CALCULATING THE CORRELATION COEFFICIENT

1 Calculate the correlation coefficient r to 2 decimal places for each of the following bivariate data sets.

(a)

Length of trip (km)	0.5	10	40	60	80	100
Number of times parents are asked 'Are we there yet?'	2	5	18	20	25	40

(b)

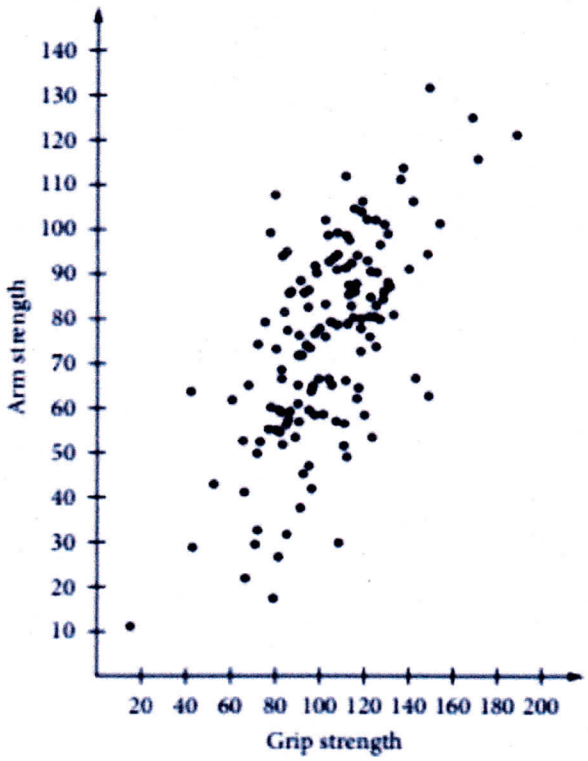
Computer time (h)	1	6	4	12	3	2
Number of 'Likes' added	24	40	50	120	50	30

a) Using EXCEL : =CORREL (Array 1, Array 2)

returns $r = 0.98$

b) $r = 0.94$

2 Consider the following scatterplot.



(a) Which pattern describes the scatterplot?

- **A** strong, positive linear
- **B** moderate, negative linear
- **C** no association
- **D** moderate, positive linear

(b) Which value would r be closest to?

- **A** 0.5
- **B** 0.2
- **C** 0
- **D** −0.2

CALCULATING THE CORRELATION COEFFICIENT

3 For each data set:

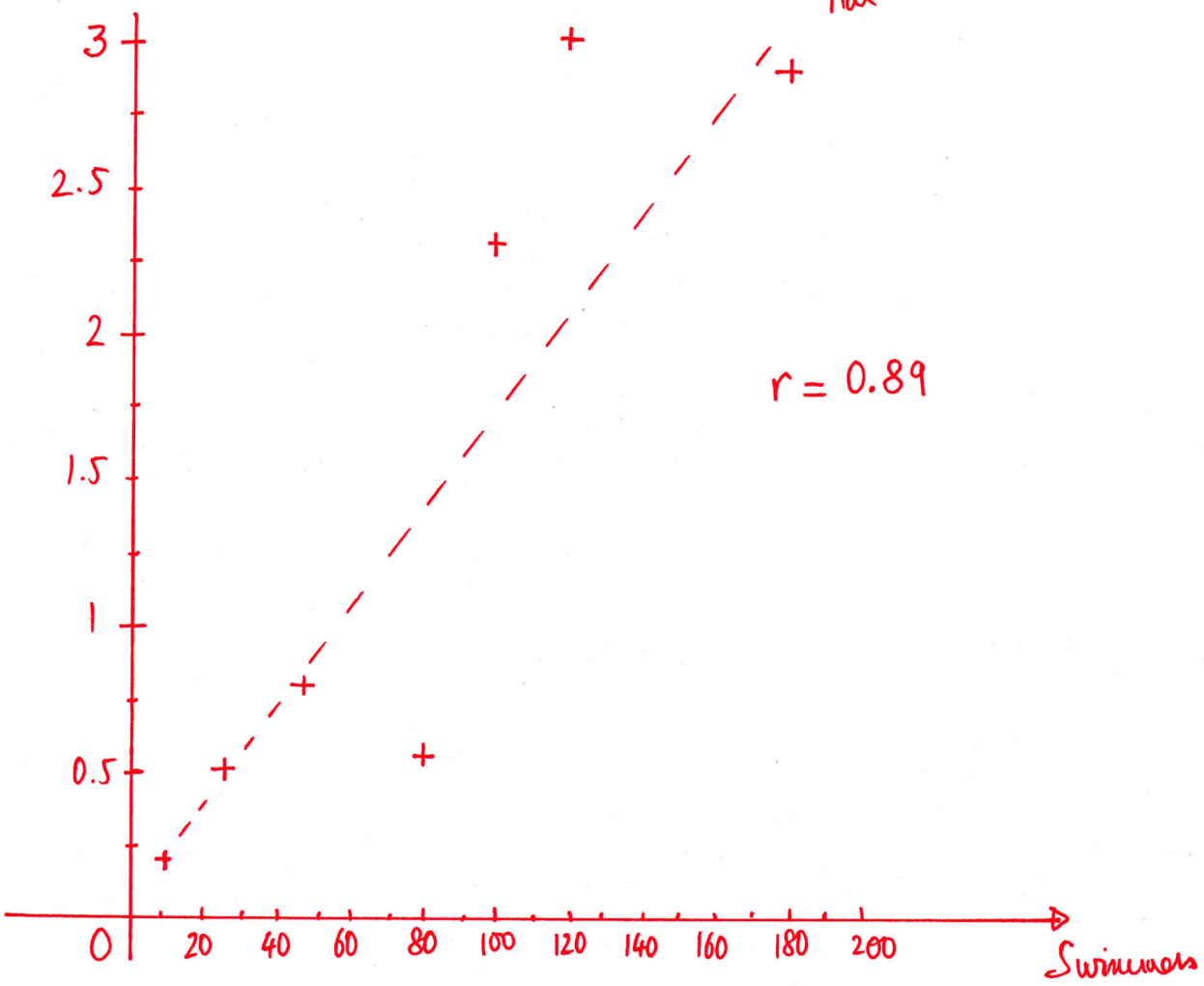
(i) use your technology to draw a scatterplot and calculate r

