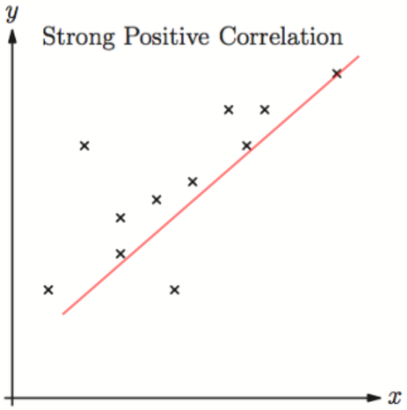
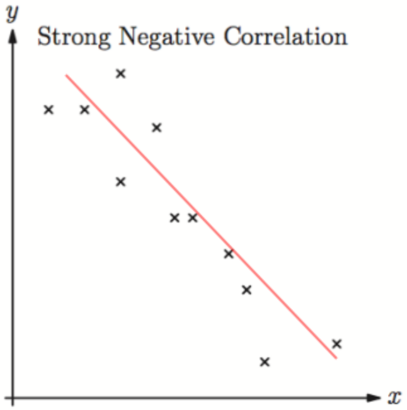
Correlation

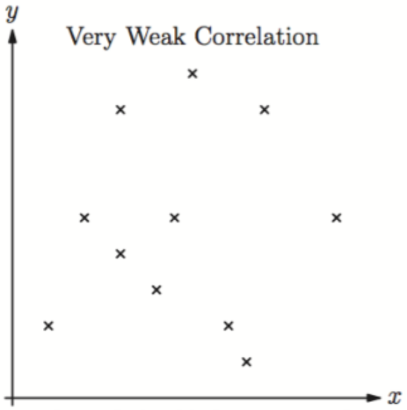
In this course we shall focus on *linear* correlation – the extent to which two variables X and Y are related by a relationship of the form $Y = mX + c$. If the gradient m of the linear relationship is positive, we say that the correlation is positive; if the gradient is negative, we describe the correlation as negative.



as x increases, y generally increases



as x increases, y generally decreases



no clear relationship between x and y

The variable to be plotted on the *x*-axis is called the **independent variable**; it is the variable that can be controlled by the experimenter.

The variable to be plotted on the *y*-axis is known as the **dependent variable**.

Measures of correlation

Pearson's product-moment correlation coefficient (r_p)

PPMCC : r_p

Which diagram shows a stronger correlation?

