UPCSE Biology – Basic Statistic course

Scatter plots, lines of best fit, and correlation

Introduction

- Last week we looked at bar charts and histograms. Today we look at a third form of data visualisation known as scatter plots.
- Here we study the relationship between two data sets, specifically between two variables.

Introduction

- Examples of this include:
  - Percentage germination of seeds vs amount of rainfall;
  - Bone density vs age;
  - Blood cholesterol level vs heart disease;
  - Plasma concentration of a drug vs time after start of intravenous release;
  - … anything where we can measure y vs x.
- Data connecting two variables is called *bivariate data*
- Such data is usually presented as a table.

Introduction

- Our first job with such data will be to plot it on an $x-y$ graph in order to see what the trend looks like.
- From this we decide if we can approximate the trend of the data by a straight line or not.
- The following examples are designed only to practice recognising if the trend is linear or not.

Scatter plots

Example 1

- Consider the lengths and areas of nine privet leaves

Length (mm)	Area (mm^2)
8	25
12	48
18	124
20	162
26	220
33	315
38	420
40	515
45	650

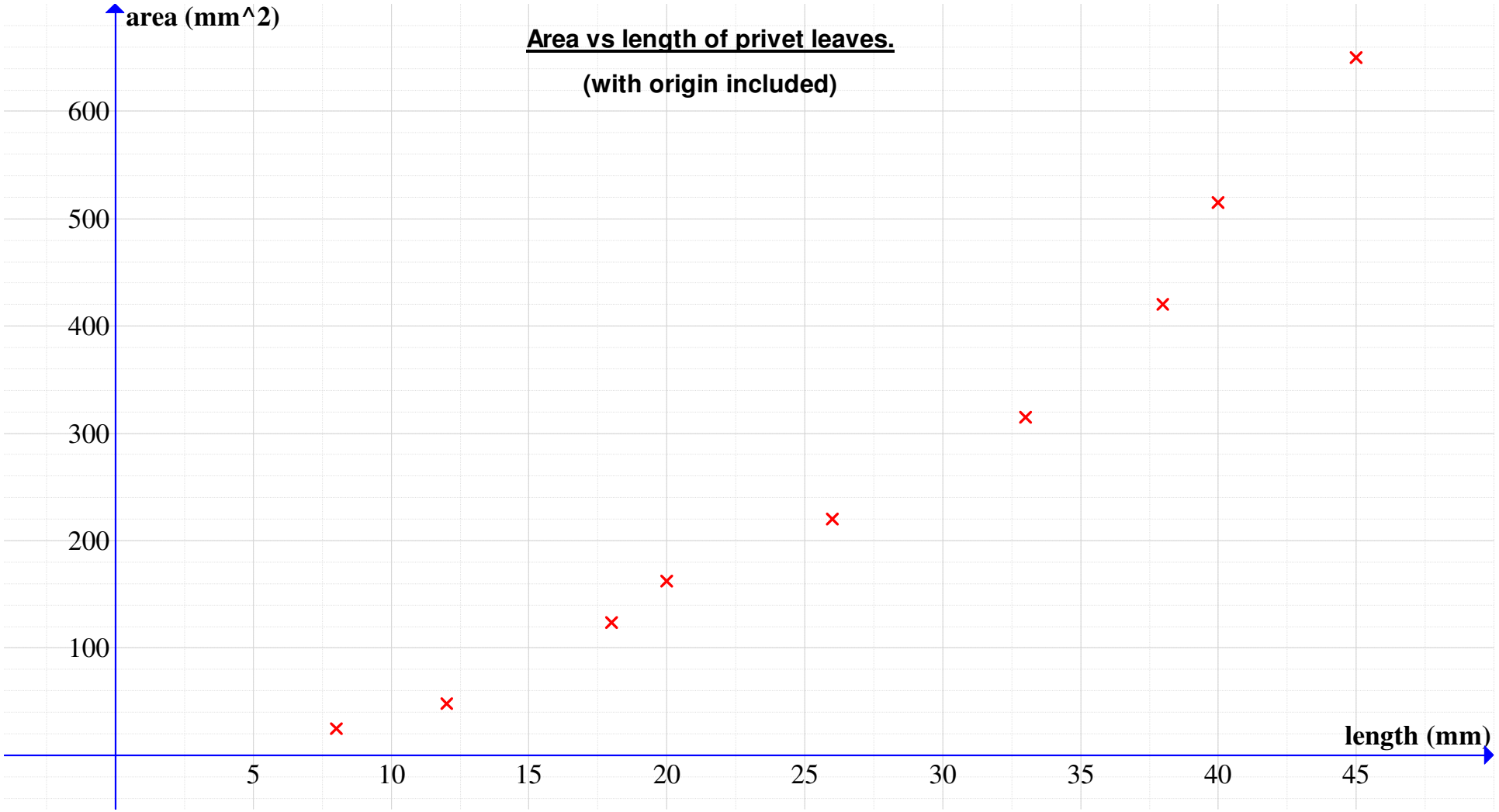
Scatter plots

Example 1

- We want to see if there is a correlation between length and area.
- To do this plot a standard x–y graph.

Length (mm)	Area (mm²)
8	25
12	48
18	124
20	162
26	220
33	315
38	420
40	515
45	650

Scatter plots



Scatter plots

- From the graph we can see that there is a linear trend …
- … so it will be worth using a best fit line for this data.

Scatter plots

Example 2

- Consider the concentration of O_2 in seawater and freshwater at different temps.

Temp (^{0}C)	O_2 in freshwater (ppm)	O_2 in seawater (ppm)
1	14.0	11
10	11.5	9.0
15	10.0	8.0
20	9.0	7.5
25	8.0	7.0
30	7.5	6.0

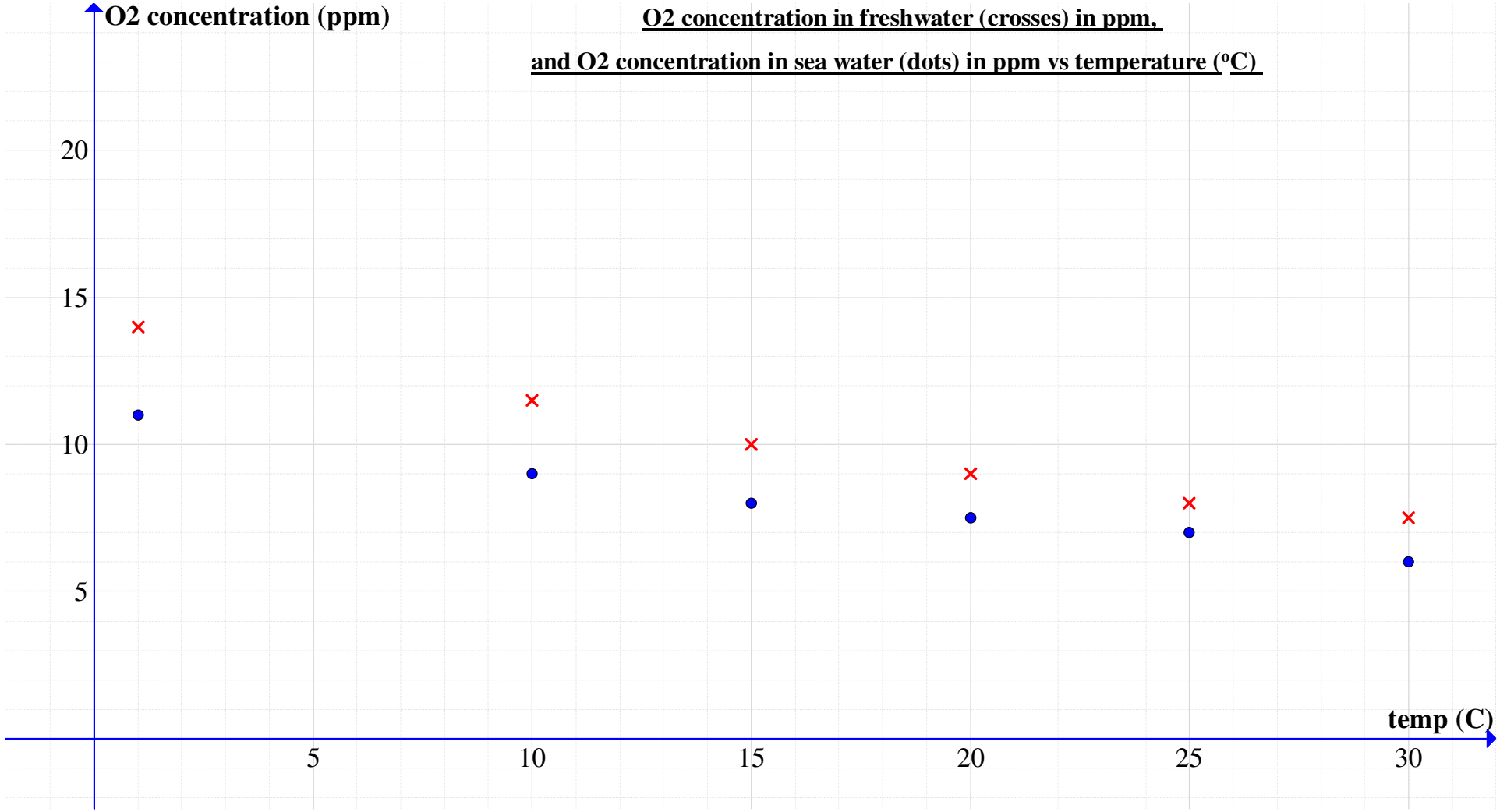
Scatter plots

Example 2

- We want to see if there is a correlation between temp and water type.
- To do this plot a standard x–y graph.

Temp (^{0}C)	O_2 in freshwater (ppm)	O_2 in seawater (ppm)
1	14.0	11
10	11.5	9.0
15	10.0	8.0
20	9.0	7.5
25	8.0	7.0
30	7.5	6.0

Scatter plots



Scatter plots

- From the graph we can see that both sets of data have an approximately linear trend …
- … so it will be worth using a best fit line for this data.

Scatter plots

Example 3

- Consider the weight and length of randomly selected newborn turtles.

- We want to see if there is a correlation between length and weight.

Length l (mm)	Weight w (g)
49	29
52	32
53	34
54.5	39
54.1	38
53.4	35
50	30
51.6	31
49.5	29
51.2	30

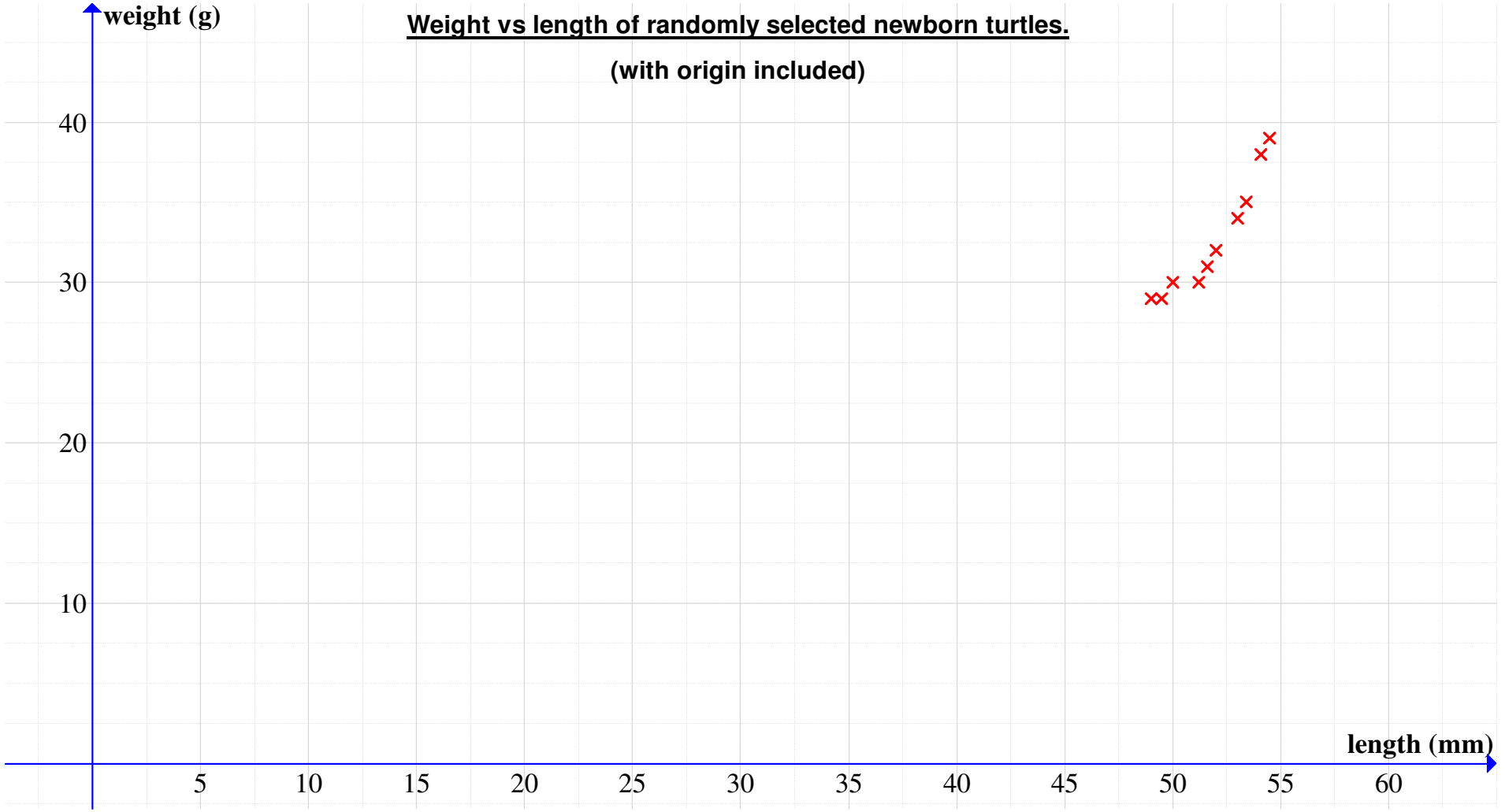
Scatter plots

Example 3

- To do this plot a standard x–y graph.

Length l (mm)	Weight w (g)
49	29
52	32
53	34
54.5	39
54.1	38
53.4	35
50	30
51.6	31
49.5	29
51.2	30

Scatter plots

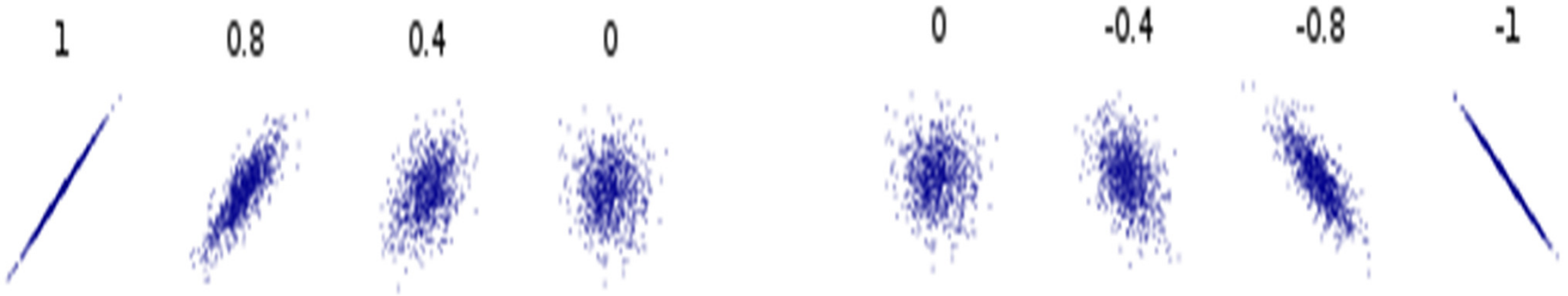


Scatter plots

- From the graph we can see that there is a linear trend …
- … so it will be worth using a best fit line for this data.

