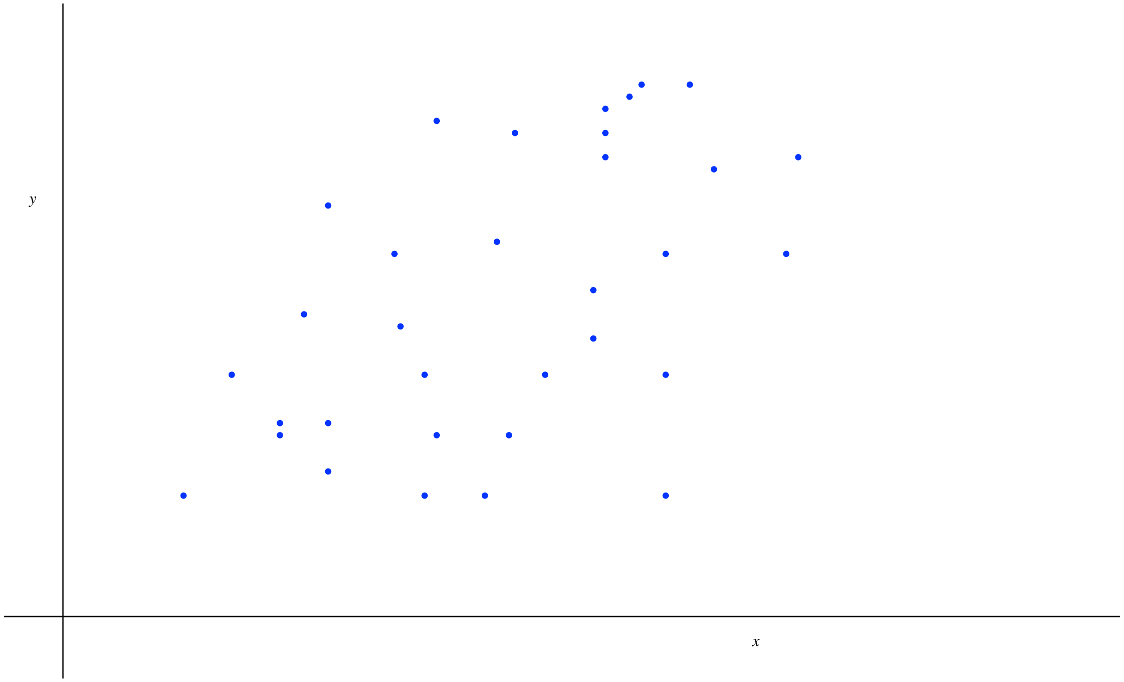
Linear Regression

(*) Given a set of paired data, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, we want a method (formula) for predicting the (approximate) y-value of an observation with a given x-value.



Problem: There are many observations with the same x-value but different y-values... $\Rightarrow$ *Can't predict one y-value from x.*

(*) More realistic goal: a method (formula) for predicting the (approximate) ***average*** y-value for all observations having the same x-value.

Notation:

- $\overline{y}(x_j)$ = average y-value for all observations with x-value $= x_j$.
- $\widehat{y}_j$ = *estimated* value of $\overline{y}(x_j)$.
- We want a method for calculating $\widehat{y}_j$ from x_j.

(*) We want the method (formula) to be *linear* — this means that there is a specific line and the points $(x_j, \widehat{y}_j)$ all lie on this line.

(*) In other words, we want to find a formula of the form

$$\widehat{y}_j = \beta_0 + \beta_1 x_j.$$

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

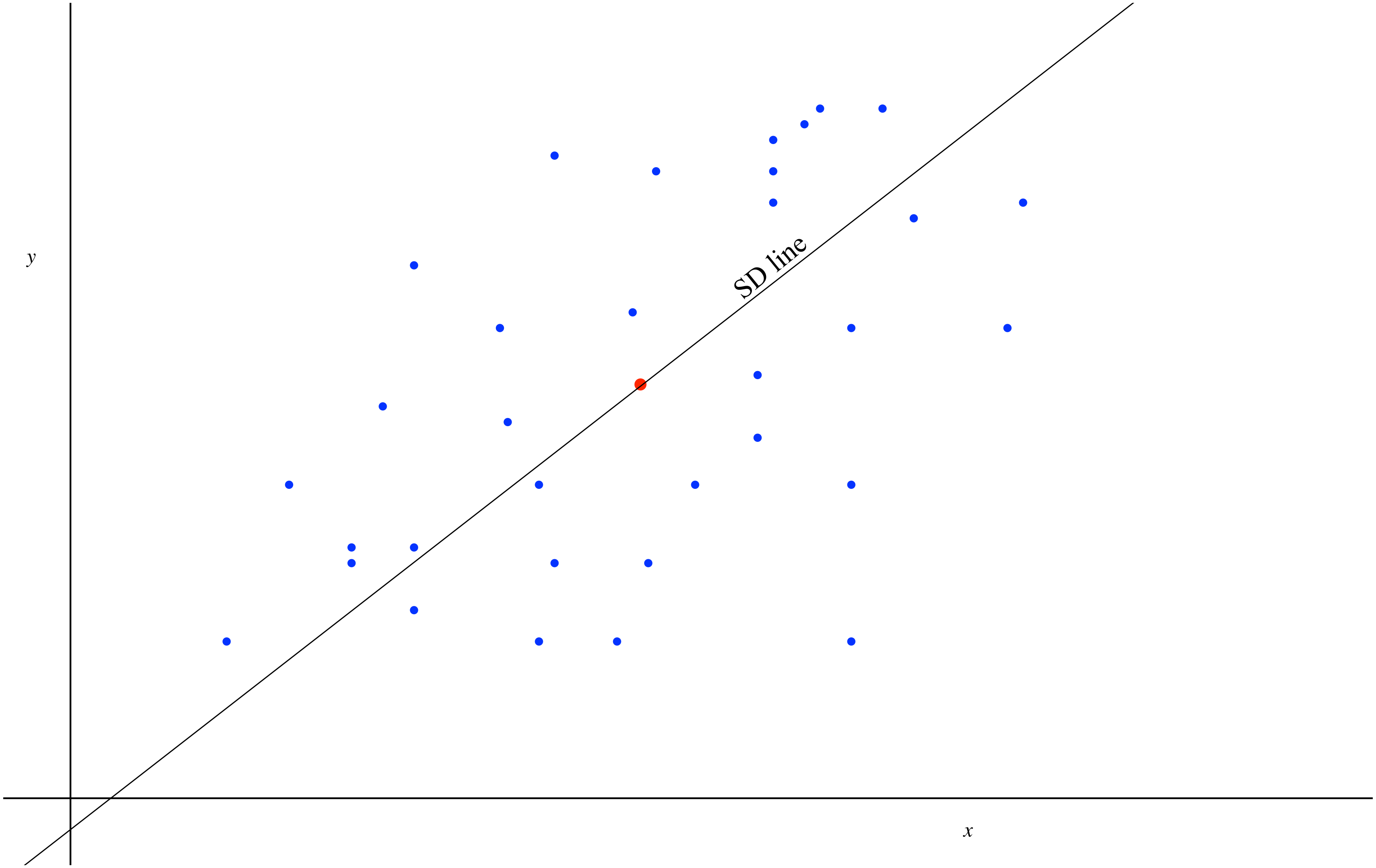
(*) The SD-line is the line that passes through the *point of averages* $(\overline{x}, \overline{y})$ with ***slope*** $m = \pm\frac{SD_y}{SD_x}$.