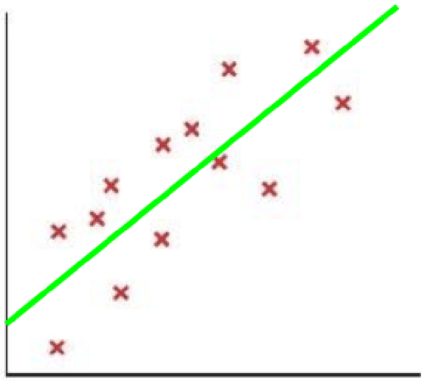


Diagram A

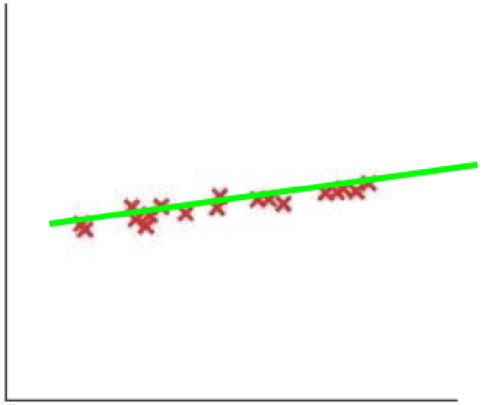


Diagram B

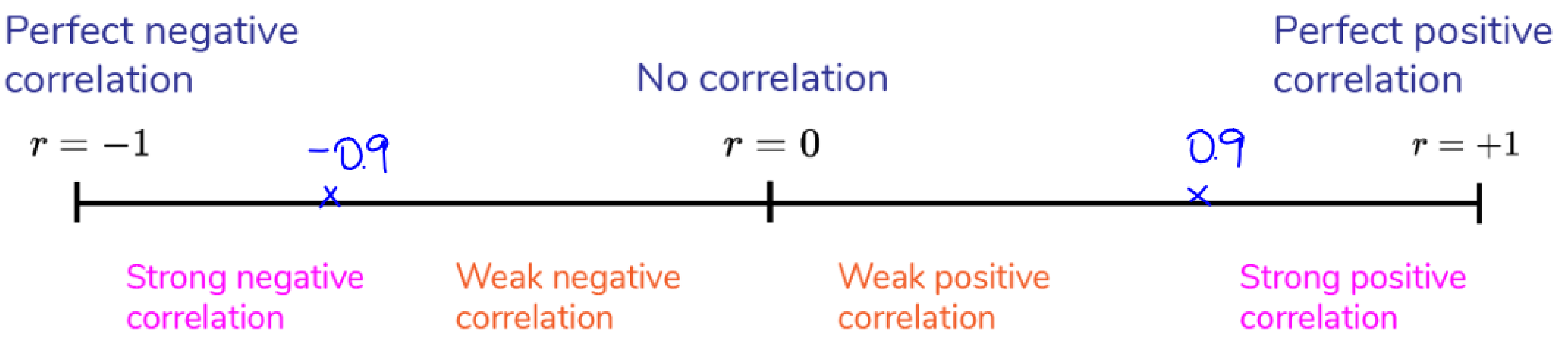
In fact, both diagrams use the exact same data; only the scales are different.
For this reason, we should try to quantify the strength and direction of an association.

This is exactly what Pearson’s product-moment correlation coefficient aims to do.

By convention, we refer to this statistic simply as r.

Facts about r

The value of r is always in betweeen -1 and 1.
r has no units,and is not a percentage.
The sign of r tells us the direction of the correlation: positive or negative or zero.
The size of r tells us the strength of the linear correlation,as shown.



Perfect negative correlation

No correlation

Perfect positive correlation

$r = -1$ -0.9 $r = 0$ 0.9 $r = +1$

Strong negative correlation

Weak negative correlation

Weak positive correlation

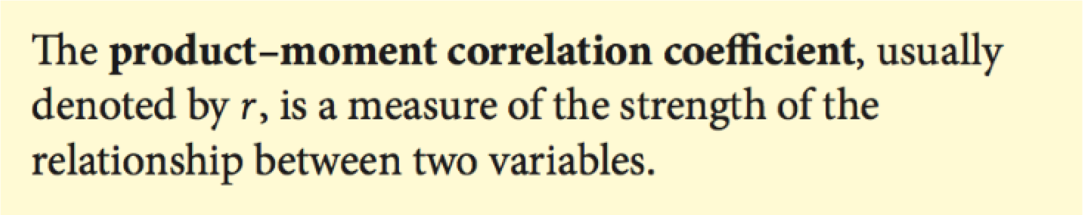
Strong positive correlation

if r = +1, there is a perfect positive linear correlation; all the points fall on a line with positive slope.
If r = 0, there is zero correlation.
If r =-1, there is a perfect negative linear correlation; all the points fall on a line with negative slope.

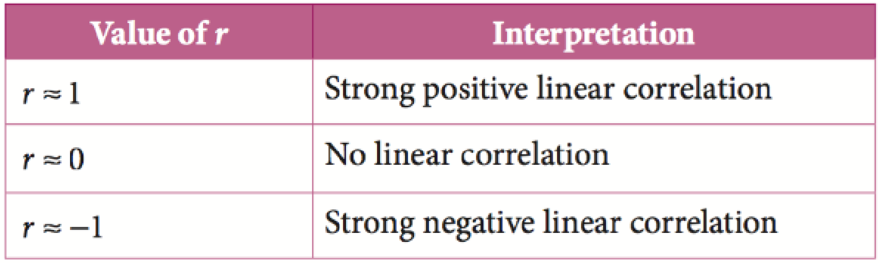
What can go wrong?

When using linear regression models, or any other model, we need to be careful to avoid a few common pitfalls.

- Don't try to predict x from y using the model (especially if the correlation is not strong).
- The model is only as good as the strength of the correlation.
- Don't extrapolate.



The **product–moment correlation coefficient**, usually denoted by r, is a measure of the strength of the relationship between two variables.