Scatter plots

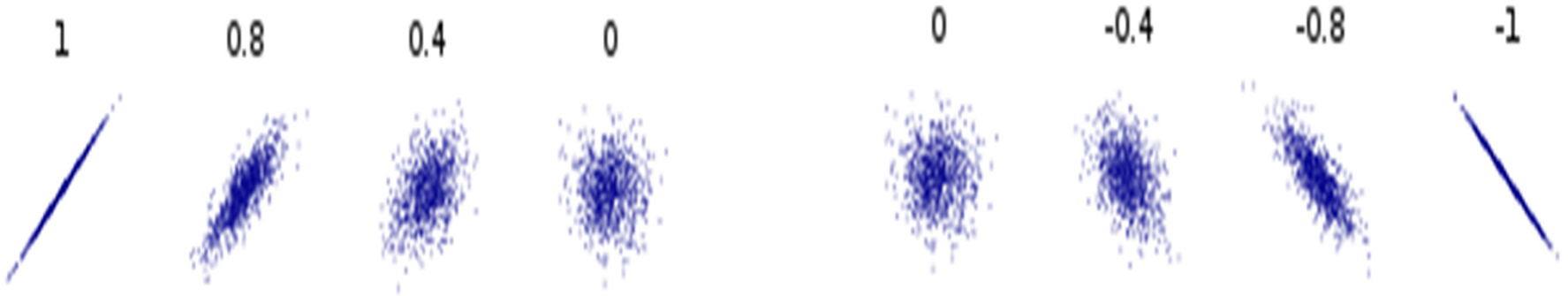
- In general we have the following types of scatter plots.



- The numbers at the top of each scatter plot are called correlation coefficients, given by the letter "r".
- The numbers you see relate to the degree to which there is a linear relationship or not.
- Then we can only have $-1 \leq r \leq +1$

Scatter plots

- In general we have the following types of scatter plots.



- ±1 indicates a perfect linear relationship; 0 indicates no linear relationship.
- Scatter plots allow us to see the trend of the data, and the type of connection between two variables: linear or not.

Examples

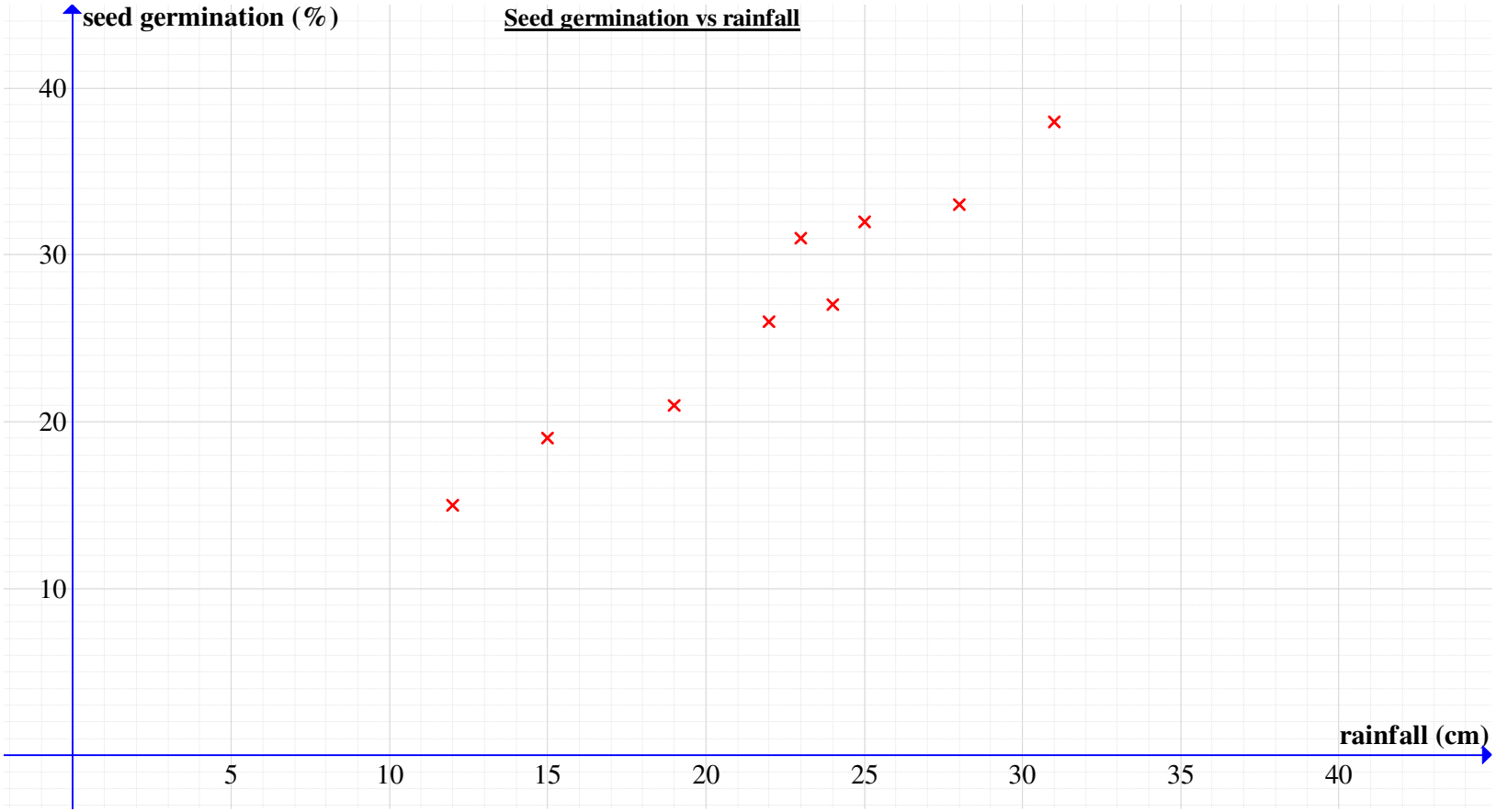
1) Seed germination

The table below shows the percentage of seed germination sown in areas with different amounts of monthly rainfall. Is there a linear relationship between rainfall and % germination? If so find the best fit line.

Rainfall (cm)	12	22	19	15	31	25	28	24	23
Germination (%)	15	26	21	19	38	32	33	27	31

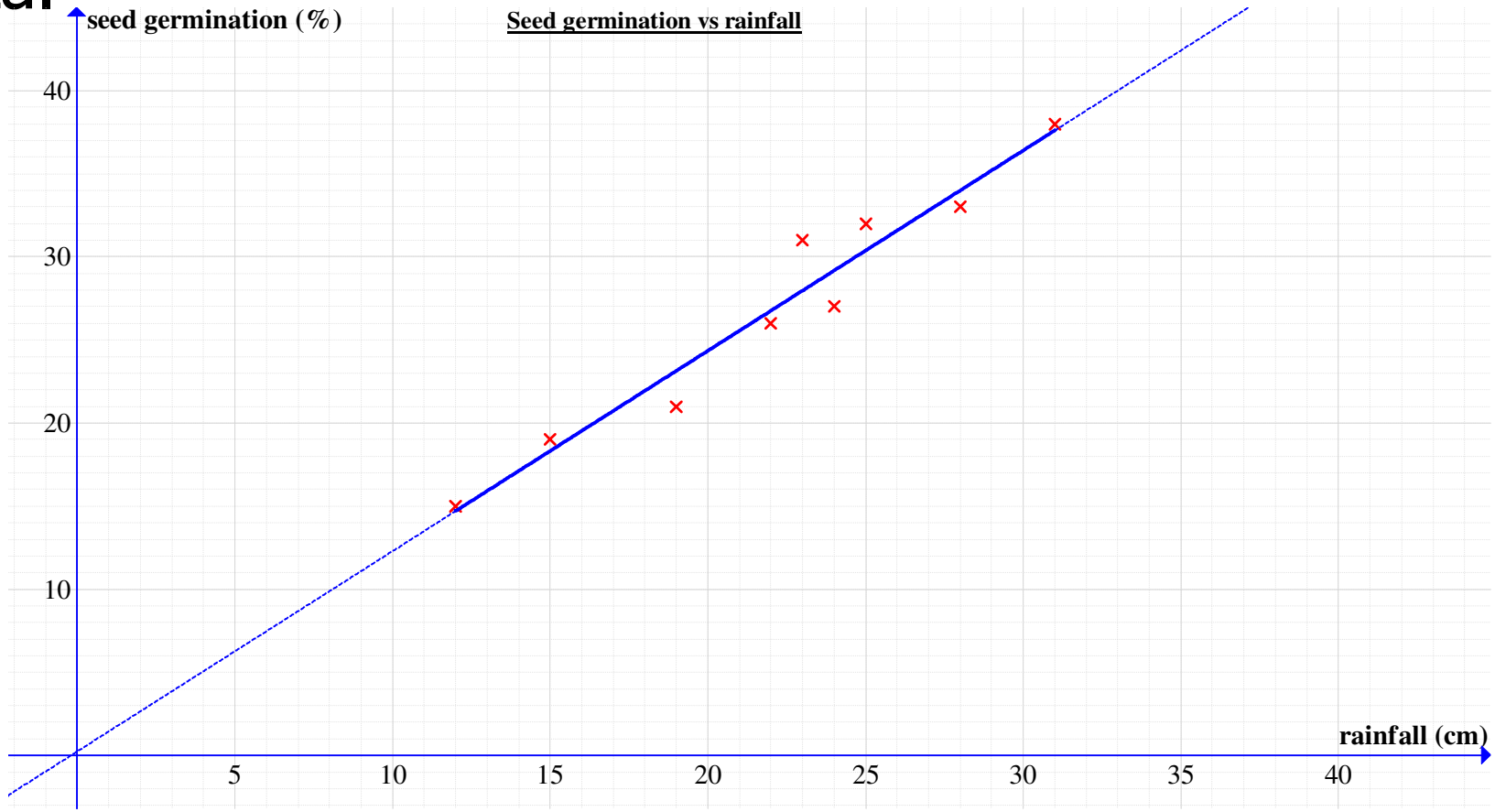
Examples

Firstly plot a scatter graph. Is it worth fitting a best fit line?



Examples

Yes, so draw the best fit line through the "middle" of the data.



Examples

- Now we can make some general predictions:
  - What is the predicted percentage germination for a rainfall of 30cm?
  - If the percentage germination is found to be 25% what is the estimated amount of rainfall?
  - What is the predicted percentage germination for a rainfall of 44cm? Can we trust this answer? If not, why not?
- This is the point of using lines of best fit: So that we can make general predictions (with one exception)

Examples

2) Thickness of eggshells.

Data was collected on the thickness of eggshells laid by birds of prey exposed to pollutants. A random sample was collected from 6 different nests, and tests for pollutant level p, and shell thickness w, was recorded as shown in the table below:

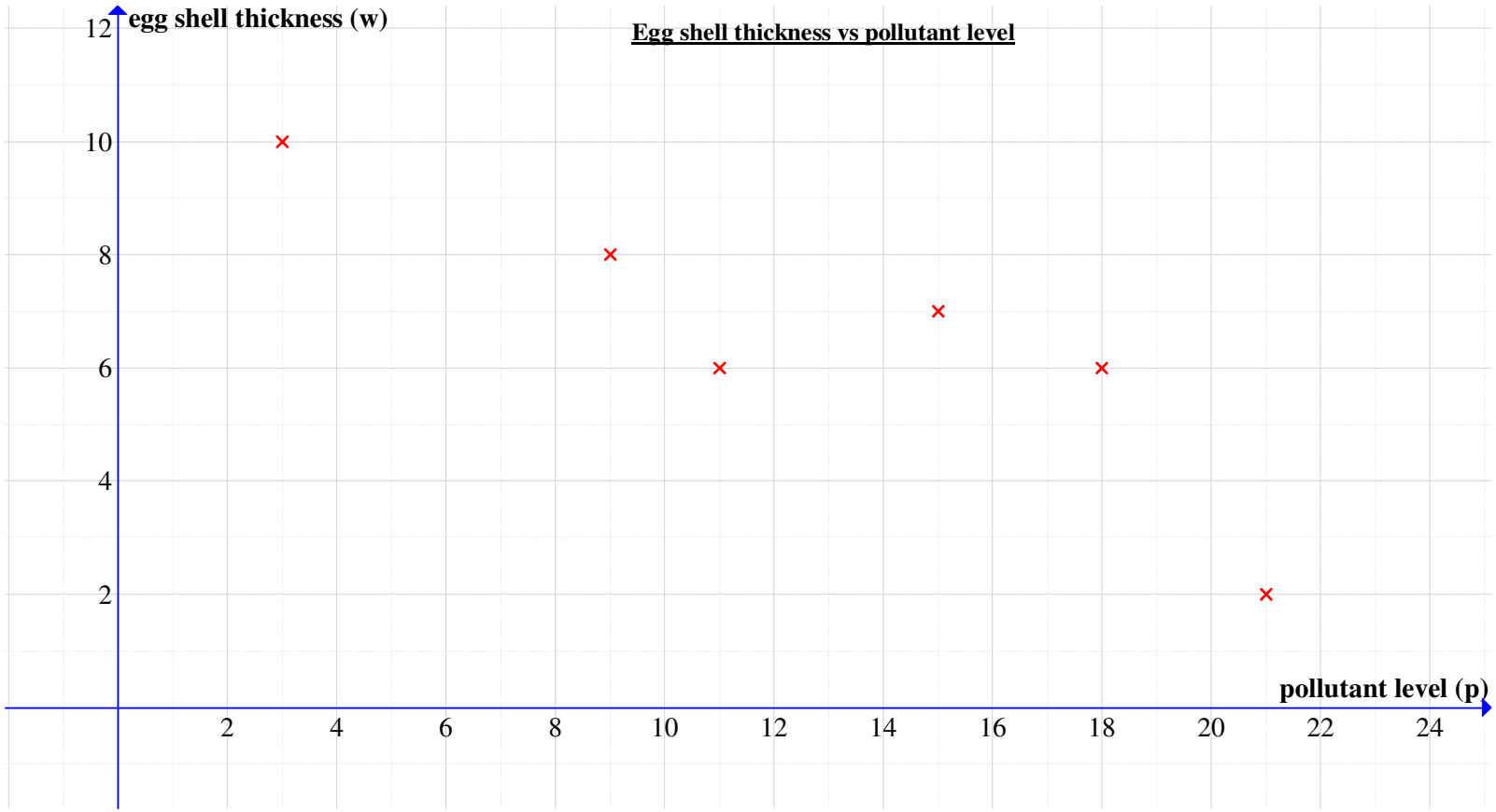
Examples

2) Thickness of eggshells.