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

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

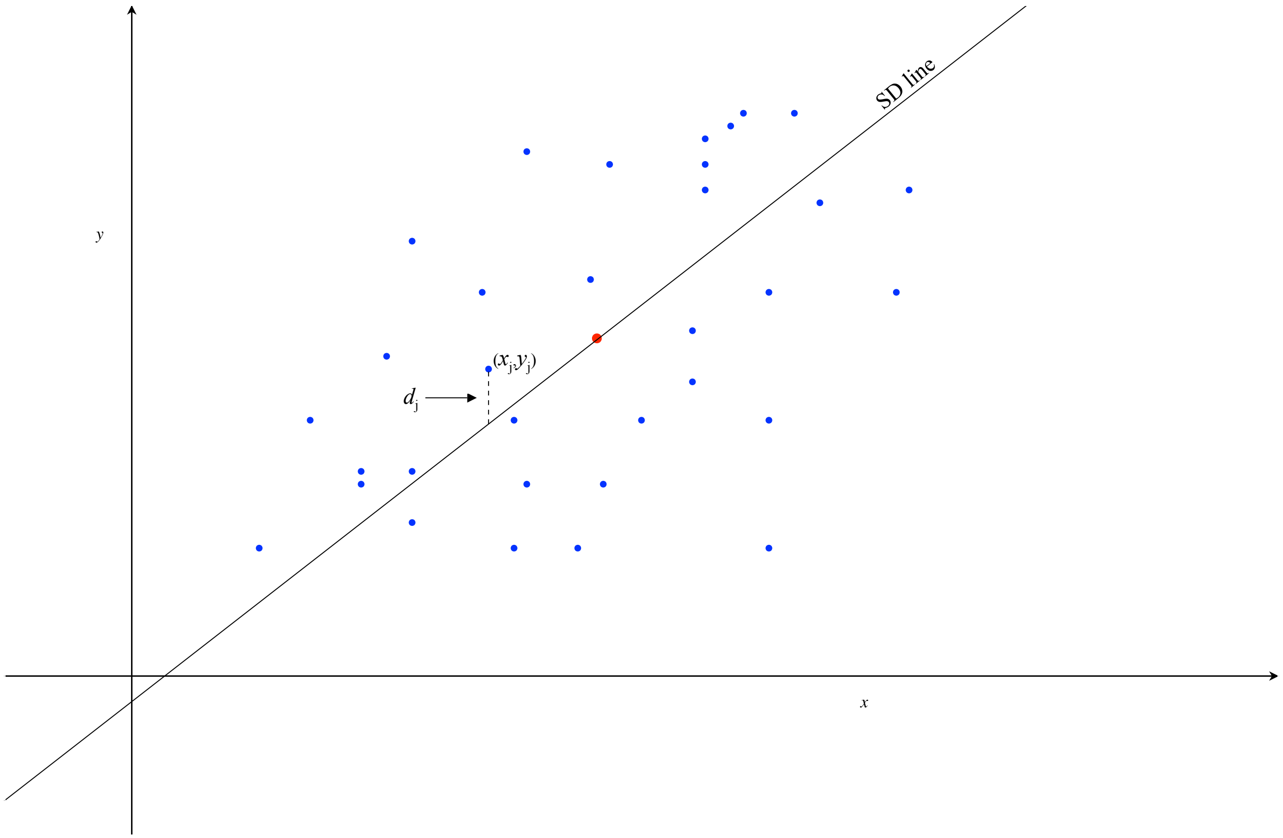
On this line, y, if x increases by one SD_x, then y increases by one SD_y (if $r > 0$) or y decreases by one SD_y (if $r < 0$).

(*) Roughly speaking, the SD-line runs down the middle of the 'cloud' of points, in the 'long' direction.



(*) The red dot represents the *point of averages*.

(*) The correlation coefficient gives a measure of how the data is clustered around the SD-line. For each point (x_j, y_j) in the scatter plot, find the *vertical* distance d_j of (x_j, y_j) from the SD-line...



... then the clustering around the SD-line is quantified by $\sqrt{\frac{1}{n}\sum_j d_j^2}$.

