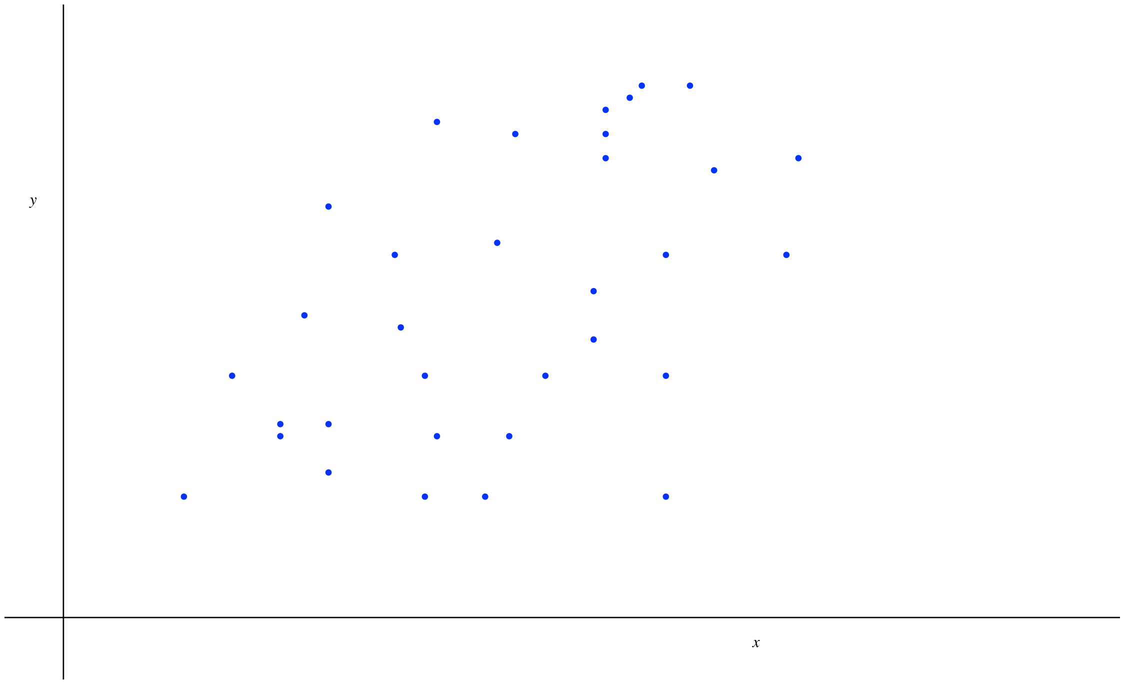
Linear Regression

(*) 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.