Pollutant level (p)	Shell thickness (w)
3	10
9	8
11	6
15	7
18	6
21	2

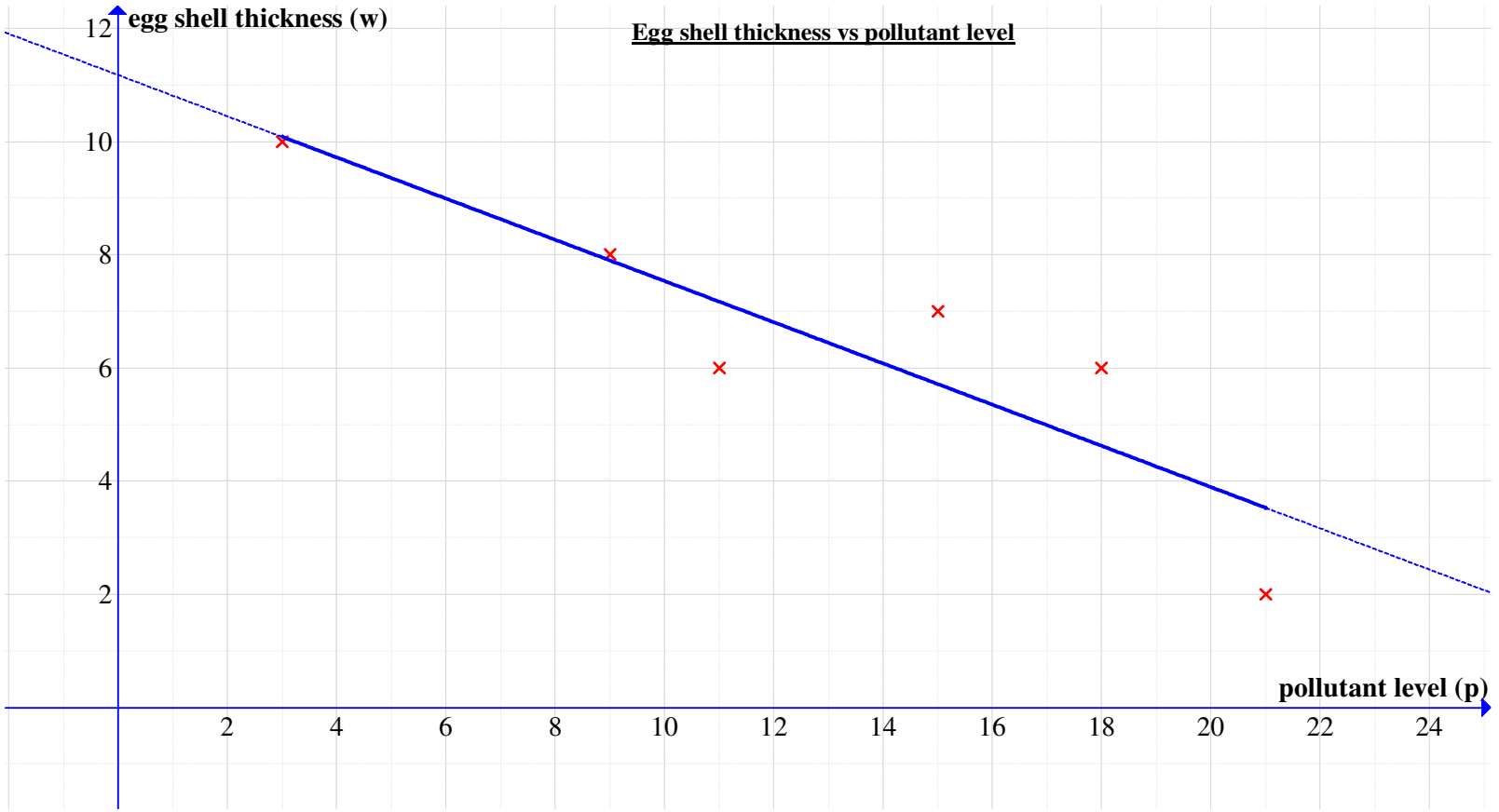
Examples

Firstly plot a scatter graph. Is it worth fitting a best fit line?



Examples

Yes, so draw the best fit line through the "middle" of the data.



Examples

Questions

- What can you deduce about the overall effect of pollution on eggshell thickness for these types of birds?
- If you measured the thickness of an egg shell to be 4, what approximate pollutant level could you deduce?
- What level of pollutants is required to decrease the thickness of an eggshell to 1? Can you trust this answer? If so, this would have to be biologically justified. If not, why not?

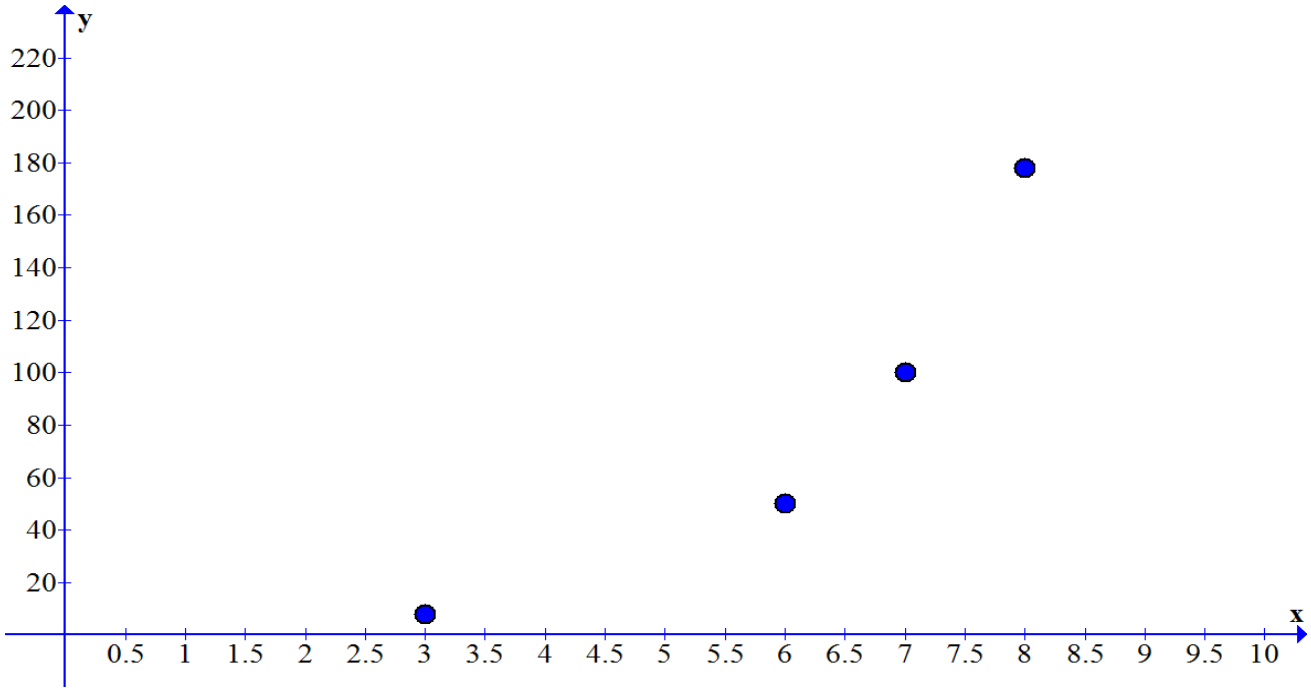
Is the trend really linear?

- Recall that r is the correlation coefficient.
- This tells us the degree to which the data has a linear tend.
- The larger the value of r (in terms of size) the more the trend is linear.
- But there are exceptions.

Is the trend really linear?

- A high value for r does not automatically mean that our data has a linear relationship.
- To see this consider the data and graph below

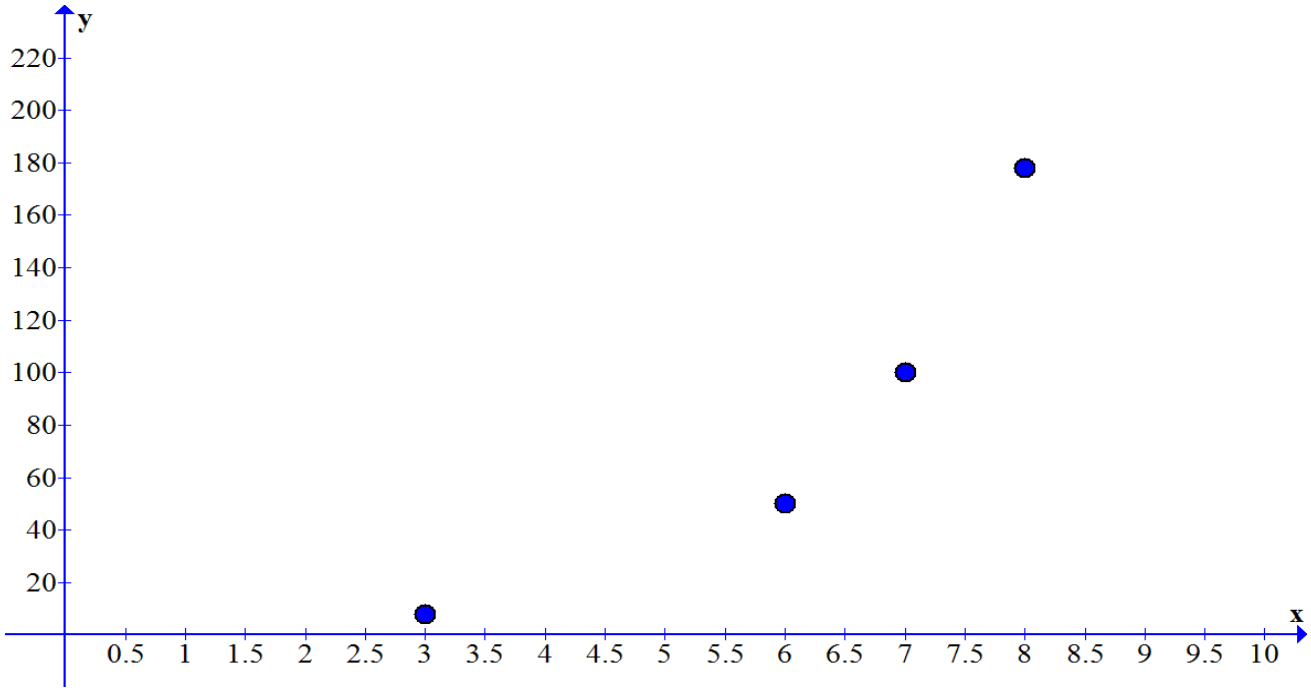
x	y
3	8
6	50
7	100
8	178



Is the trend really linear?

- There looks to be a linear relationship between x and y.
- The correlation coefficient is $r = 0.912$ (very high)

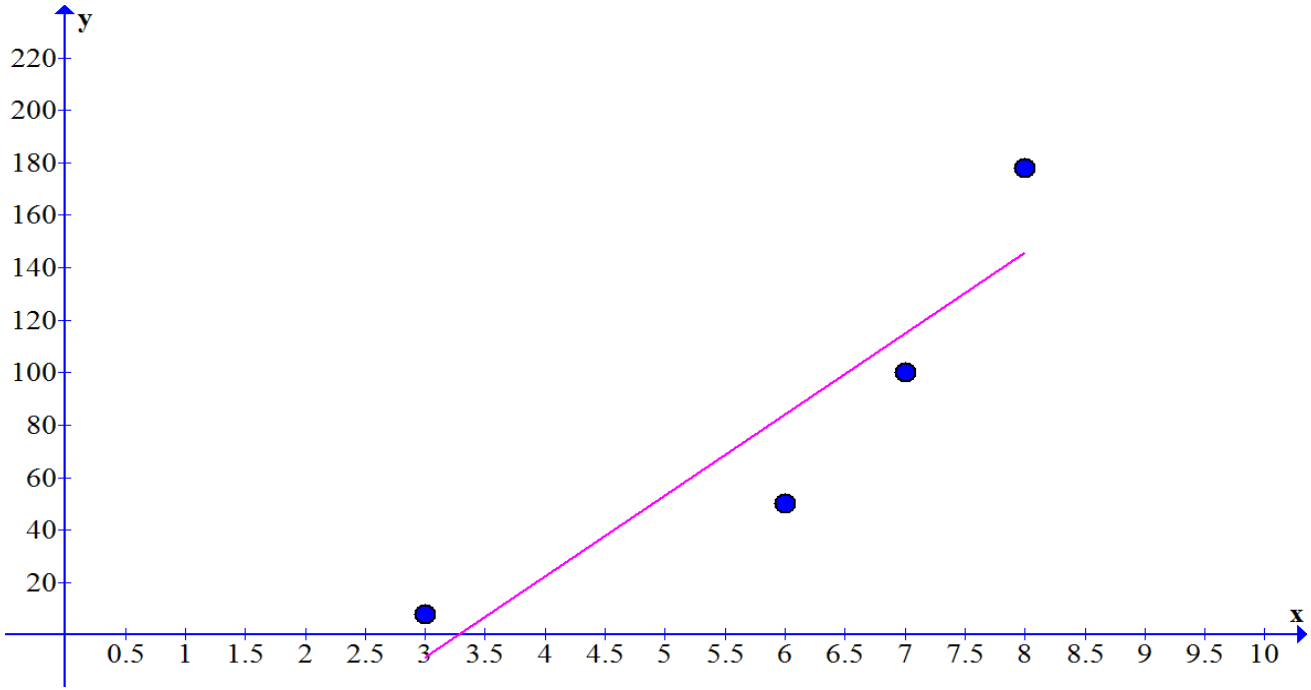
x	y
3	8
6	50
7	100
8	178



Is the trend really linear?

- There looks to be a linear relationship between x and y.
- The correlation coefficient is $r = 0.912$ (very high)

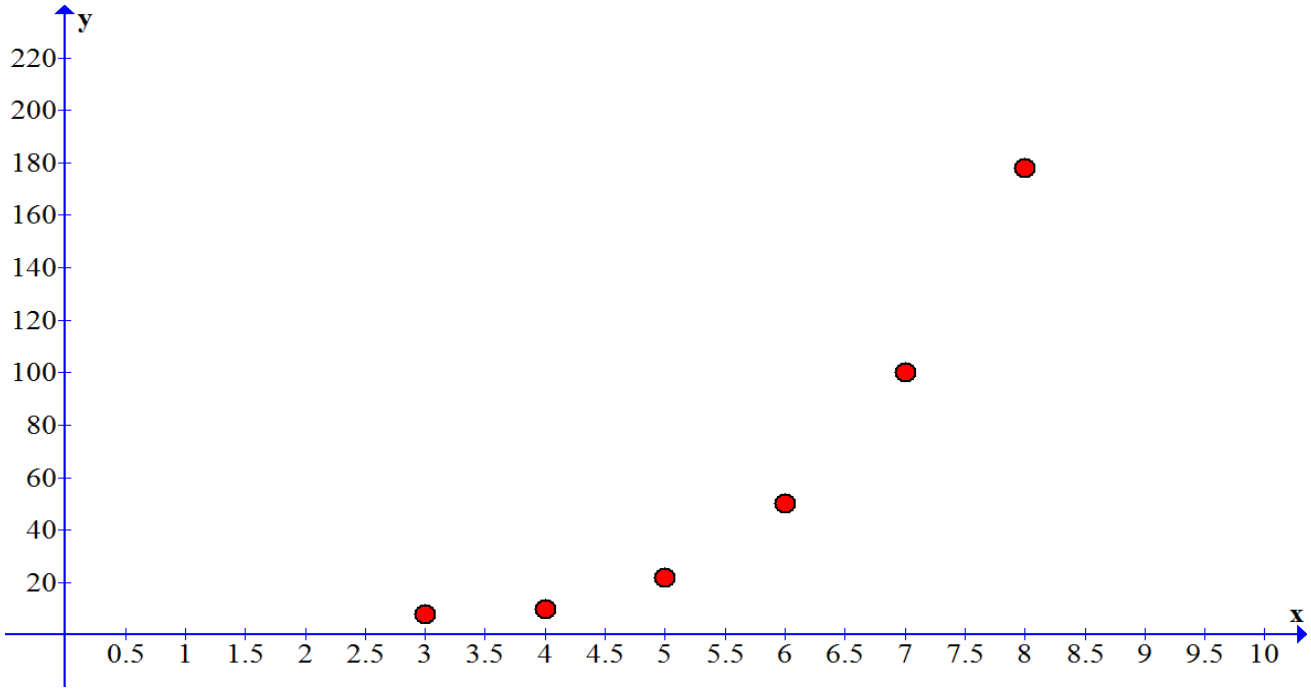
x	y
3	8
6	50
7	100
8	178



Is the trend really linear?