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

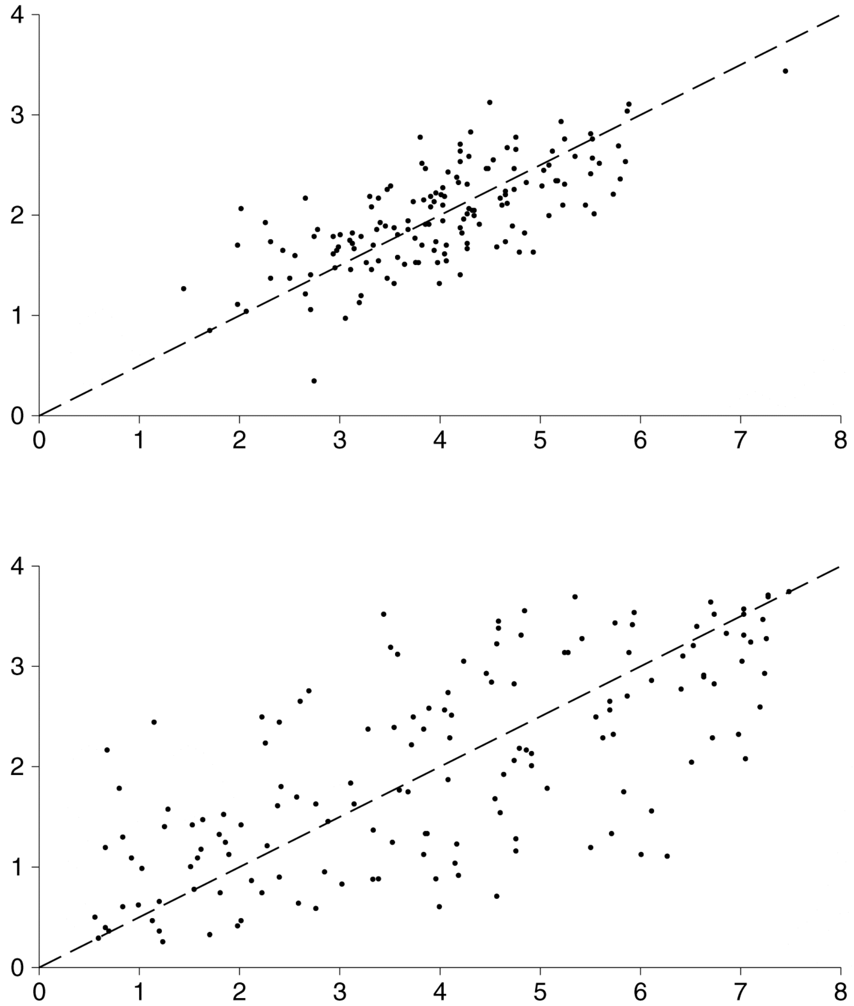
$$\sqrt{\frac{1}{n}\sum_j d_j^2} = \sqrt{2(1-|r_{xy}|)} \times SD_y.$$

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

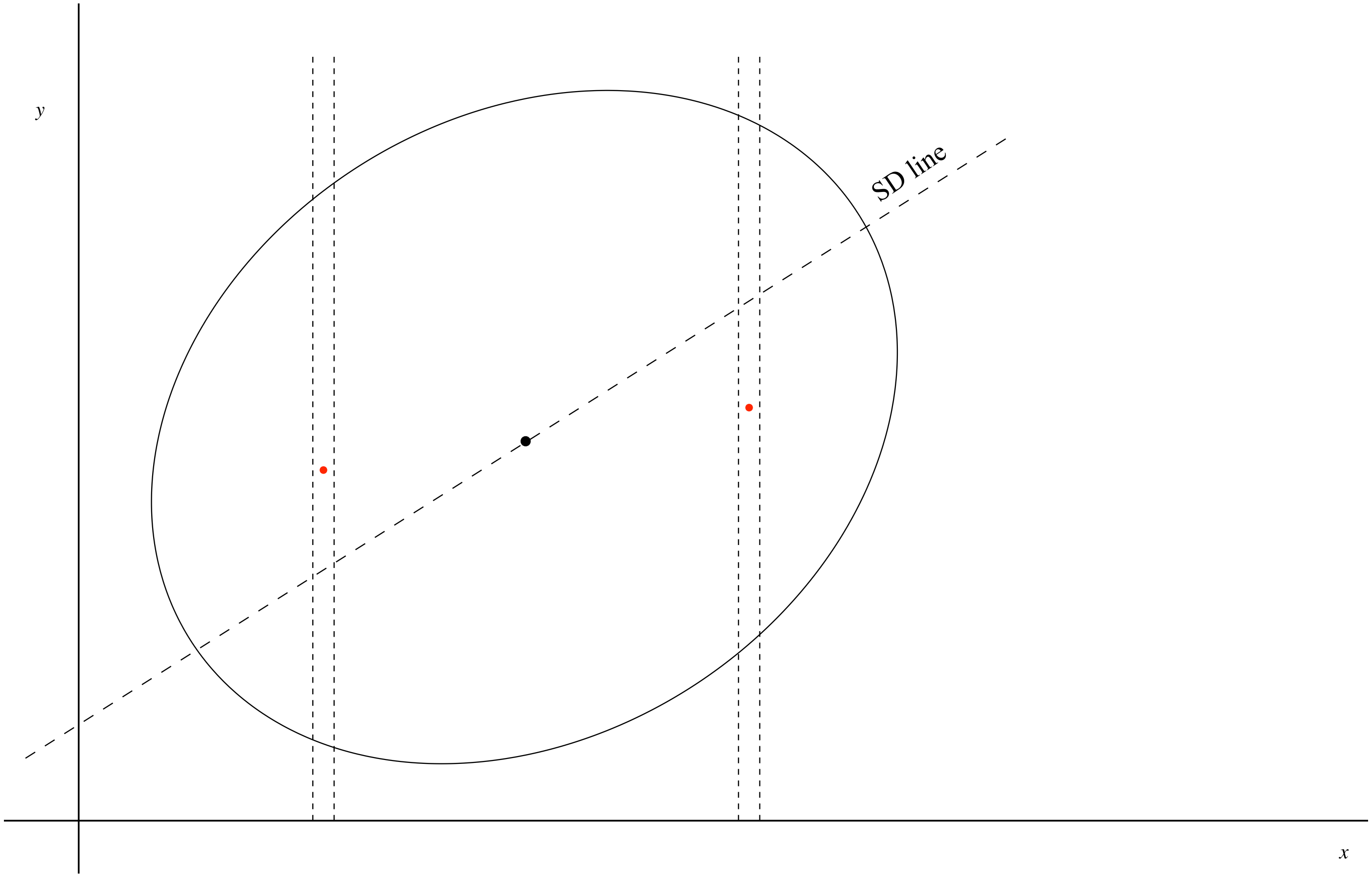
(*) The closer $|r_{xy}|$ is to 1, the more tightly clustered the data will be around the SD line. But this measure of clustering around the SD-line depends on both r_{xy} ***and*** SD_y.