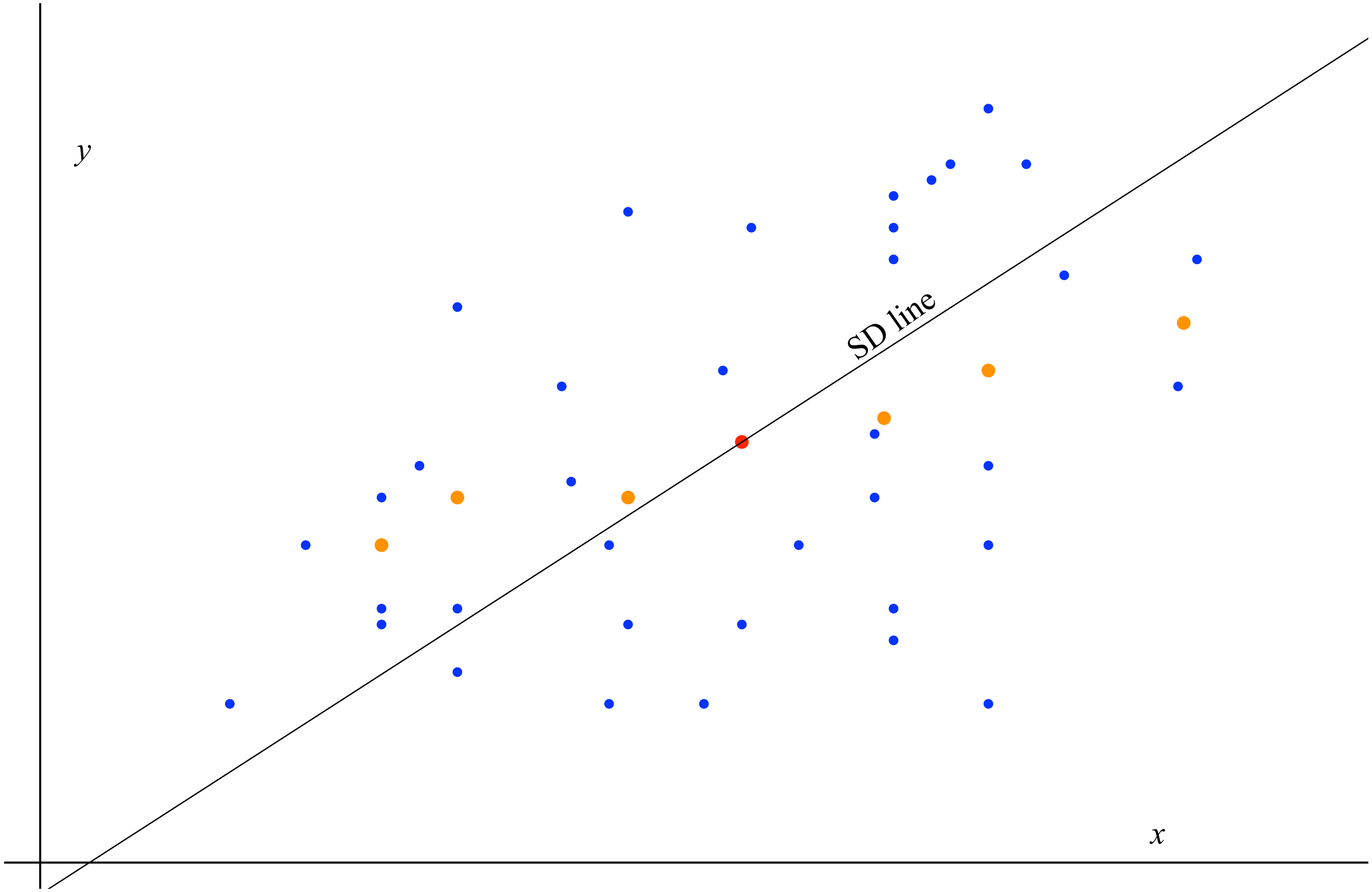
Notation:

- $\overline{y}(x_j)$ = 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.