(ii) comment on the association (if any).

Swimmers in the pool	10	25	45	80	100	120	180
Bacteria level (ppm)	0.2	0.5	0.8	0.6	2.3	3.0	2.8

Max

Max



bi) Strong, positive linear

CALCULATING THE CORRELATION COEFFICIENT

4 The following experimental data was collected.

Rate of reaction (mol/min)	4.2	6.8	10.3	15.6	182	19.5
Temperature (K)	300	350	400	450	500	550

(a) Without removing any outliers in the data, calculate the value of the correlation coefficient r to 2 decimal places.

(b) Identify the outlier in the data.

(c) An outlier exerts a large influence on the data. Which statement is true?

A Outliers are generally excluded from the data set, so calculations without it are an accurate representation of the study.

B Outliers must be included because they form a significant part of the data set.

(d) After excluding the outlier, use technology to draw a scatterplot of the data.

(e) How is it evident in this scenario that excluding the outlier will mean that the points line up with a stronger linear relationship?

(f) What error may have occurred when recording the data?

a) without removing outliers, $r = 0.46$

b) (182, 500) is an outlier

c) [A]

d) after removing the outlier $r = 0.99$

e) excluding the outlier means the points line up with a stronger linear relationship. This is evident since r changes from 0.46 to 0.99.

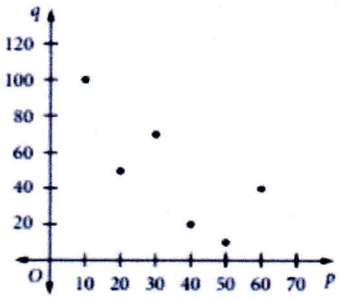
f) The number 182 mol/min was most likely meant to be 18.2 mol/min

CALCULATING THE CORRELATION COEFFICIENT

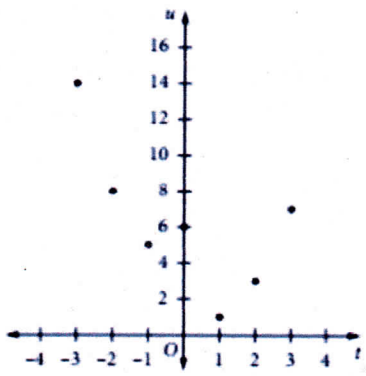
5 Three bivariate data sets and their scatterplots are shown below.

Set A

Set B



Set C



Answer the following questions about the bivariate associations. Note the same response may be true for more than one question.

(a) For which data set is the sample correlation coefficient r closest to -1? → Set B
(b) For which data set is the sample correlation coefficient r closest to 1? → Set A
(c) Which data set is non-linear? → Set C
(d) Which data set indicates the strongest correlation between its two variables? → Set B

6 Look at the pairs of variables in each row of the table. Where causation is observed, complete the following statement, replacing 'variable 1' and 'variable 2' with the actual names. Where causation cannot be verified, discuss the reason.

'__% of the change in variable 2 is due to the change in variable 1. __% of the change is due to other factors.'

	Variable 1	Variable 2	r	r^2
(a)	Hours of rehearsal	Quality of performance	0.8	0.64
(b)	Number of fast food outlets in a town	Number of hospitals	0.95	0.90
(c)	Number of registered cars	Number of registered motorcycles	0.6	0.36
(d)	Exam result in Physics	Exam result in Maths	0.9	0.81
(e)	Episode number of a reality TV series	Number of contestants remaining	0.98	0.96

a) 64% of the change in the quality of the performance is due to the change in hours of rehearsal. 36% is due to other factors

b) The number of fast food outlets and the number of hospitals in a town depend upon the population

c) 36% of the change in the number of registered motorcycles is due to the change in the number of registered cars. 64% of the change is due to other factors

d) The results in Maths and Physics exams depend upon other factors such as interest, hours of study and intelligence

e) 96% of the change in the number of contestants remaining in a reality TV series is due to the change in the episode number. 4% is due to other factors.