⇒ 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).

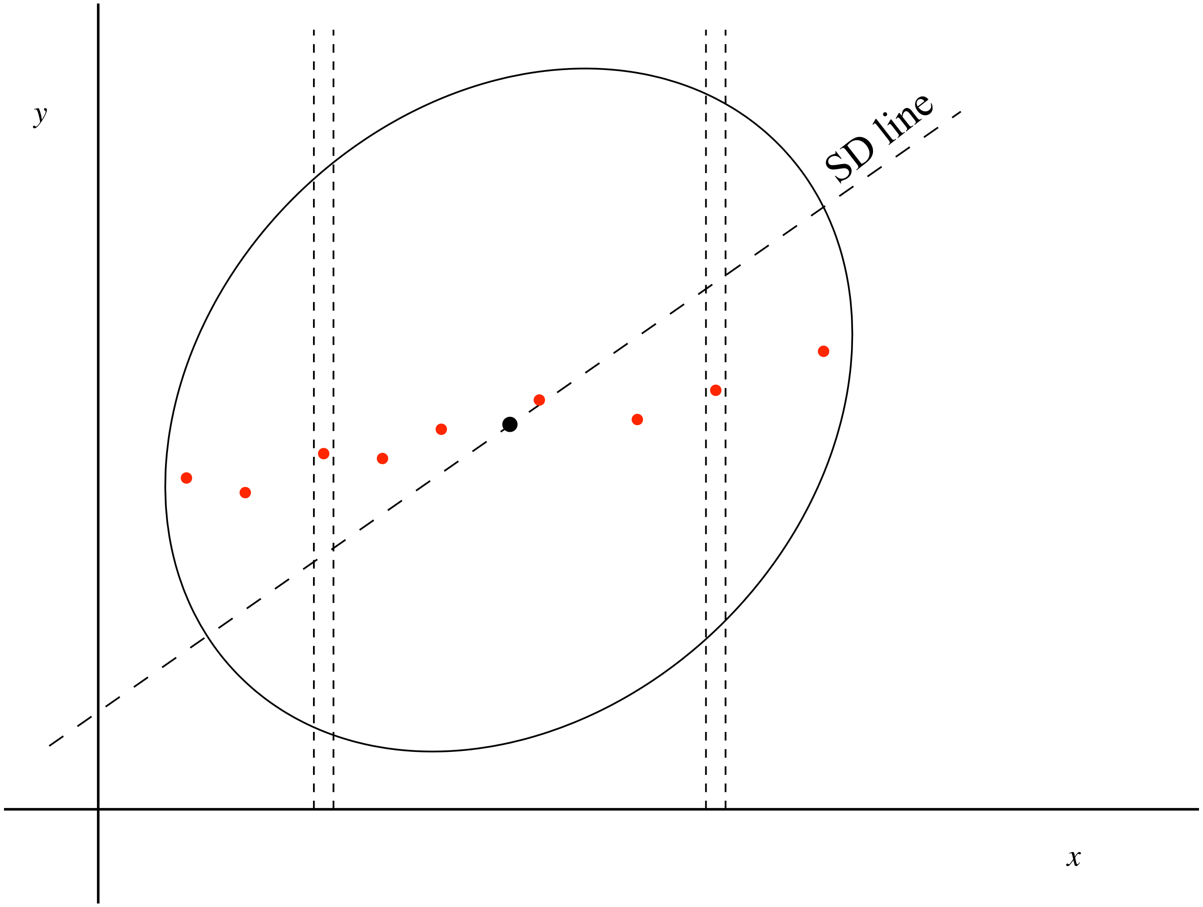
Figure 3. The effect of changing SDs. The two scatter diagrams have the same correlation coefficient of 0.70. The top diagram looks more tightly clustered around the SD line because its SDs are smaller.