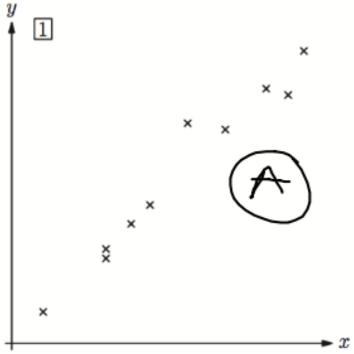
Value of r	Interpretation
$r \approx 1$	Strong positive linear correlation
$r \approx 0$	No linear correlation
$r \approx -1$	Strong negative linear correlation

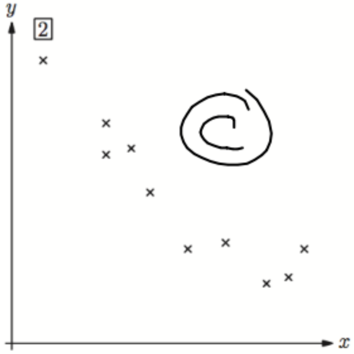
example 263

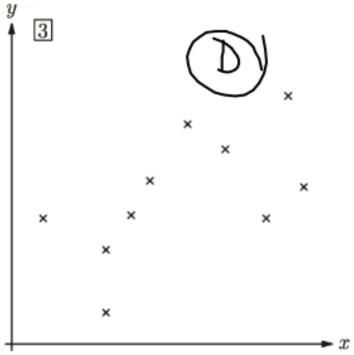
Match the scatter diagrams with the following values of r:

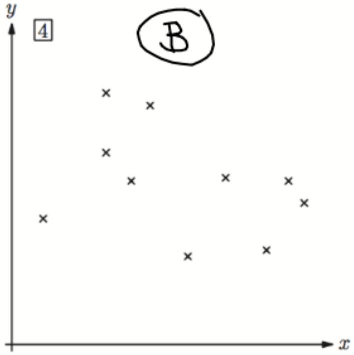
A : $r = 0.98$ B : $r = -0.34$ C : $r = -0.93$ D : $r = 0.58$

$r \approx 0.98$









THE LEAST SQUARES REGRESSION LINE

RESIDUALS

When we fitted a line of best fit by eye, we tried to minimise the distances between the line and each data point.

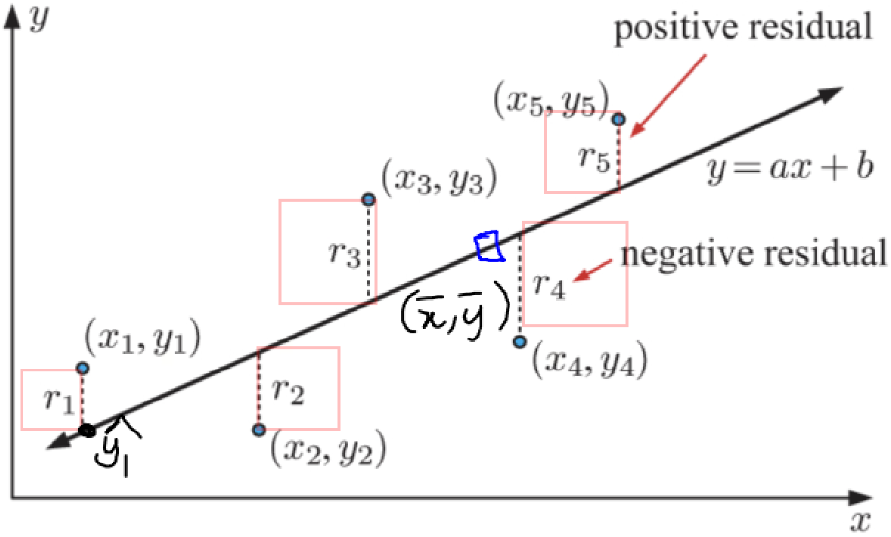
We can formally define each of these distances as a residual:

The **residual** of the ith data point (x_i, y_i) is

$$\boldsymbol{r_i = y_i - \widehat{y}_i}$$

where $\widehat{y}_i$ is the predicted value of y at $x = x_i$.

The goal of least squares linear regression is to minimise the sum of the squared residuals.