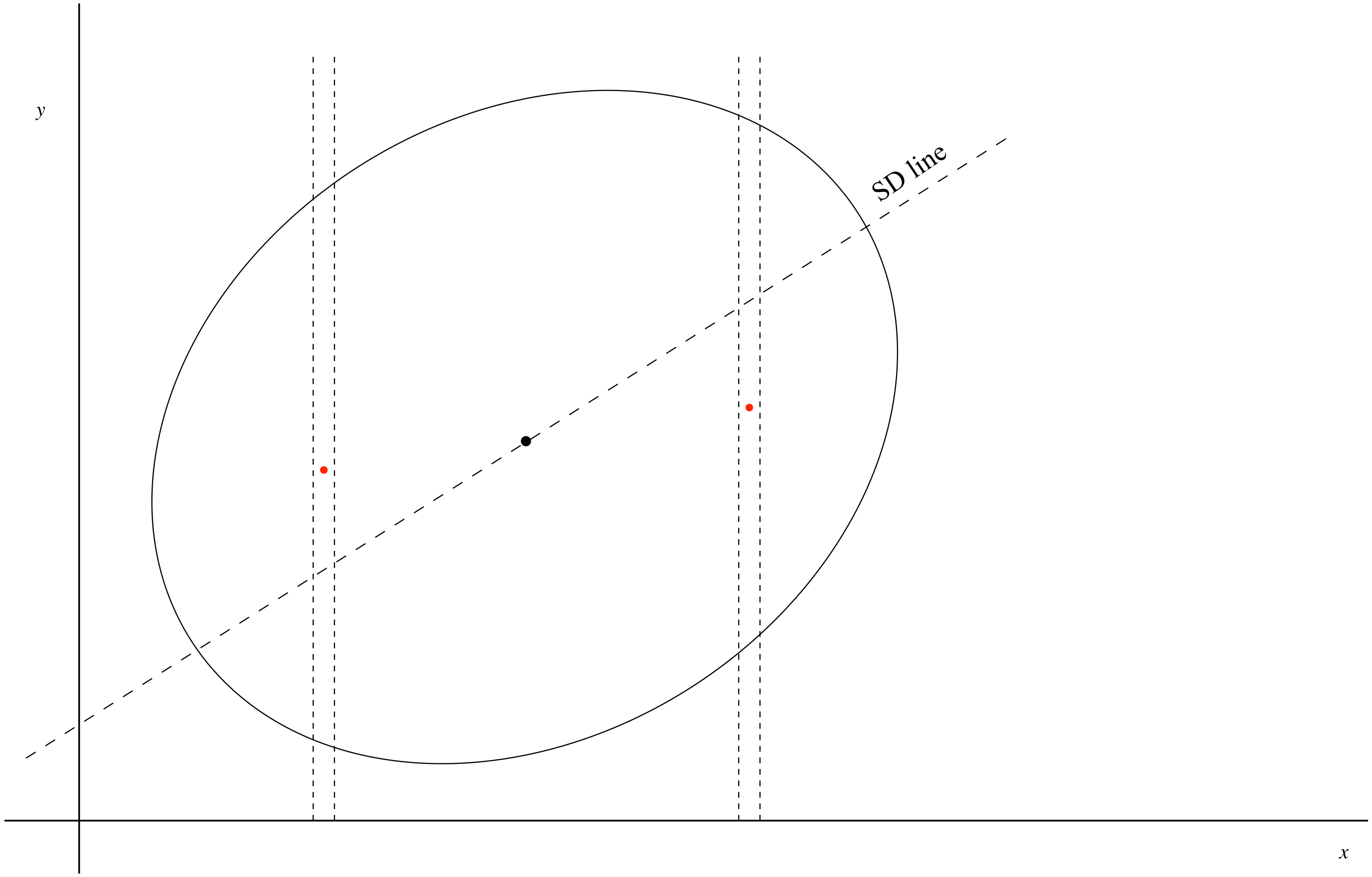
(*) First Guess: *The SD-line.*

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

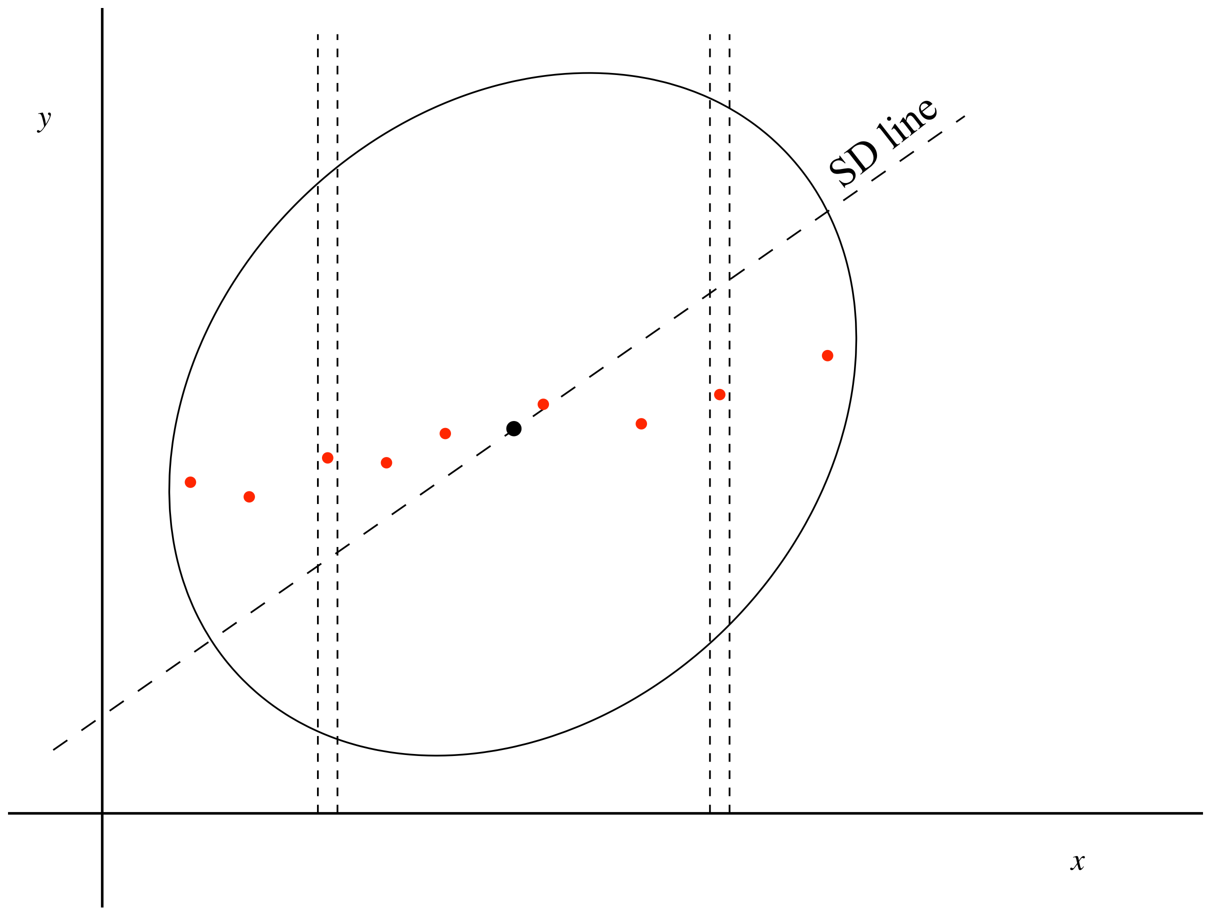


Not so well: The SD line is *underestimating* the