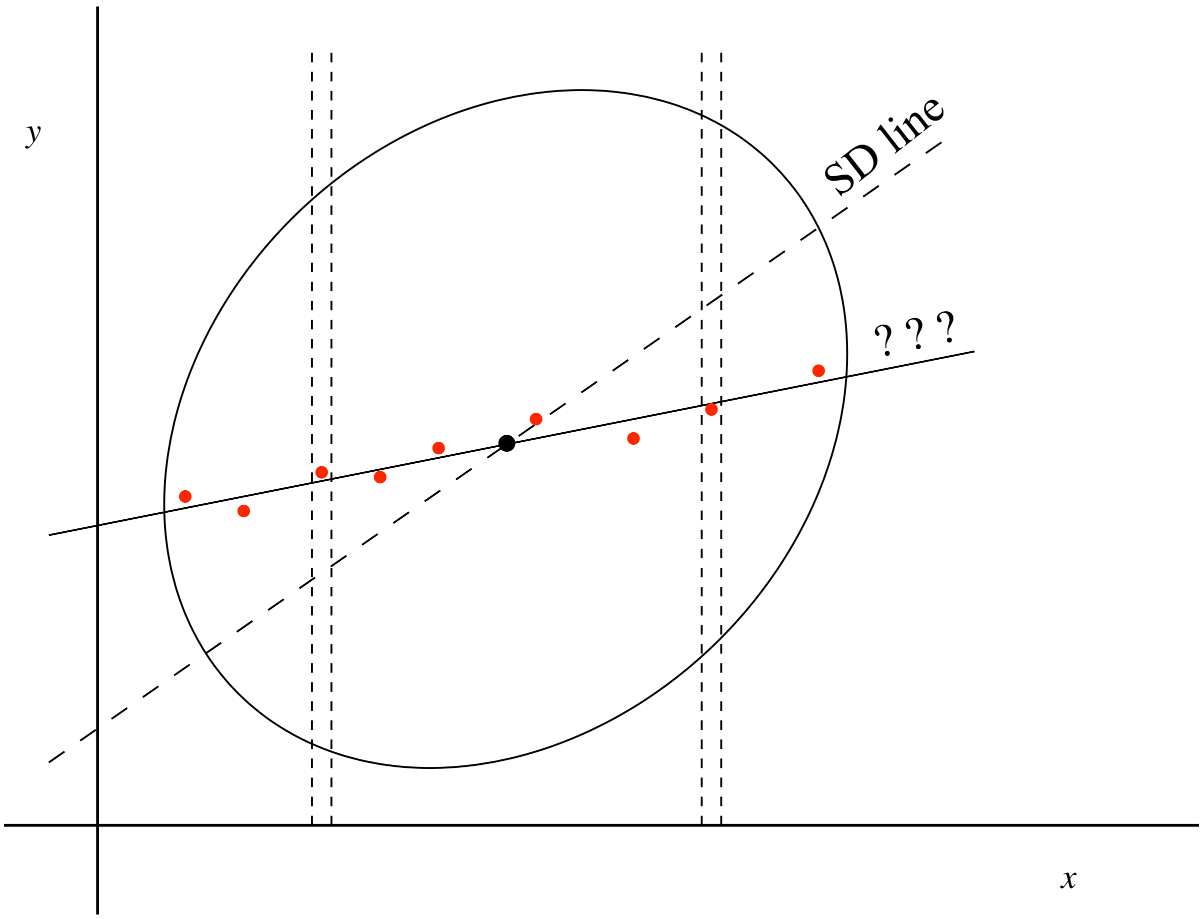
Question: How well does the SD-line approximate the averages $\overline{y}(x_j)$?



Not so well: 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 well as possible.

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

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

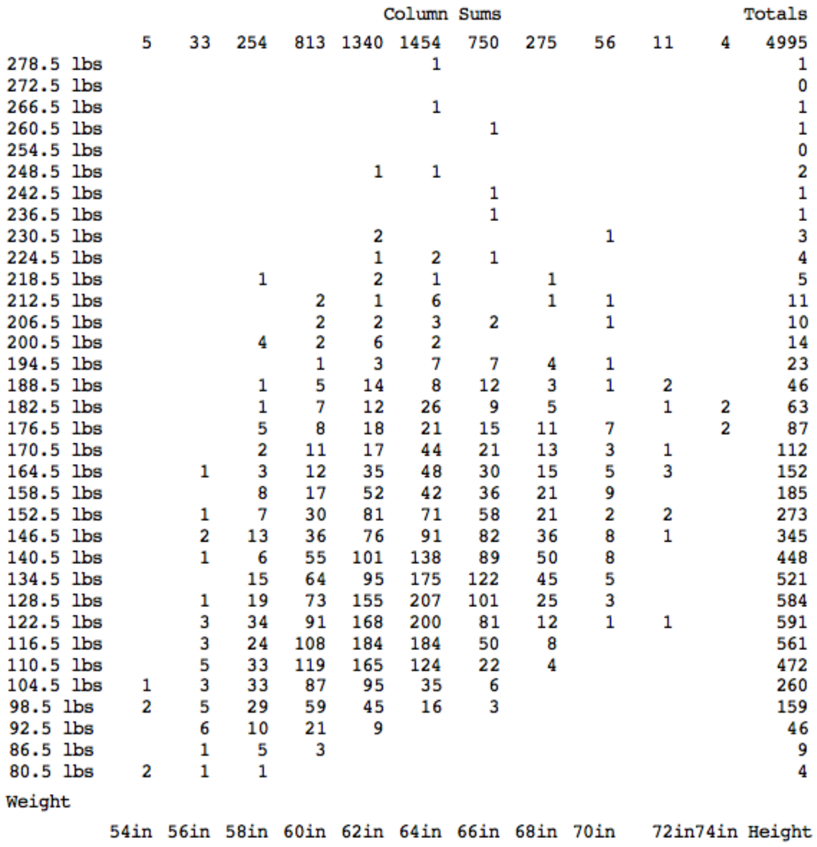
(*) Taking correlation into account leads to the ***regression*** line.

- 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 one SD_x change in x, there is an approximate one $r_{xy} \cdot SD_y$ change in the ***average value*** of the corresponding y s.

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

$$\overbrace{0.32}^{r} \times \overbrace{1.11}^{\#SD_h} \approx 0.355\, SD_w \text{ above the average weight.}$$

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 the overall average blood pressure — 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.

Question: How accurate is this estimate likely to be?

To answer this question we need to have an idea of the ***spread*** around the average in the vertical strip corresponding to $x = 30$.

Example. Students in a certain kindergarten class are given an IQ test in the fall and then again in the Spring. Researchers want to know if the academic program in this kindergarten helps boost the children's IQ.