- But if we continue to collect data according to the pattern connecting x and y we obtain this:

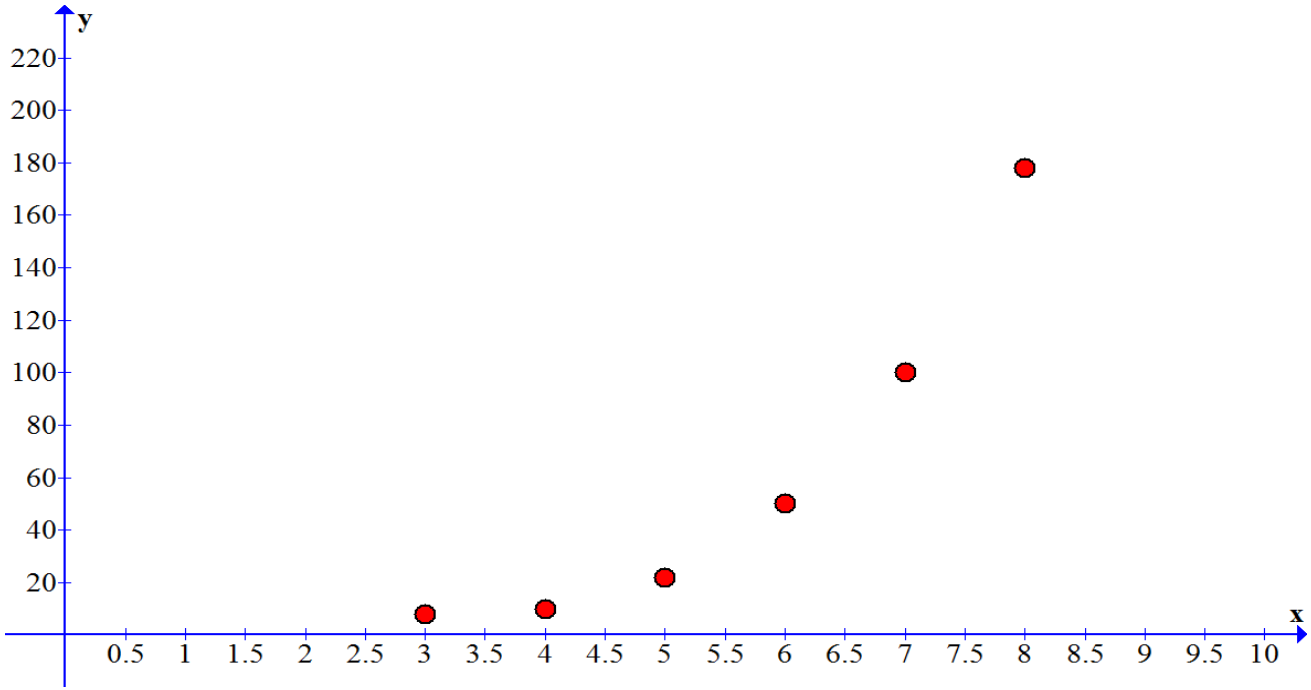
x	y
3	8
4	10
5	22
6	50
7	100
8	178



Is the trend really linear?

- This is not a linear relationship so our correlation coefficient is not valid.
- The actual correlation coefficient for this data is 1

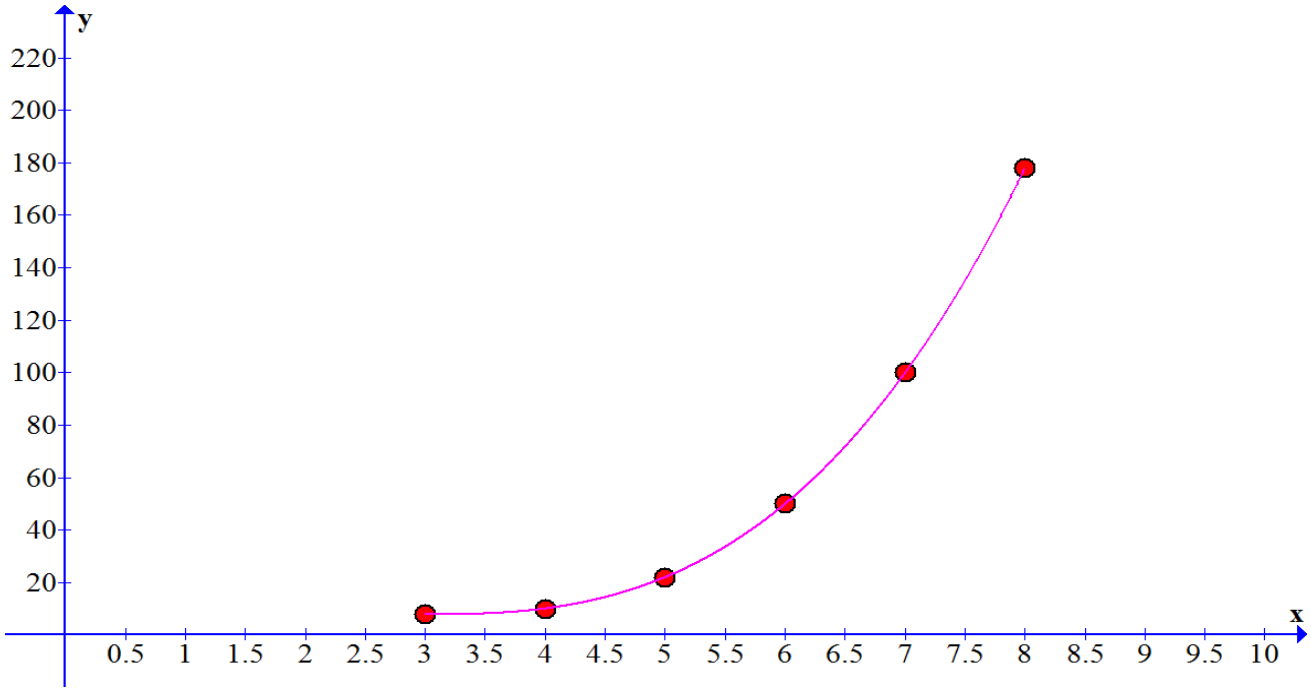
x	y
3	8
4	10
5	22
6	50
7	100
8	178



Is the trend really linear?

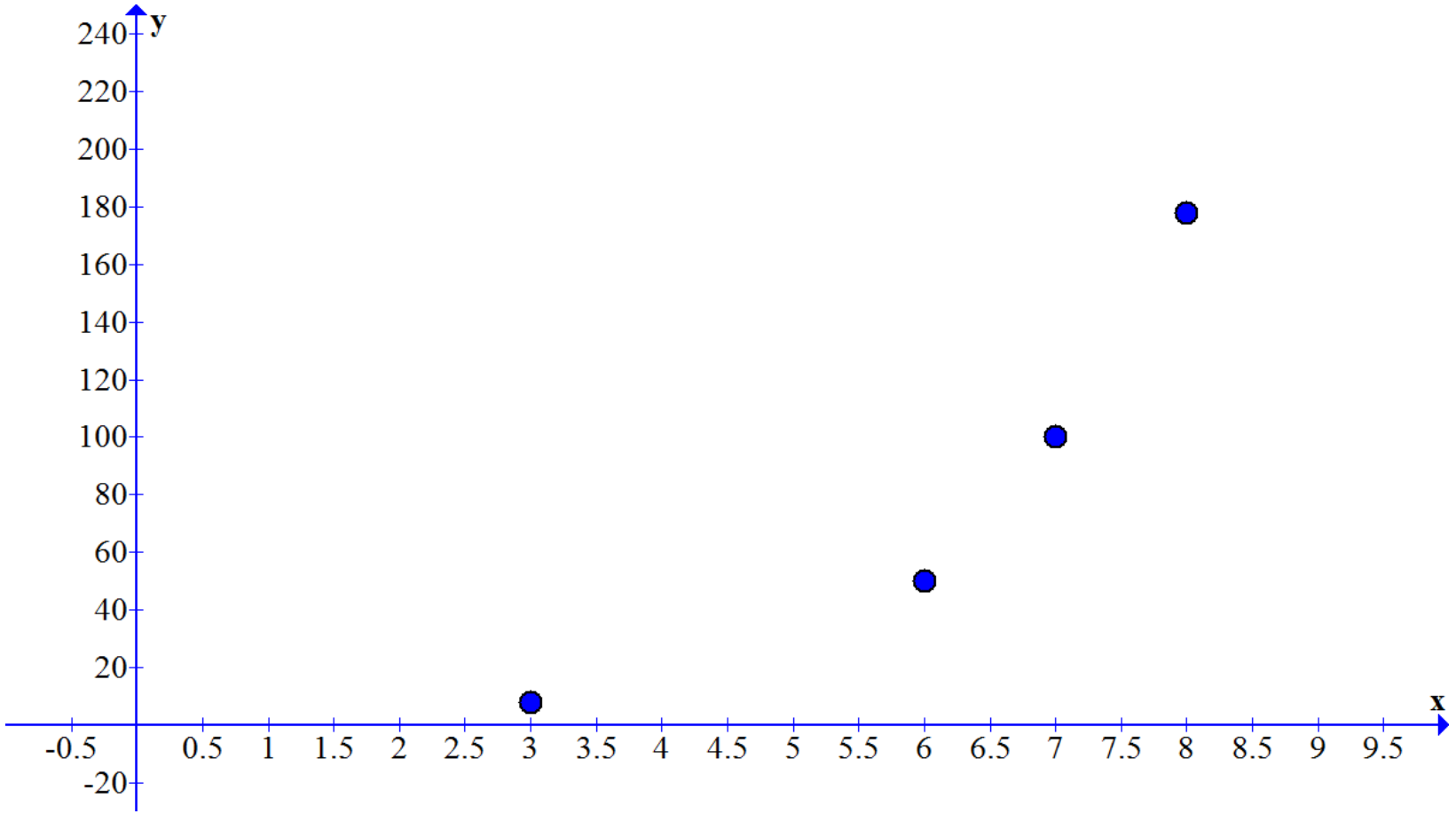
- In other words there is a perfect relationship between x and y but it is not linear.
- So what went wrong here?

x	*y*
3	8
4	10
5	22
6	50
7	100
8	178



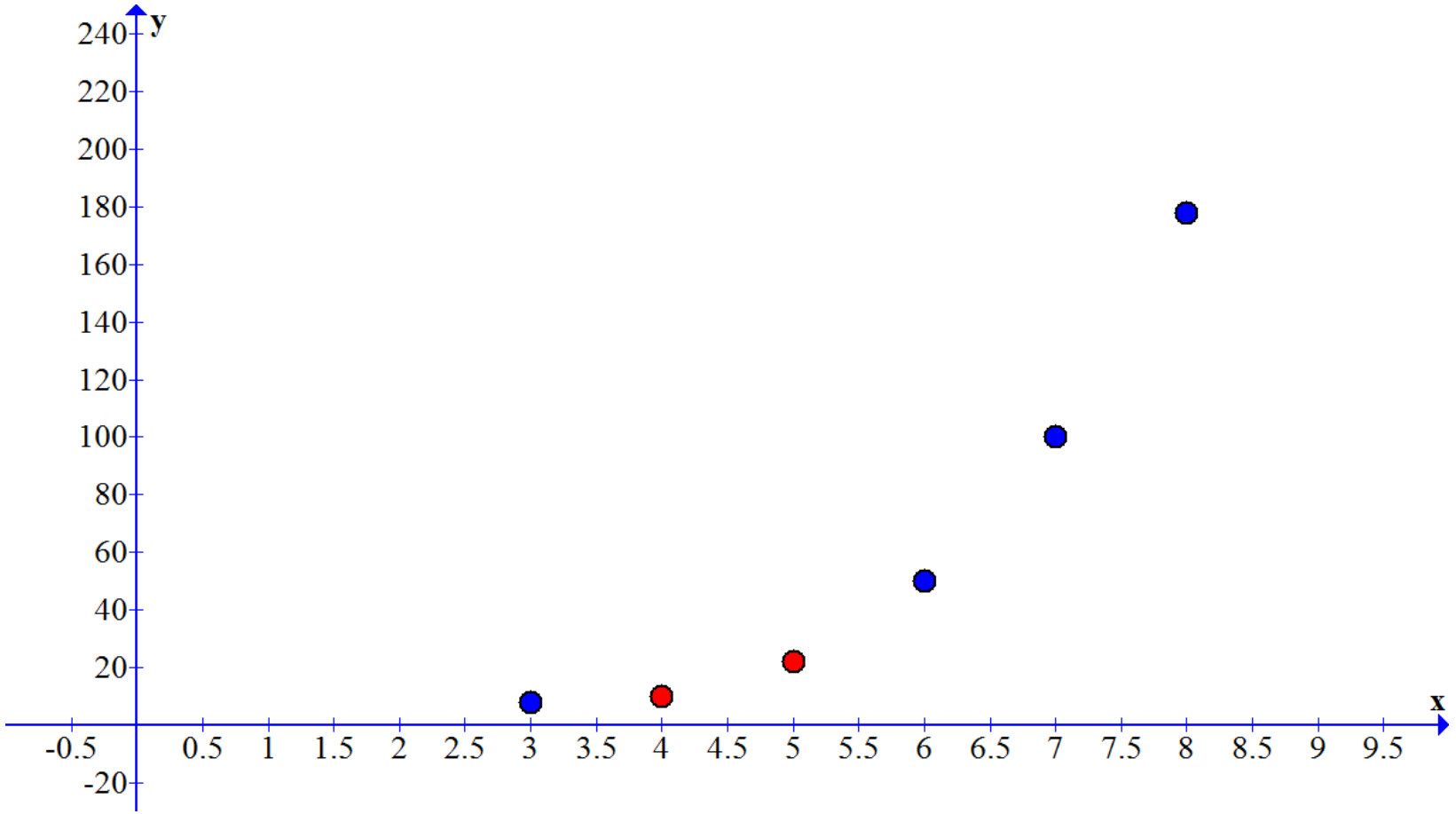
Is the trend really linear?

- Answer: Our original data is too widely spread out …



Is the trend really linear?

- Answer: … and we didn't collect enough data.



Is the trend really linear?

- So the moral of the story is:
  - The more data we have the more we can trust values of r as representing a linear correlation.
  - The less data we have the less we can trust values of r as representing a strong linear correlation.