averages to the left of the point of averages and *overestimating* the averages to the right of the point of averages.

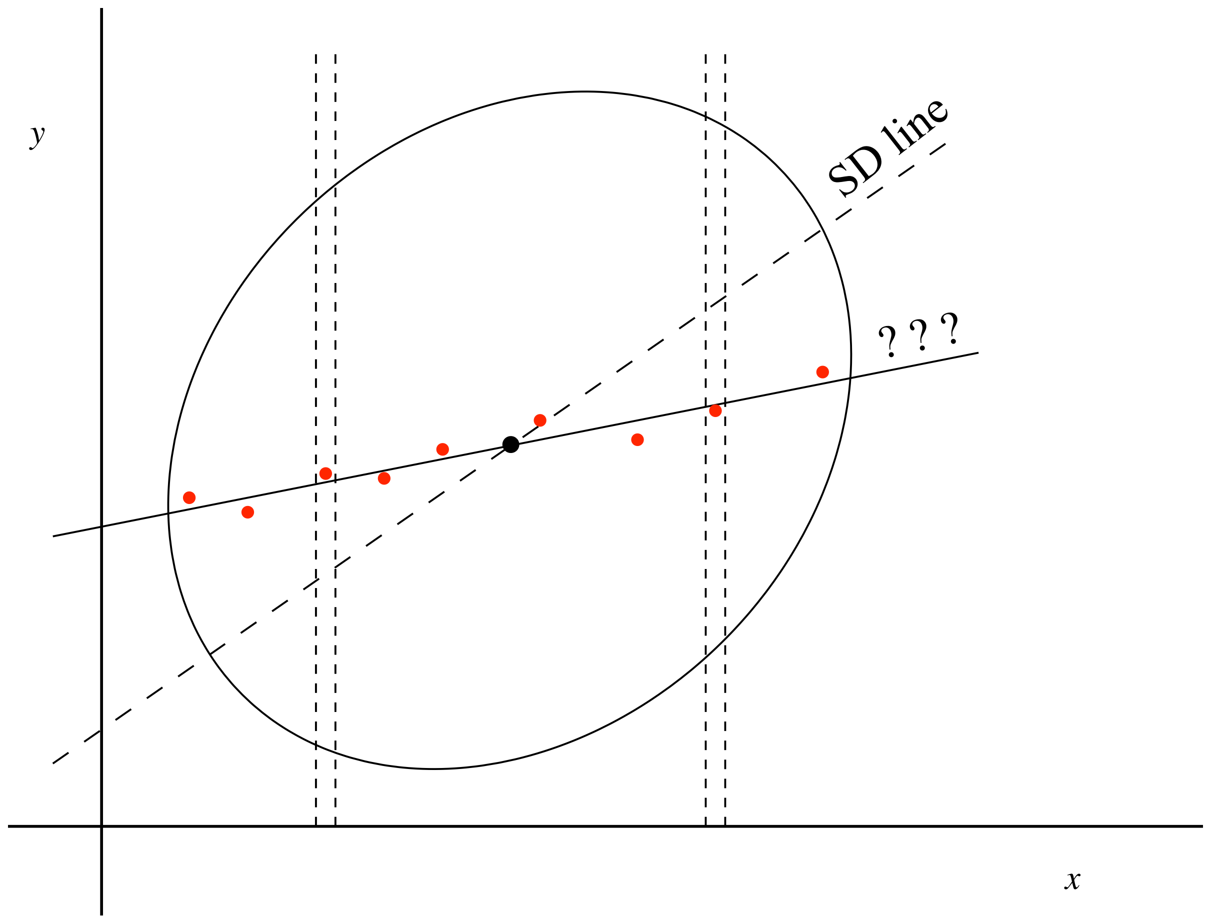
It may be easier to see the forest if we remove the trees...