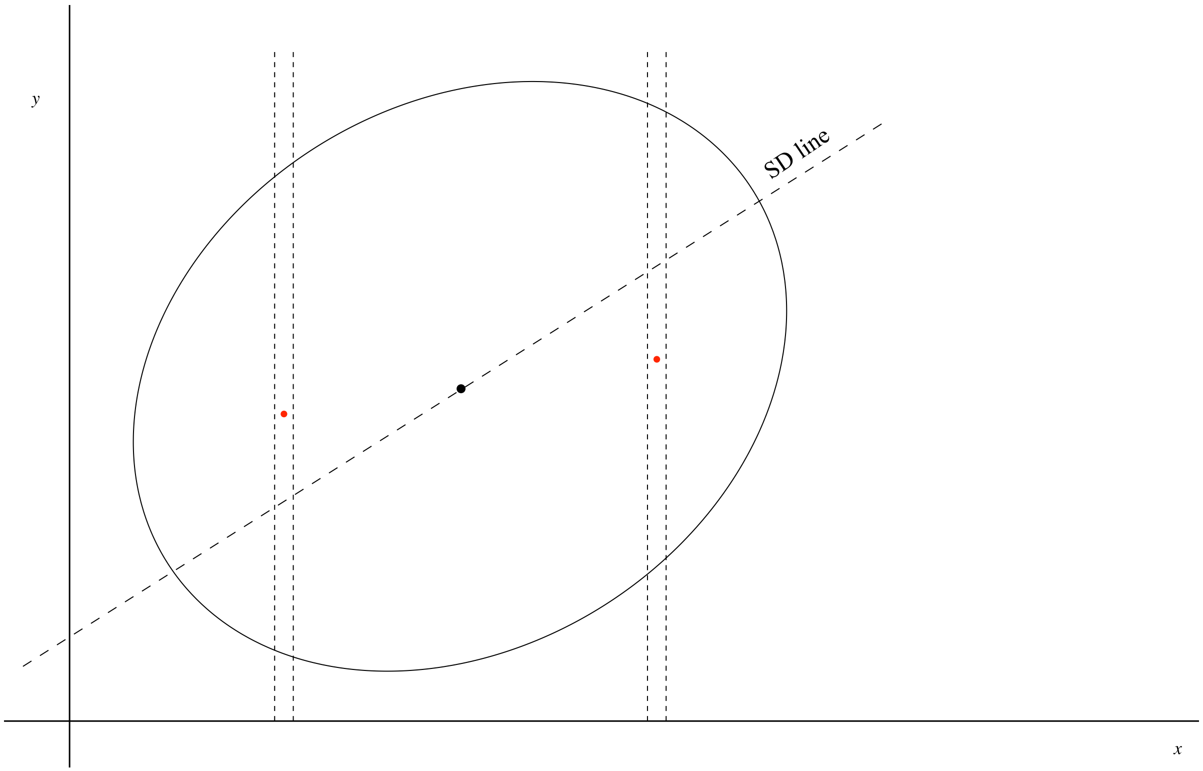
(*) The average on both tests is about 100 and both SD s are about 15, so at first glance it seems that a year of kindergarten had no overall effect.

(*) A closer look at the data finds shows that students with high scores on the first test, tended to have lower scores on the second test, on average. Also, students with lower scores on the first test did better, on average, on the second test.

(*) Why?

(*) Suppose that the relation between x and y is positive.

The ***regression effect*** is caused by the spread around the SD line: if x_j is one SD_x above $\overline{x}$, then y_j will be greater than $\overline{y}$, but only by $r \times SD_y$ on average. If x_j is one SD_x below $\overline{x}$, then y_j will be less than $\overline{y}$ on average, but only by $r \times SD_y$.



The regression effect – a famous example.

Example: Heights of sons on heights of fathers.

$$\text{average height of fathers} \approx 68 \text{ inches}, \quad \text{SD} \approx 2.7 \text{ inches}$$

$$\text{average height of sons} \approx 69 \text{ inches}, \quad \text{SD} \approx 2.7 \text{ inches} \quad r \approx 0.5$$

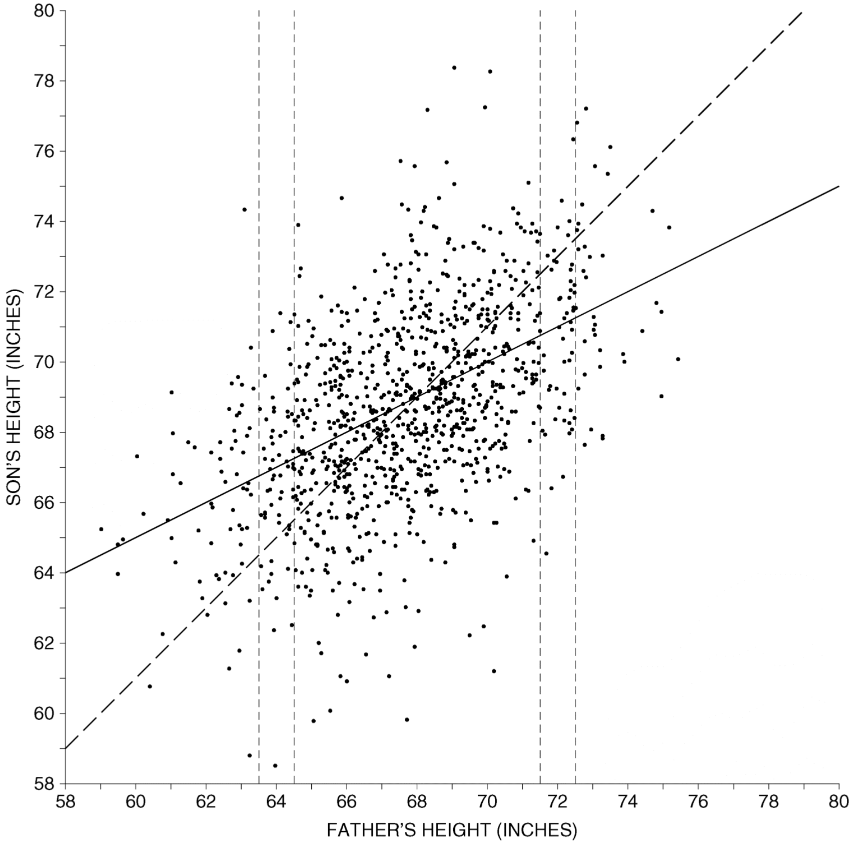


Figure 5., p.171 in FPP, sons and fathers' heights, with SD line and regression line.