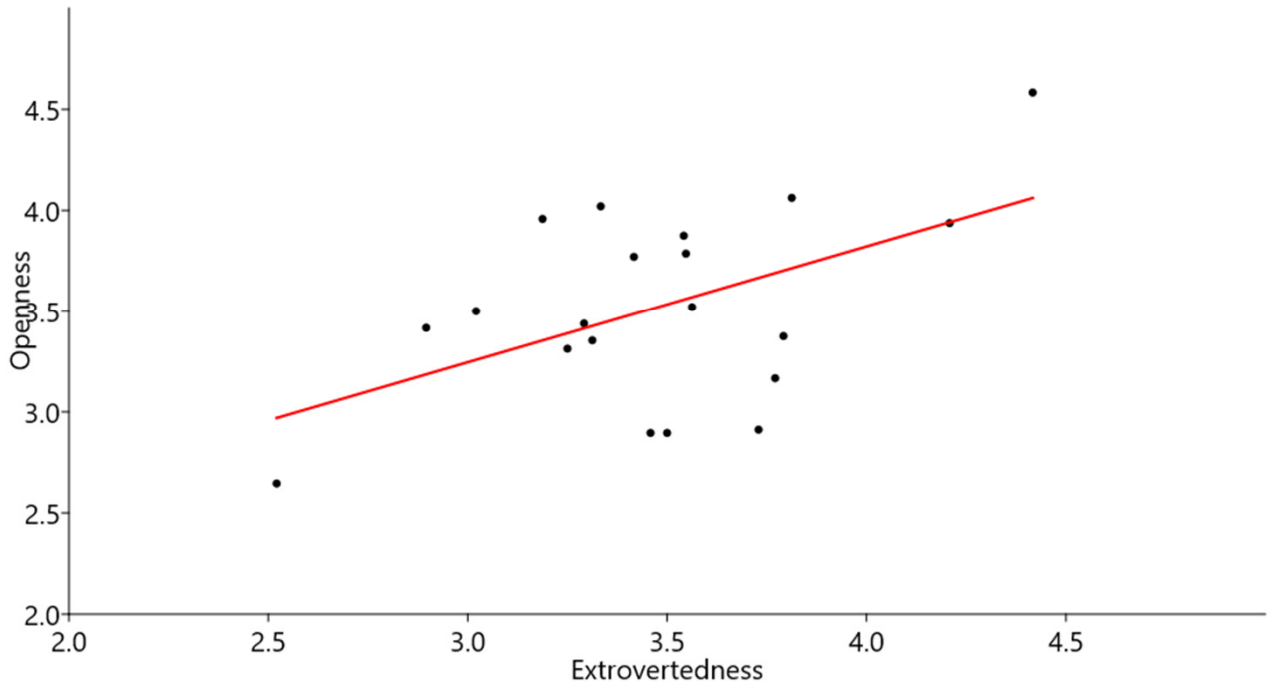
Is the trend really linear?

- So the moral of the story is:
  - The closer our data points are the more we can trust values of r as representing a linear relationship (if there is such a relationship).
  - The more spread out our data point the less we can trust values of r as representing a linear relationship.

Be careful about scaling

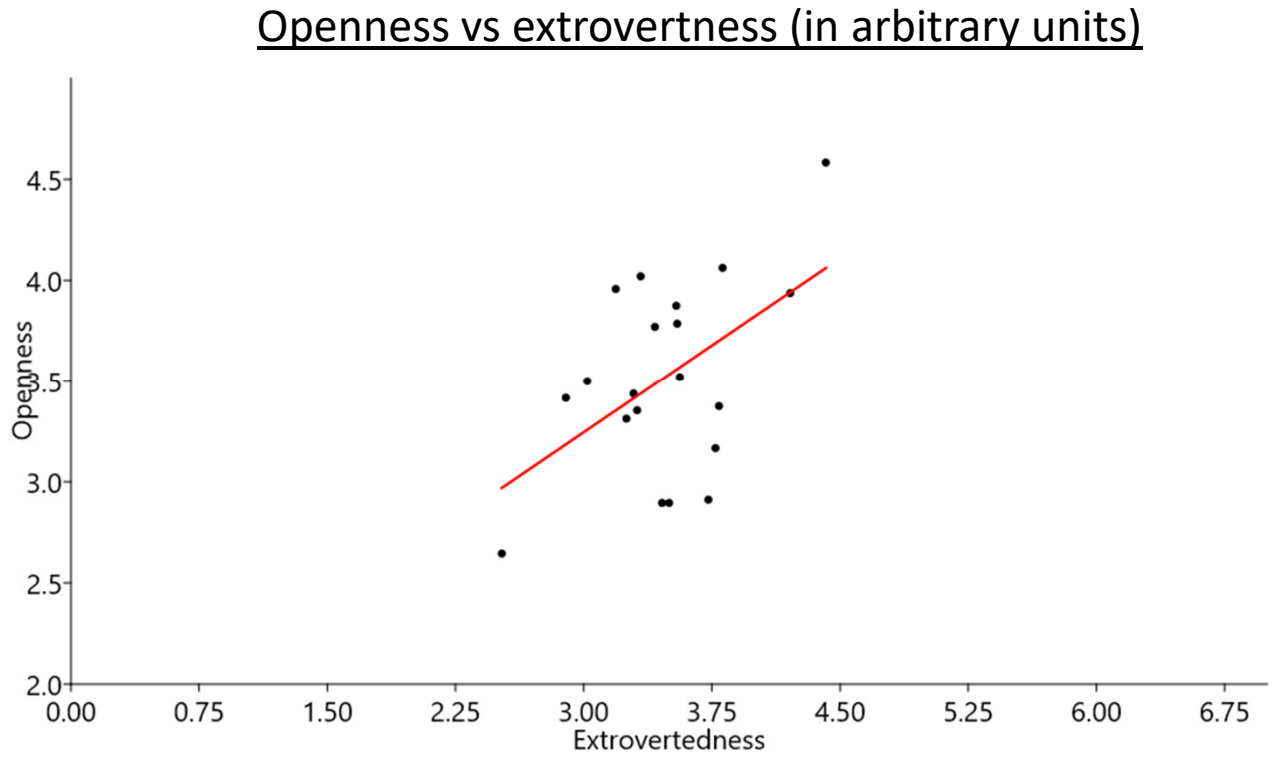
- *Example:* The following is a graph of people's openness compared to their extrovertedness:

Openness vs extrovertness (in arbitrary units)



Be careful about scaling

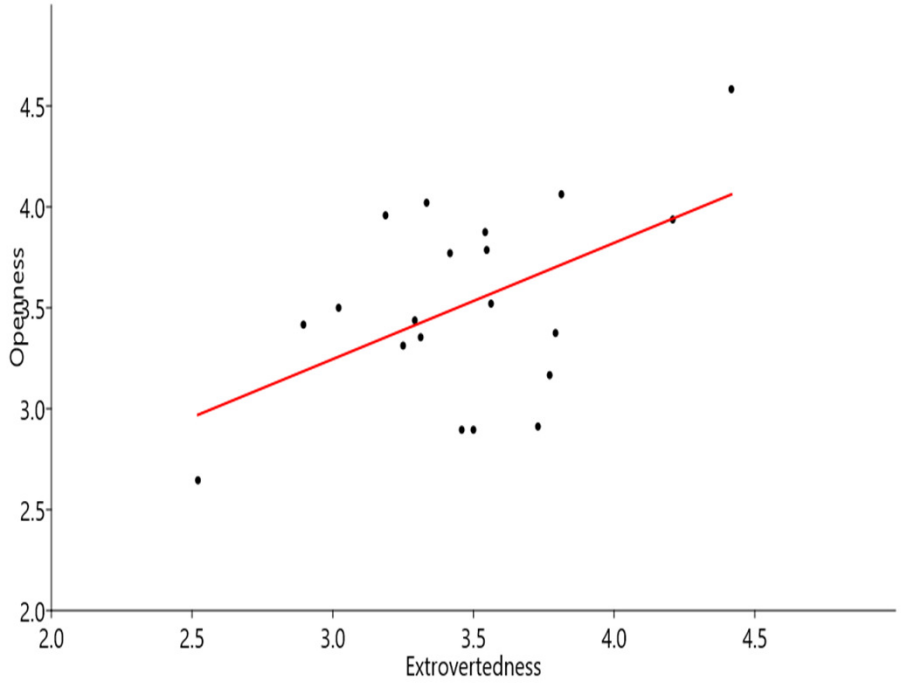
- The following is another graph of people's openness compared to their extrovertedness, using the same data as above.



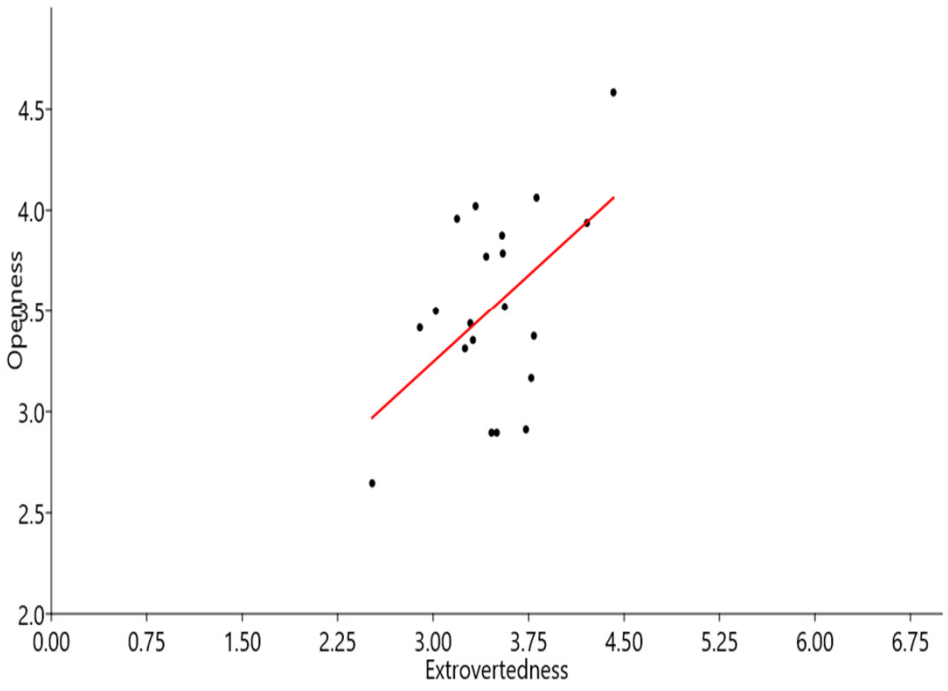
Be careful about scaling

- If the same data has been used for both graphs, why do they look different? Which one is correct?

Openness vs extrovertness (in arbitrary units)



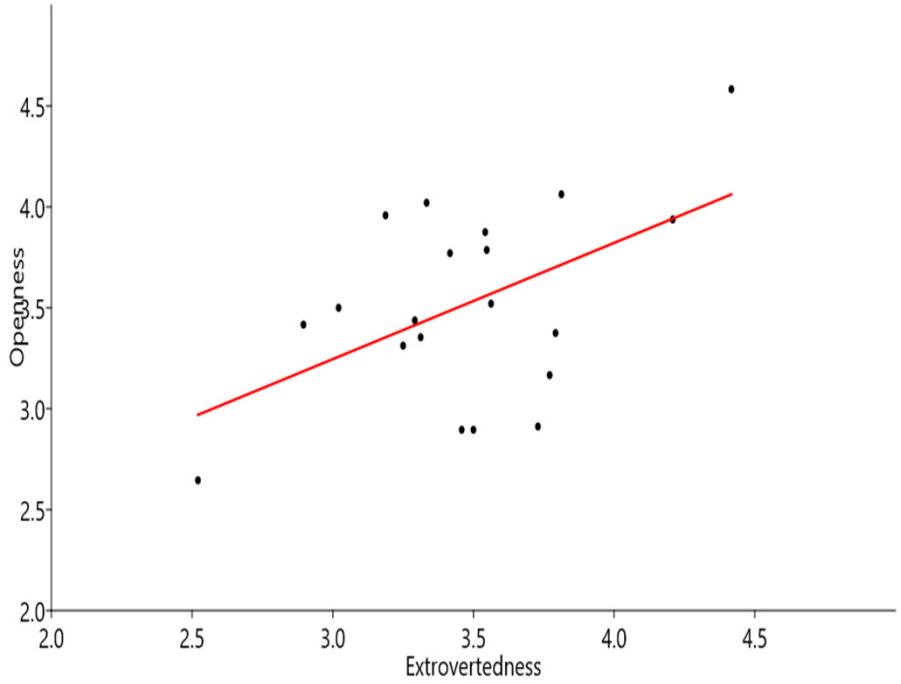
Openness vs extrovertness (in arbitrary units)



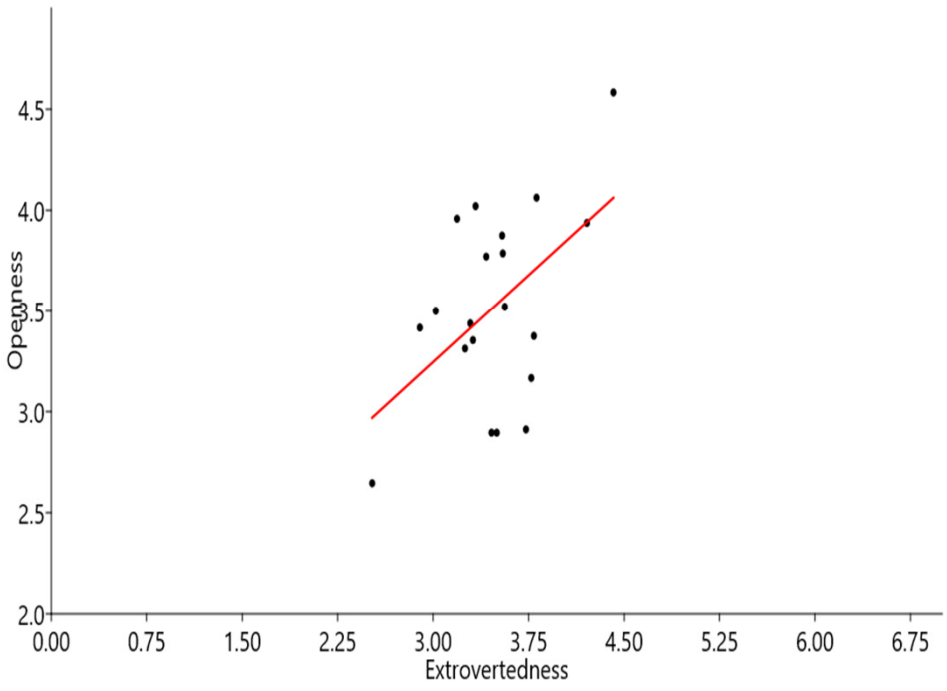
Be careful about scaling

- If you calculate the slope of the line mathematically, you will get the same slope from both graphs.