The average y-value for each x-value lies near the middle of the *vertical strip* above that x. The further the vertical strip is from the point of averages, the more the SD-line tends to miss the middle of the strip.



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

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. 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 method 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 method 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 for later: What is the margin of error for this estimate?

Question: Kevin is a 53-year old man who doesn't smoke. What is Kevin's predicted blood pressure?

Answer: The data from our (hypothetical) study shouldn't be used to predict Kevin's blood pressure. His age falls outside the range of ages for which the study was done.

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.

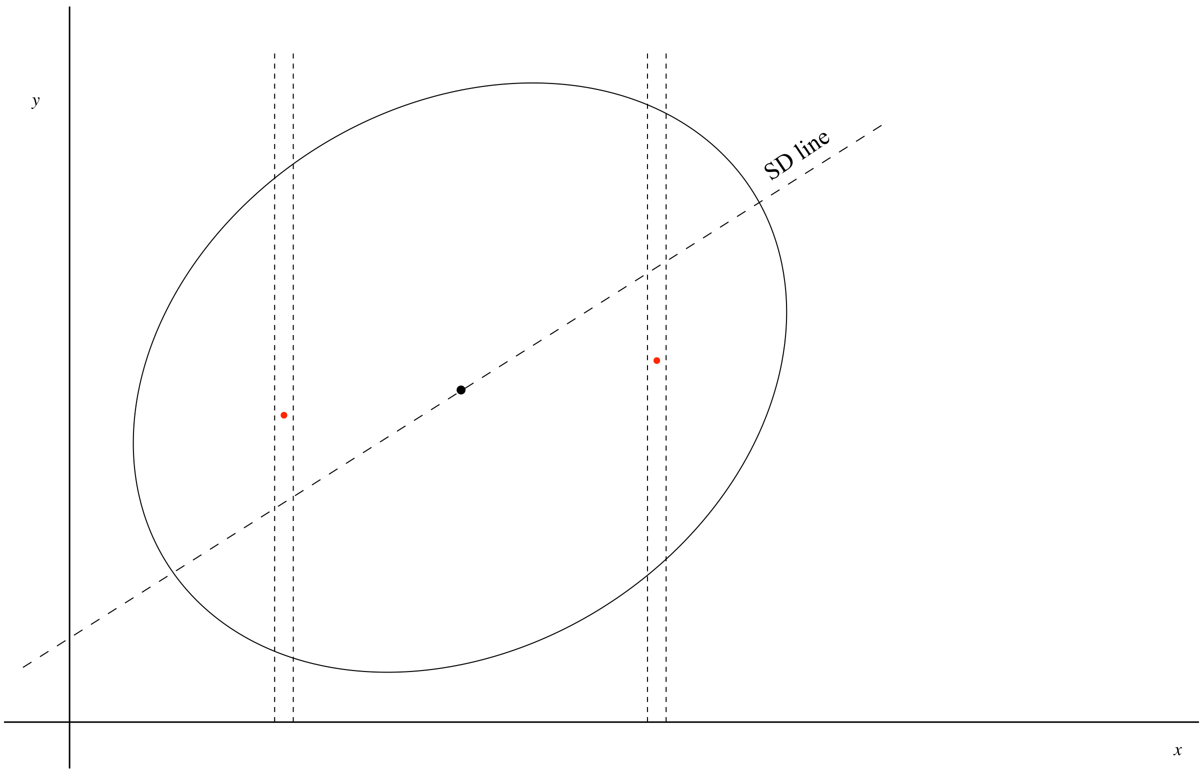
(*) A closer look at the data finds shows that students with high scores on the first test, tended to score lower (than their first score) on the second test, on average. Also, students with lower scores on the first test tended to improve on the second test.