- The average heights of the sons for each height class of the fathers follow the regression line, not the SD line.
- The average height of the sons grows more slowly than the height of their fathers.
- Fathers that are much taller than 70 inches, will have sons that are, on average, shorter than them.
- Fathers that are shorter than 70 inches will have sons that are, on average, taller than them.
- The same logic applies, for example to of test-retest scenarios: higher than average scores on the first test will be followed by somewhat lower scores on the second test, on average. Likewise, lower than average scores on the first test will be followed by somewhat better scores on the second test, on average.
- The belief that the regression effect is anything more than a statistical fact of life is the ***regression fallacy***.

Example: A study of education and income is done for men age 30 - 35. A representative sample of 10000 men in this age group is surveyed, and the following statistics are collected:

$$\overline{E} = 13 \quad SD_E = 1.5$$

$$\overline{I} = 46 \quad SD_I = 10 \quad r_{\mathrm{I,E}} = 0.45$$

where E = years of education and I = annual income, in \$1000s.

(*) A 32-year old man is observed, our best guess for his income is ...

... the average income for all men in this age group, $\overline{I} = 46$.

(*) A 32-year old man with 16 years of education is observed, our best guess for his income is ...

... the average income of all men in this age group with 16 years of education:

$$\overline{I}(16) \approx 46 + 0.45 \times \frac{16-3}{1.5} \times 10 = 55.$$

Example: A study of education and income is done for men age 30 - 35. A representative sample of 10000 men in this age group is surveyed, and the following statistics are collected:

$$\overline{E} = 13 \quad SD_E = 1.5$$

$$\overline{I} = 46 \quad SD_I = 10 \quad r_{\mathrm{I,E}} = 0.45$$

where E = years of education and I = annual income, in \$1000s.

(*) A 32-year old man is observed, our best guess for his income is ...

... the average income for all men in this age group, $\overline{I} = 46$.

(*) A 32-year old man with 16 years of education is observed, our best guess for his income is ...

... the average income of all men in this age group with 16 years of education:

$$\overline{I}(16) \approx 46 + 0.45 \times \frac{16 - 3}{1.5} \times 10 = 55.$$

(*) How can we make our estimates more informative?

(*) We can make our first answer more informative by saying that most men in this age group (probably) have incomes in the range $\overline{I} \pm SD_I = 46 \pm 10$.

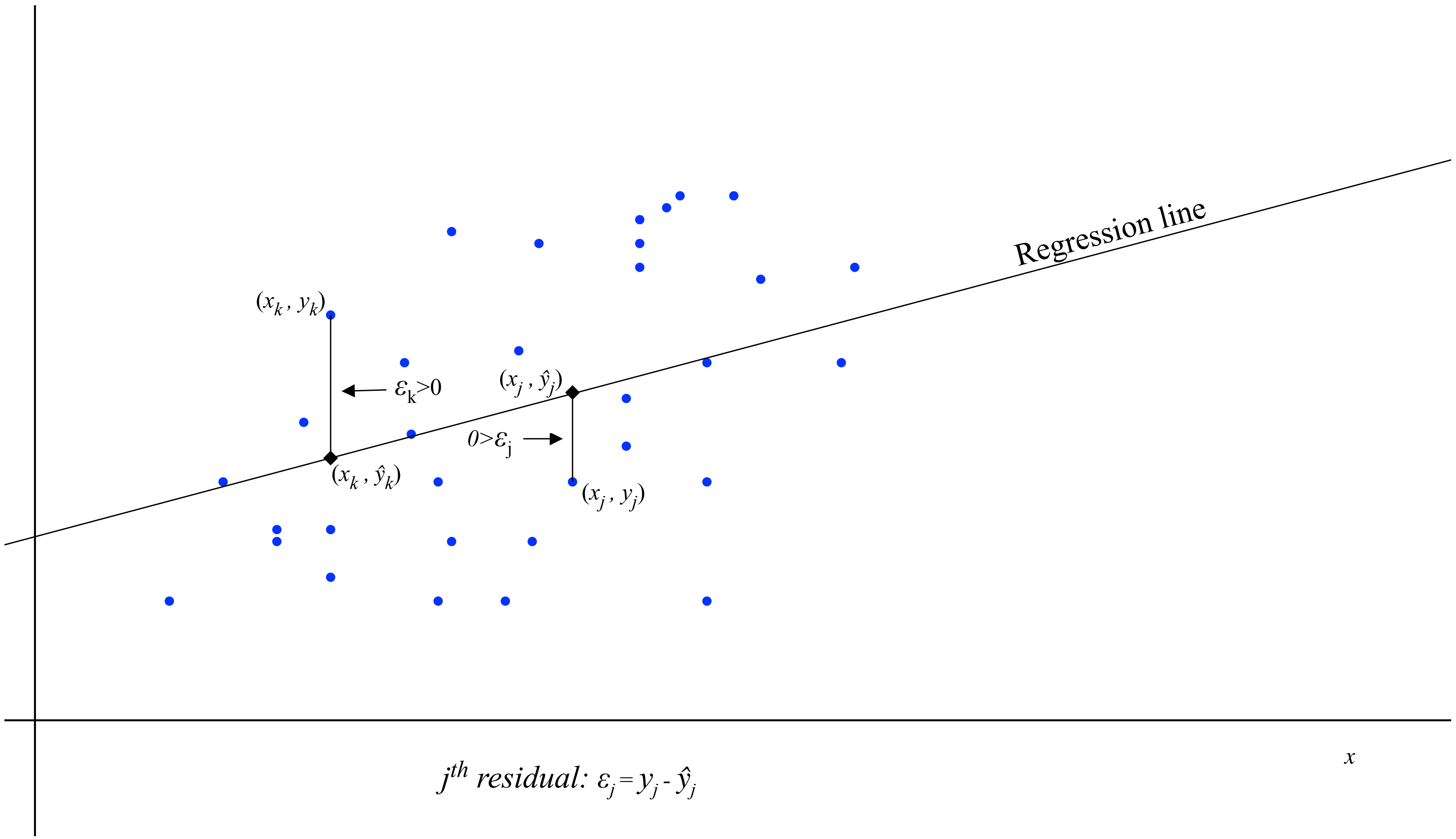
(*) We would like to do the same for the second estimate, when we know the man's level of education.