Openness vs extrovertness (in arbitrary units)



Openness vs extrovertness (in arbitrary units)



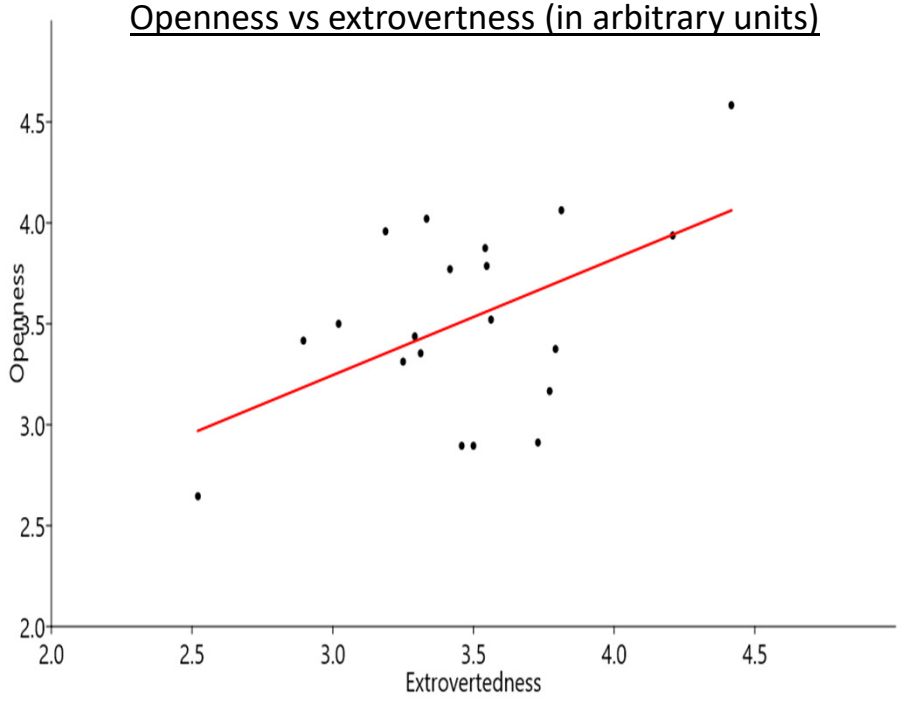
Be careful about scaling

- But if you measure the slope of the line using a ruler, you are much more prone to error.
- This is because your ruler measurement will be based on the shape/scale of the diagram, and the scale of the x-axis in the left graph is different to that on the right hand graph (it is more stretched out in the left graph compared to the right graph)

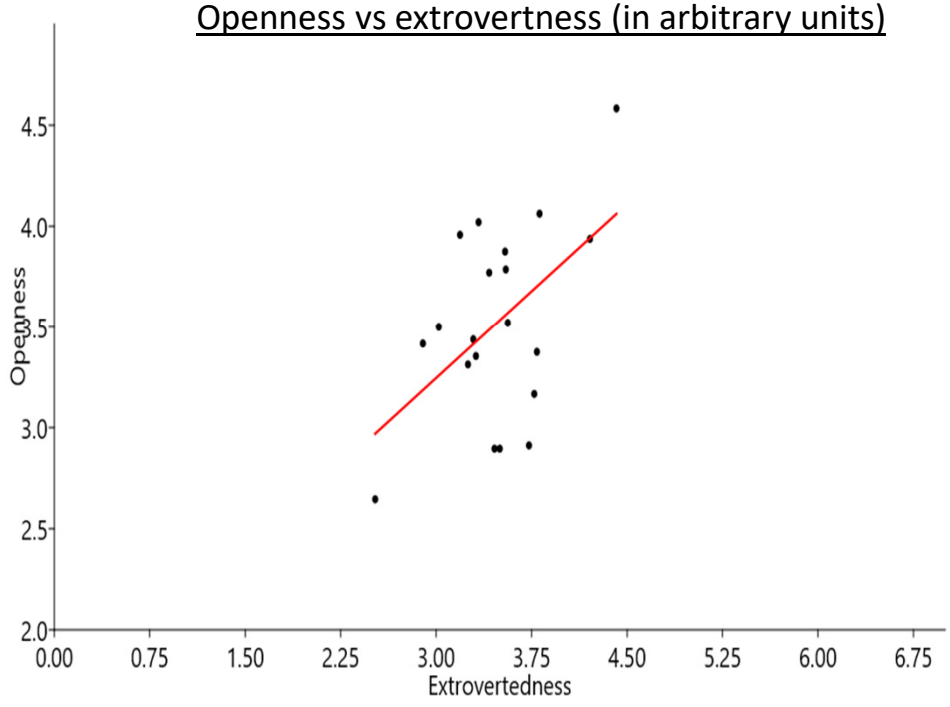
Be careful about scaling

- The rate of change in openness to extrovertness looks less steep in the left graph compared to the right graph.

Openness vs extrovertness (in arbitrary units)



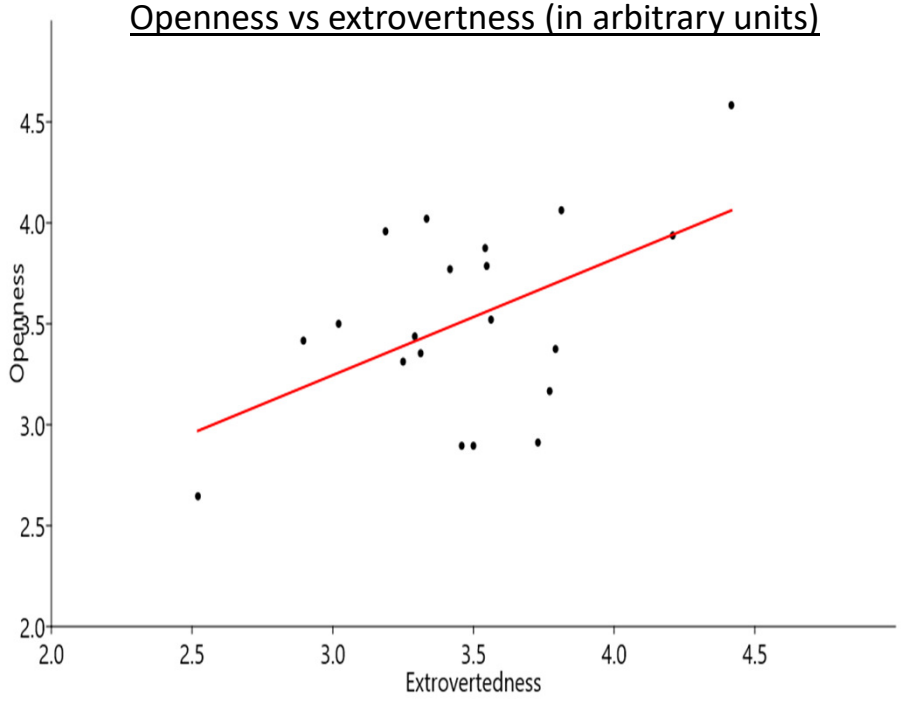
Openness vs extrovertness (in arbitrary units)



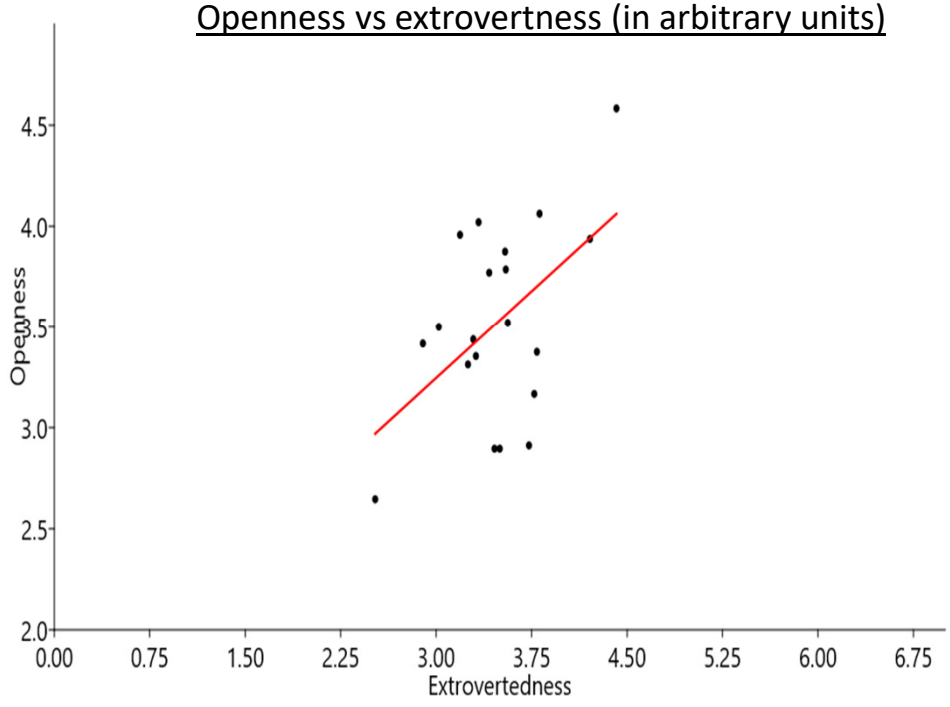
Be careful about scaling

- So, one of these graphs is a better visual representation, and the other is a bit of an illusory representation.

Openness vs extrovertness (in arbitrary units)



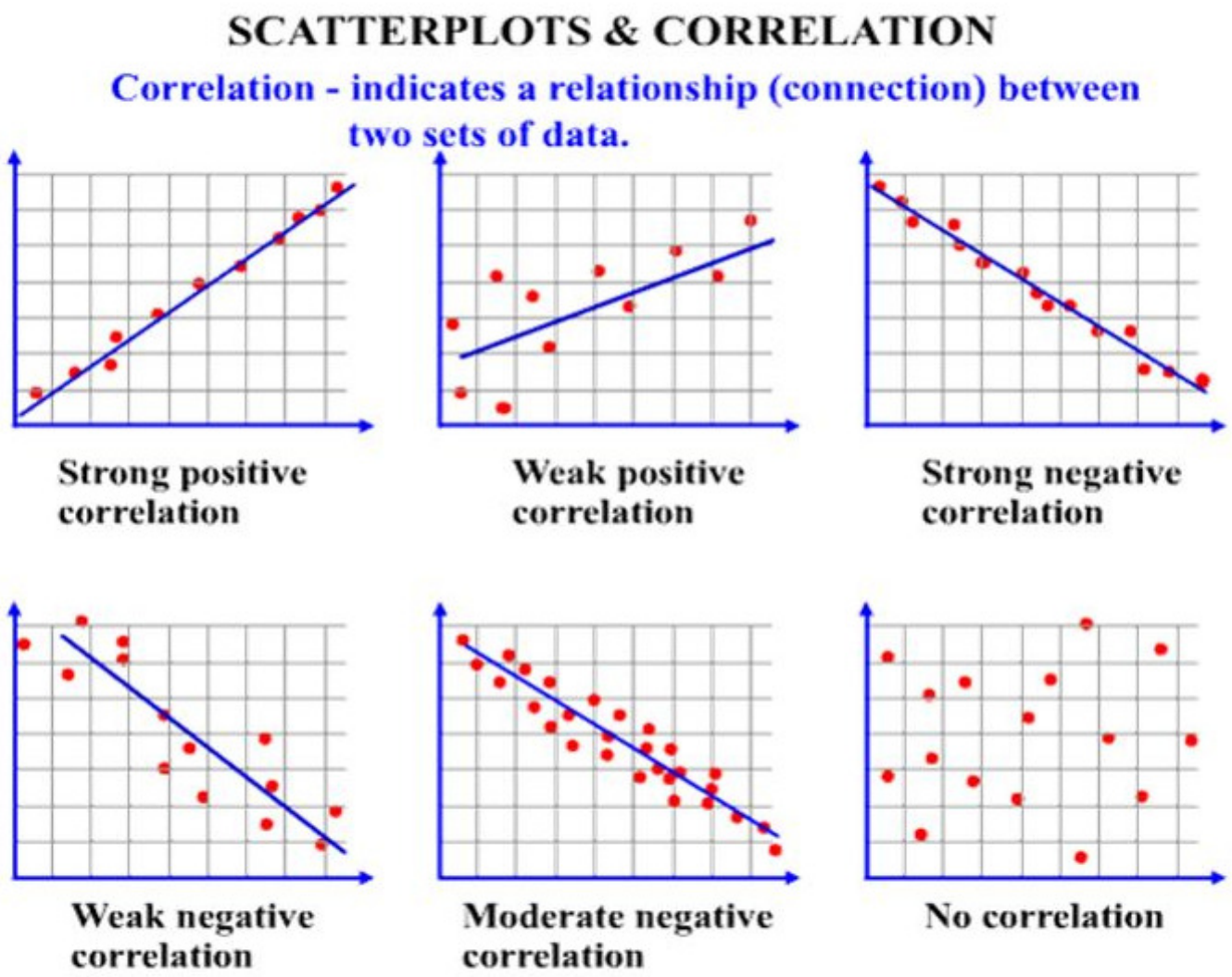
Openness vs extrovertness (in arbitrary units)



Be careful about scaling

- So, which graph should we use for visually presenting our data? See lesson.

Summary


SCATTERPLOTS & CORRELATION
Correlation - indicates a relationship (connection) between two sets of data.
Strong positive correlation
Weak positive correlation
Strong negative correlation
Weak negative correlation
Moderate negative correlation
No correlation

==========

The following slides are optional,

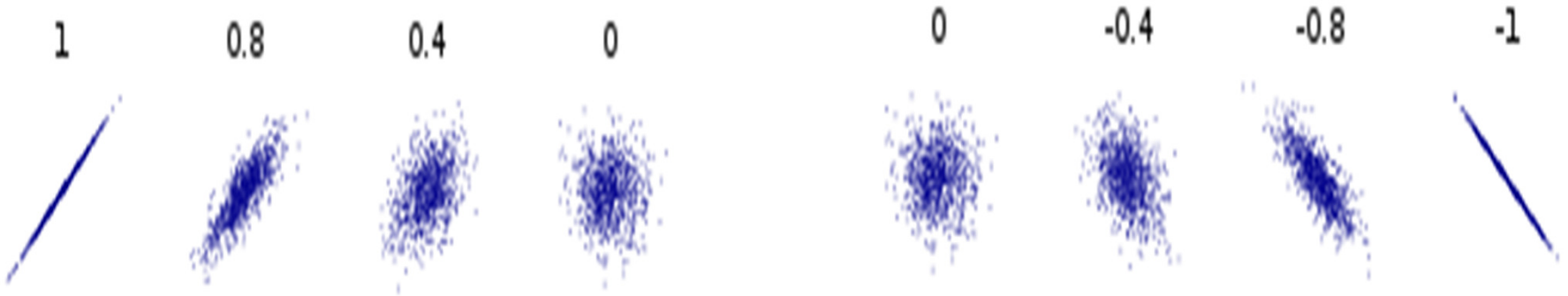
and do not form part of your biology course.

They provide more detail to the topic of best fit lines.

==========

The strength of the linear relationship: Correlation coefficient (optional)

- Returning to scatter plots:



we need to be able to calculate a value for the strength of the linear relationship between any two variables.

The strength of the linear relationship: Correlation coefficient (optional)

- This value is called the *correlation coefficient*, and its symbol is r.

- The higher the value of r (within ±1) the more linear the data is as a whole, and the stronger the linear relationship (provided we have enough data)

The strength of the linear relationship: Correlation coefficient (optional)

- The value for r is calculated as:

$$r = \frac{1}{n-1}\sum\left[\frac{\left(x_i - \bar{x}\right)}{s_x}\frac{\left(y_i - \bar{y}\right)}{s_y}\right] \qquad (1)$$

- Where s_x and s_y are the standard deviations of the x and y data.
- This formula helps us understand the concept of correlation coefficient, but is very slow to use when we need to do the arithmetic calculation.