(*) Why?

$\Longrightarrow$ The *regression effect.*

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

The ***regression effect*** is caused by the *vertical* spread of the data 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.

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

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

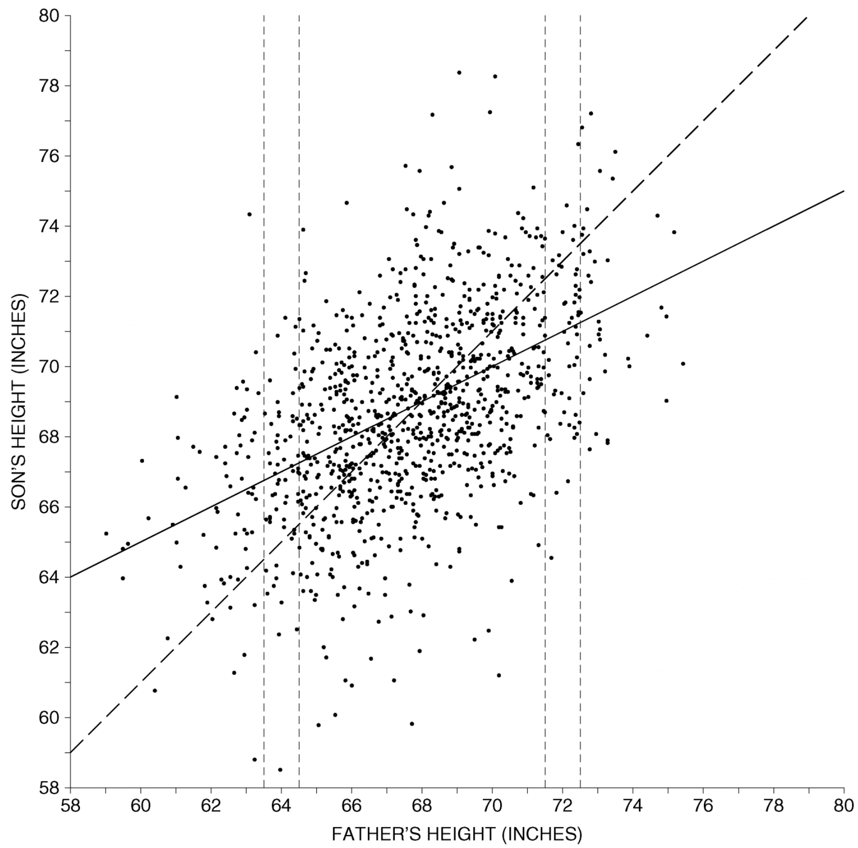


Figure 5., p.171 in FPP, sons and fathers' heights, with SD line and regression line.