(*) To do this, we need an estimate for the SD of the incomes for all men in this study with 16 years of education.

⇒ With access to all the data, we could isolate the men with 16 years of education and calculate the average and SD of their incomes (in which case, we wouldn't need the regression method either!).

(*) Question: is there some way to generate an estimate for the SD we want using only the five basic statistics

$$\overline{E}, \quad SD_E, \quad \overline{I}, \quad SD_I \;\text{ and }\; r_{\mathrm{I,E}}?$$



j^{th} residual: $\varepsilon_j = y_j - \hat{y}_j$

The “Root-Mean-Square” error of regression (also called the *Standard error of regression*, SER) is given by

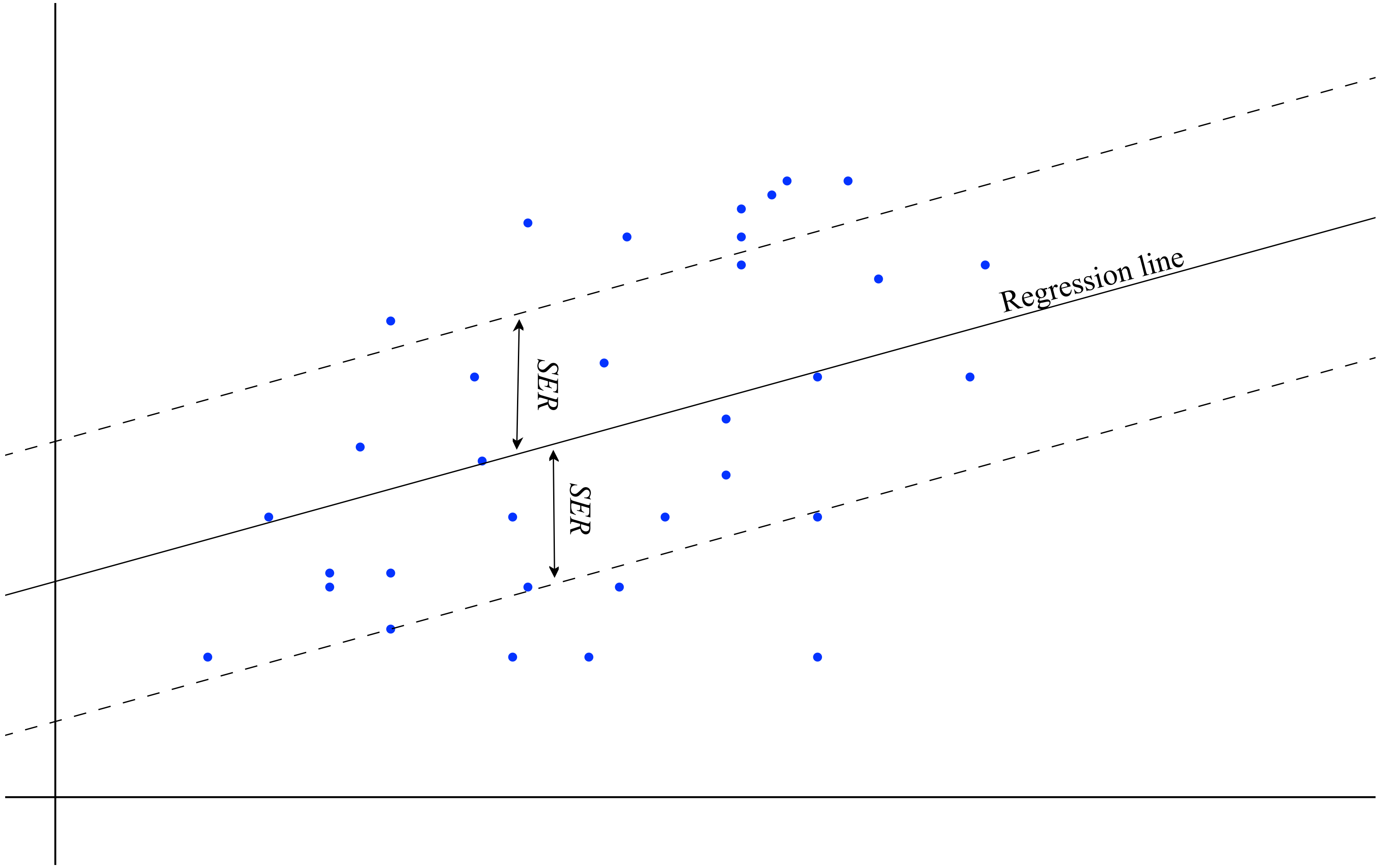
$$SER = \sqrt{\frac{1}{n}\sum_j \varepsilon_j^2} = \sqrt{\frac{1}{n}\sum_j (y_j - \widehat{y}_j)^2},$$

where

- y_j is the y-value of the j^{th} observation,
- $\widehat{y}_j$ is the predicted value of y_j (by the regression line),
- $\varepsilon_j = y_j - \widehat{y}_j$ is the j^{th} residual or *prediction error*.

(*) The R.M.S. Error for regression (SER) is like an SD for the points in a scatter plot, with the regression line playing the role of the average.

Rule of thumb: Most of the points in a scatter plot lie in a strip around the regression line — from one SER below the regression line to one SER above the regression line.



(*) Back to the question of finding an SD for the set of y_js of all observations with the same x_j....

For certain types of paired data, the SER gives a good approximation of the SD in every vertical strip

In other words we can use the regression estimate together with the SER to give more informative predictions for y_j given x_j.

(*) The regression method says that $\overline{y}(x_j) \approx \widehat{y}_j(x_j)$. This gives the 'point estimate' (prediction) $y_j \approx \overline{y}(x_j) \approx \widehat{y}_j(x_j)$.

(*) The SER adds a margin of error, and we can say that ...

... y_j is likely to fall in the range $\widehat{y}_j \pm SER$.

This becomes particularly useful because there is a convenient shortcut for the SER:

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