The strength of the linear relationship: Correlation coefficient (optional)

- To do the arithmetic more quickly, use

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right]\left[n\sum y_i^2 - \left(\sum y_i\right)^2\right]}} \qquad (2)$$

The strength of the linear relationship: Correlation coefficient (optional)

- Equation (1) is easier to look at and understand conceptually, but takes much longer to run by hand.
- Equation (2) looks more messy but is much quicker to use by hand, especially if you have a lot of data.
- Speed of calculation is an important issue even when using software to do these calculations since these calculations may need to be done on 1000s or 10,000s of pieces of data.

Regression (optional)

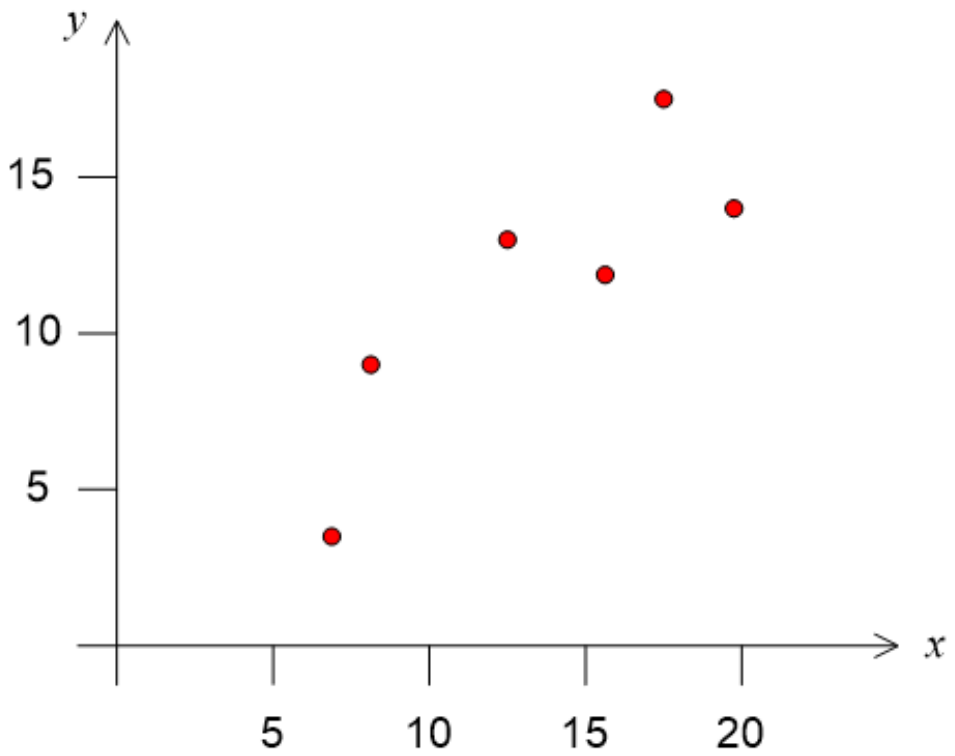
- We now need to find the equation of the straight line that best represents the trend of the data.
- This equation is called a line of best fit, since it is the best line that can be passed through all the data.
- What do we mean by “best” line?

Regression (optional)

- Suppose we have collected 6 pieces of data:

x	y
x_1	y_1
x_2	y_2
x_3	y_3
x_4	y_4
x_5	y_5
x_6	y_6

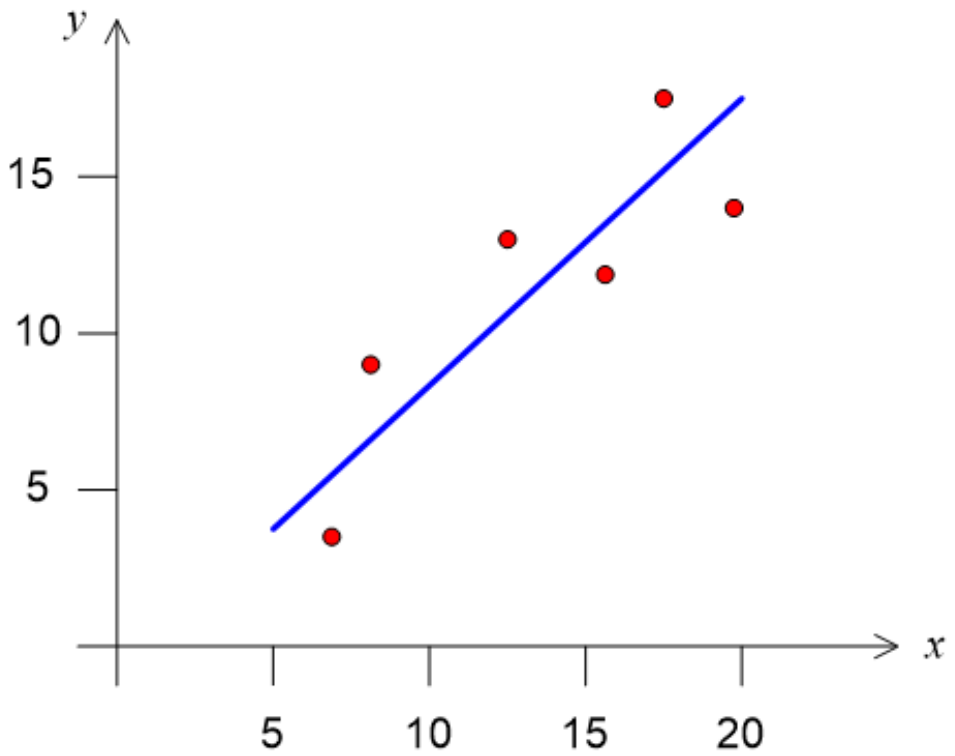
… and we obtain the graph below



Regression (optional)

- We want to find the best line which goes through this data.

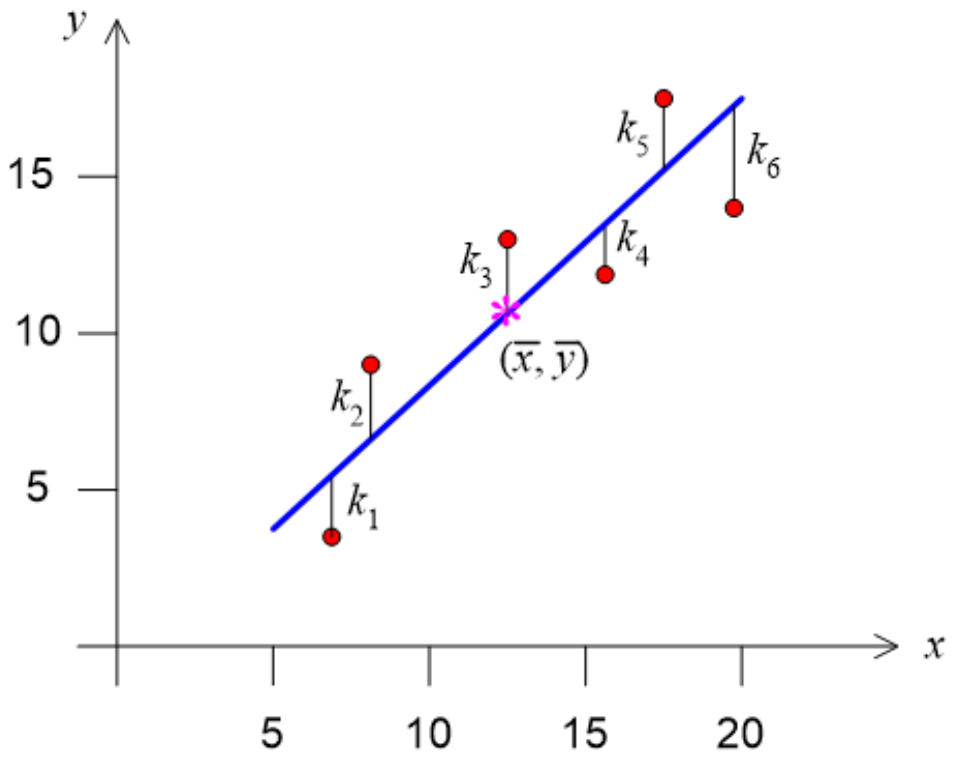
x	y
x_1	y_1
x_2	y_2
x_3	y_3
x_4	y_4
x_5	y_5
x_6	y_6



Regression (optional)

- Is there a way to measure the "closeness" of the data to the line?

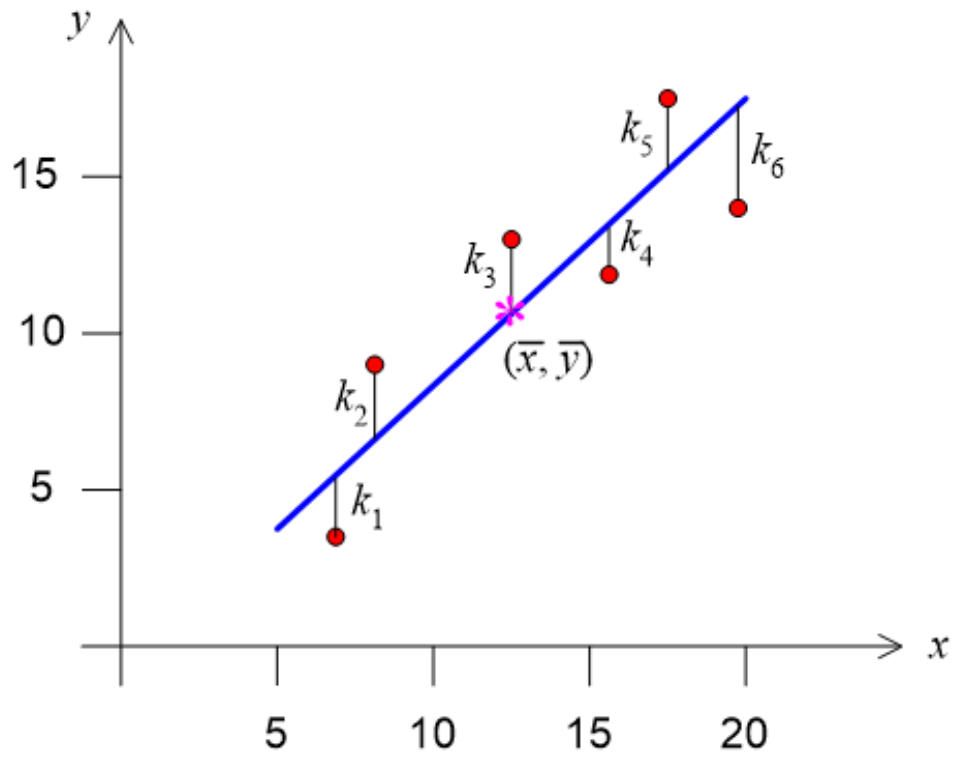
- Yes. We do this by considering the vertical distance between each value and the line.



Regression (optional)

- We want to minimise the total distance off all of these values to the straight line.

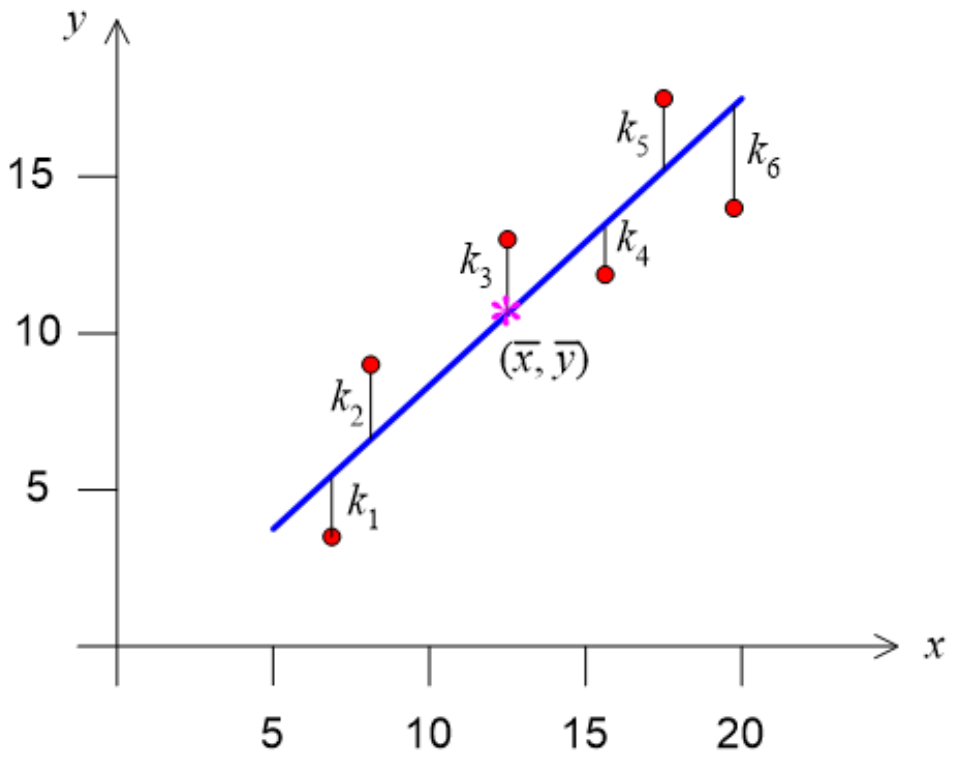
- Distances above the line will be positive and distances below the line will be negatives.



Regression (optional)

- We don't want the negative distances to have a cancelling effect on the positive distances, ...

- ... so we will square each distance, sum these squared values, and find the smallest sum possible.



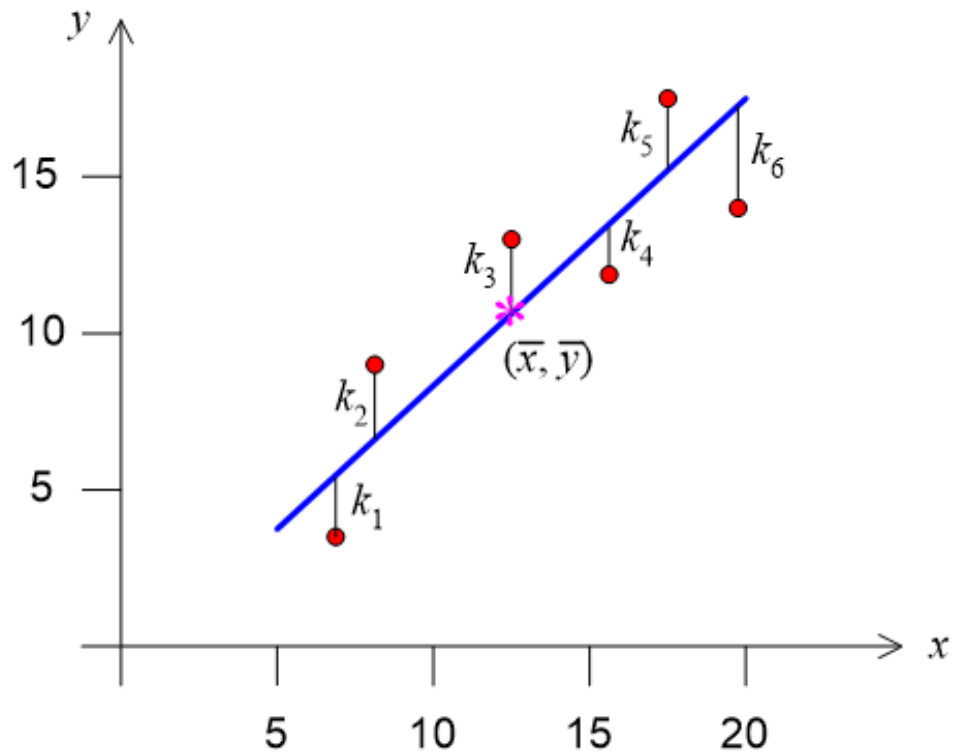
Regression (optional)

- This is why our best fit line is sometimes called *least squares line*.