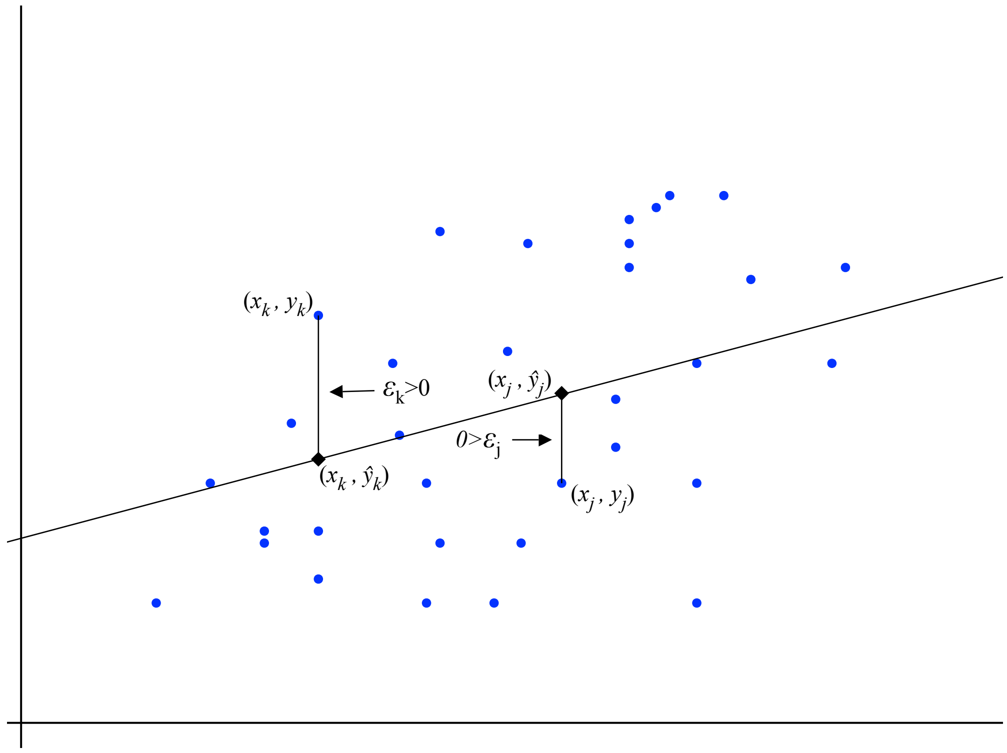
- The average heights of the sons for each height class of the fathers follow the regression line, not the SD line.
- The average height of the sons grows more slowly than the height of their fathers.
- Fathers that are much taller than 70 inches, will have sons that are, on average, shorter than them.
- 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***.

The R.M.S. error of the regression

If we compare y_j to the value $\hat{y}_j$ predicted by the regression method, then we will typically observe an error

$$\varepsilon_j = y_j - \hat{y}_j.$$

These errors may be positive or negative and reflect the fact that the data does not lie exactly on any straight line.



The square-root of the average of the squares of these errors,

$$\sqrt{\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j^2} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

is called the *R.M.S. error of the regression.*

- The *R.M.S. error of the regression* is roughly the average *vertical* distance of points in the scatterplot to the regression line.
- The R.M.S. error of regression is also called the ***standard error of regression***, or ***SER***.
- The SER is to the regression line what the SD is to the average. Most of the data in the scatter plot will be within one or two SER's from the regression line.

Using the SER

When using the regression equation to predict an *individual* y-value from an observed x-value, we can say that y is likely to be within one (or two) SER(s) of $\hat{y}(x)$.

Typically, we want to have actual numbers, so it is nice to know that there is a **shortcut** for computing the SER:

$$SER = \sqrt{1 - r_{xy}^2} \cdot SD_y.$$

Example: Returning to the '*question for later*' In the Smoking-BP example, we have $r = 0.64$ and $SD_y = 7$, so

$$SER = \sqrt{1 - (0.64)^2} \cdot 7 \approx 5.38.$$

I.e., in the case of the 31-year old man who smokes 30 cigarettes a day, we can now say that his blood pressure is likely to be in the range

$$136.28 \pm 5.38 \text{ mmHg}.$$