- Remember now that the equation of a line is

$$y = mx + c$$

where c is the y-intercept and m is the slope.



Regression (optional)

- We need to find formulae for m and c. Because this involves some maths we haven't learnt I will simply state them:

$$m = \frac{n(\sum xy) - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \qquad (3)$$

$$c = \bar{y} - m\bar{x} \qquad (4)$$

where $\bar{x}$ and $\bar{y}$ are the mean values of the x data and y data respectively.

Regression (optional)

- As with calculating standard deviation, it will be useful to set up a table, as below, to help find *m* and *c*:

	x	y	x^2	xy
	.	.	.	.
	.	.	.	.
	.	.	.	.
	.	.	.	.
	.	.	.	.
$\Sigma =$				

Examples (optional slide)

The following slides repeat the previous examples of seed germination and thickness of egg shells but this time by going through the exact calculations for finding the equation of the best-fit line.

Examples (optional slide)

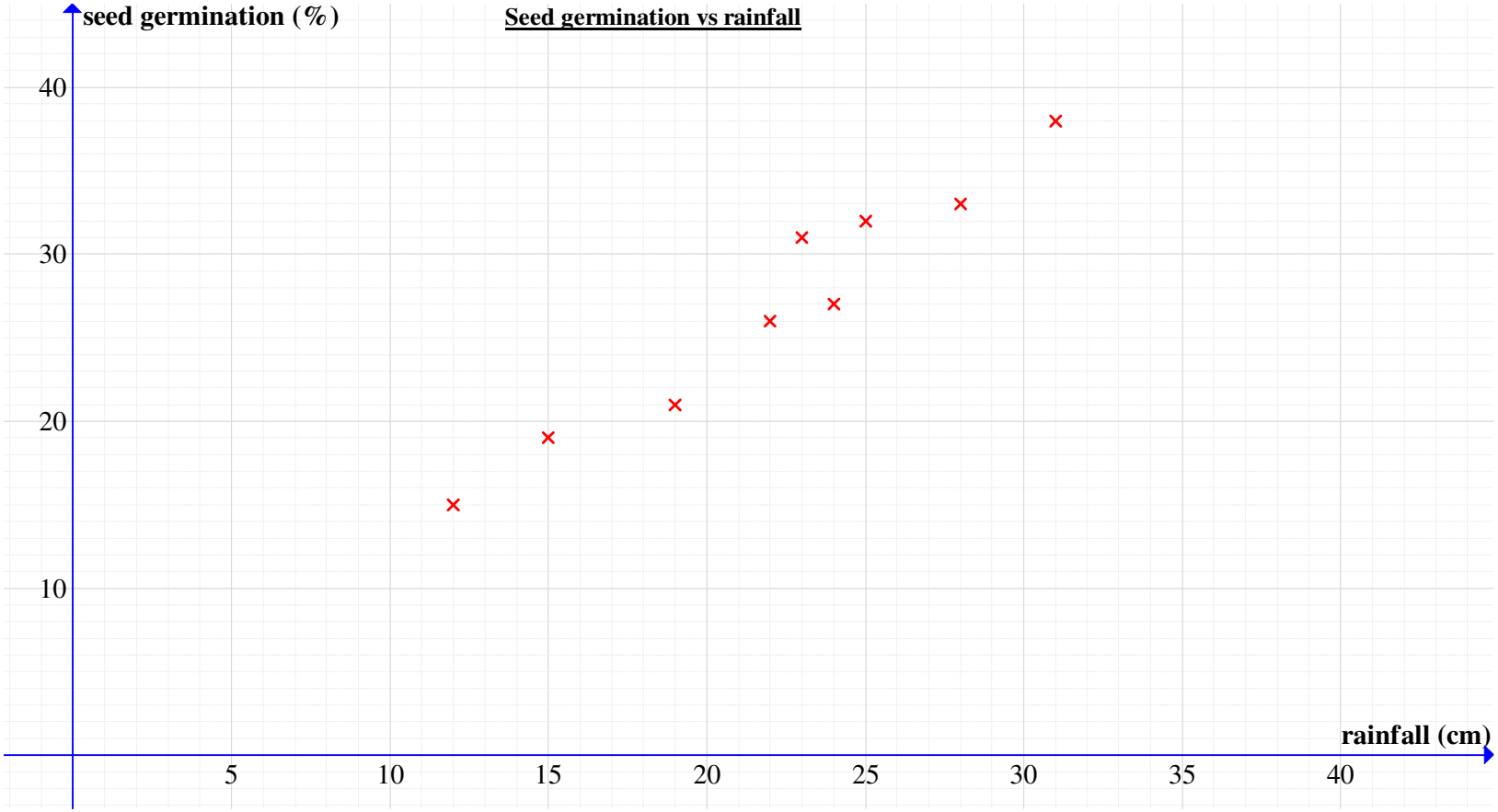
1) Seed germination

The table below shows the percentage of seed germination sown in areas with different amounts of monthly rainfall. Is there a linear relationship between rainfall and % germination? If so find the best fit line.

Rainfall (cm)	12	22	19	15	31	25	28	24	23
Germination (%)	15	26	21	19	38	32	33	27	31

Examples (optional slide)

Firstly plot a scatter graph. Is it worth find a best fit line?



Examples (optional slide)

- Is it worth find a best fit line? Yes. So find r and the line of best fit.

	x	y	x^2	xy
	12	15	144	180
	22	26	484	572
	19	21	361	399
	15	19	225	285
	31	38	961	1178
	25	32	625	800
	28	33	784	924
	24	27	576	648
	23	31	529	713
Σ =	199	242	4689	5699

Examples (optional slide)

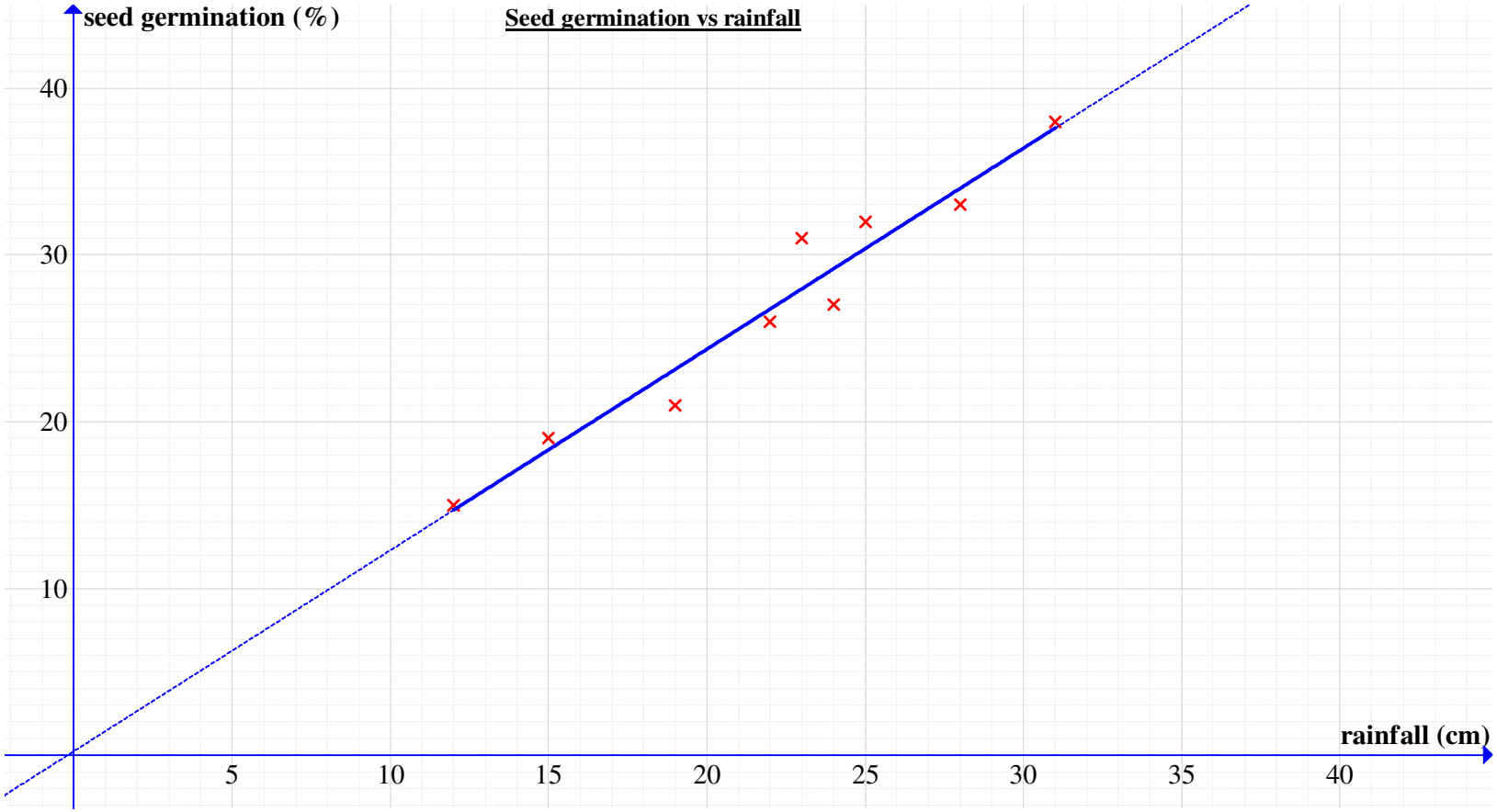
- So the equation of the line is found as

$$m = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{51291 - 48158}{42201 - 39601} = 1.205$$

$$c = \bar{y} - m\bar{x} = 26.89 - 1.205 \times 22.11 = 0.245$$

Examples (optional slide)

So the equation of the best fit line is $y = 1.205x + 0.245$



Examples (optional slide)

Questions

- What is the predicted percentage germination for a rainfall of 30cm?

- What is the predicted percentage germination for a rainfall of 44cm? Can we trust this answer? If not, why not?

Examples (optional slide)

2) Thickness of eggshells.

Data was collected on the thickness of eggshells laid by birds of prey exposed to pollutants. A random sample was collected from 6 different nests, and tests for pollutant level p, and shell thickness w, was recorded as shown in the table below:

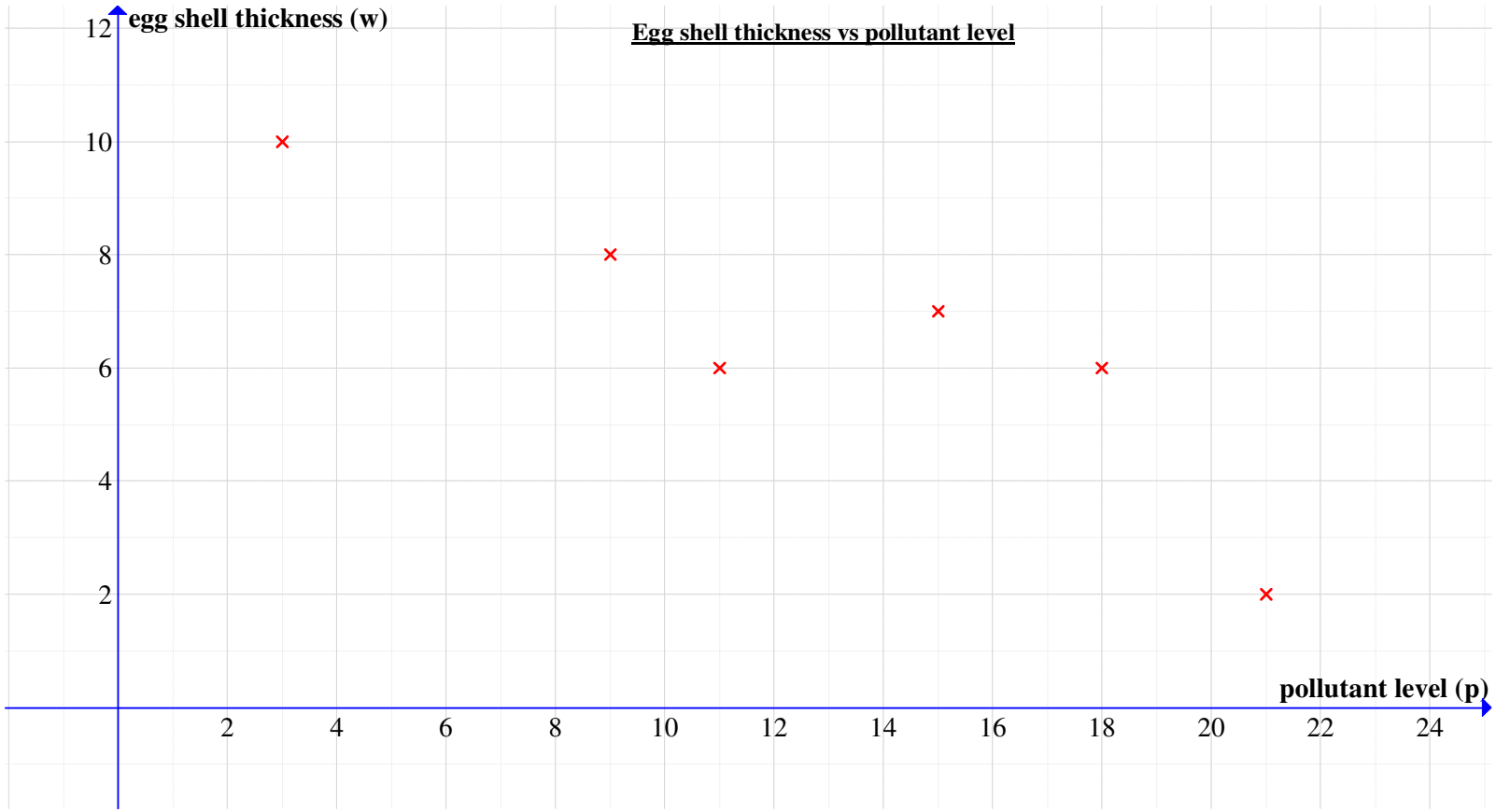
Examples (optional slide)

2) Thickness of eggshells.

Pollutant level (p)	Shell thickness (w)
3	10
9	8
11	6
15	7
18	6
21	2

Examples (optional slide)

Firstly plot a scatter graph. Is it worth find a best fit line?



Examples (optional slide)

- Is it worth find a best fit line? Yes. So find r and the line of best fit.

	x	y	x^2	xy
	3	10	9	30
	9	8	81	72
	11	6	121	66
	15	7	225	105
	18	6	324	108
	21	2	441	42
Σ =	77	39	1201	423

Examples (optional slide)

- $r = ?$
- $m = ?$
- $c = ?$

Examples (optional slide)

So the best fit line is $y = -0.364x + 11.173$