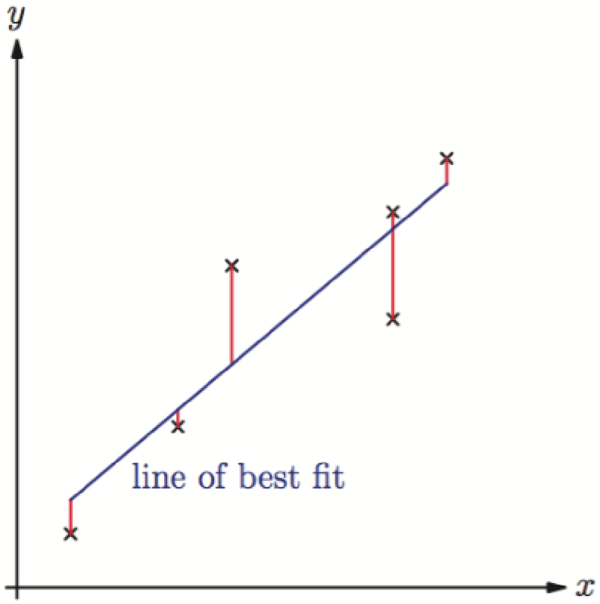


$$SS_{\text{res}} = \sum_{i=1}^{n} {r_i}^2$$

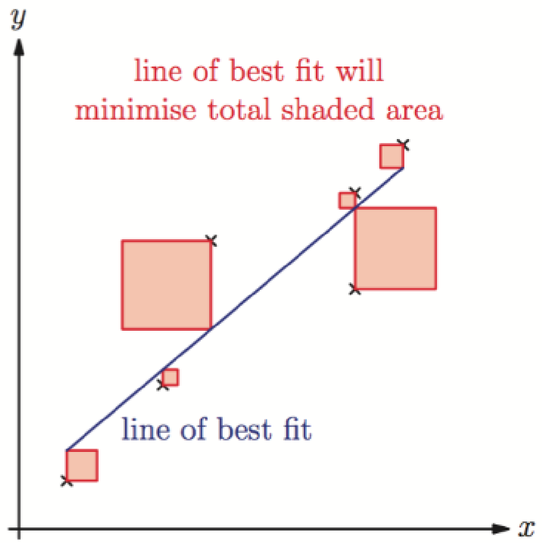
$$= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

Linear regression

Suppose we have established that there is a linear relationship between two variables. We would then want to find the equation that best describes this relationship. To do this we use a method called *least-squares regression*.



We could add up these distances and then try to minimise the total distance by varying the gradient and intercept of the line, but it turns out to be better to minimise the sum of the *squares* of these distances.



To do this minimisation requires some fairly advanced calculus. Fortunately, as with the correlation coefficient r, you can use your calculator to obtain the equation of this **line of best fit**, also referred to as a **regression line**.

example 264

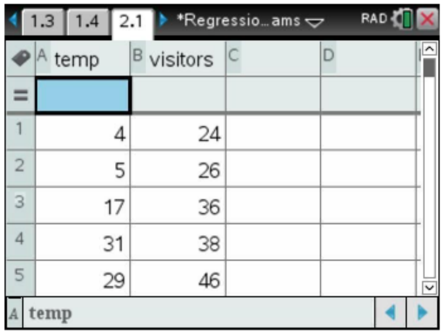
The maximum temperature T, in degrees Celsius, in a park on six randomly selected days is shown in the following table. The table also shows the number of visitors, N, to the park on each of those six days.

Maximum temperature (T)	4	5	17	31	29	11
Number of visitors (N)	24	26	36	38	46	28

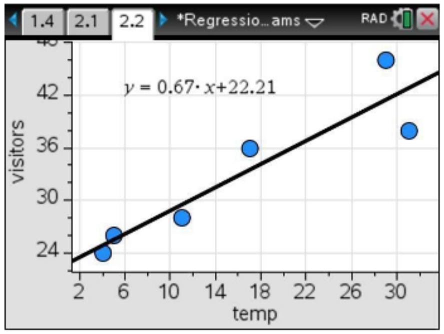
The relationship between the variables can be modelled by the regression equation $N = aT + b$.

(a) (i) Find the value of a and of b.

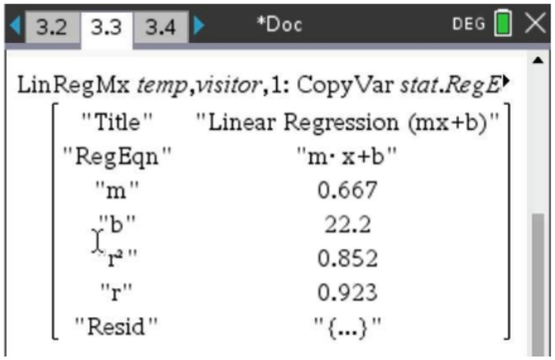
(ii) Write down the value of r. [4]



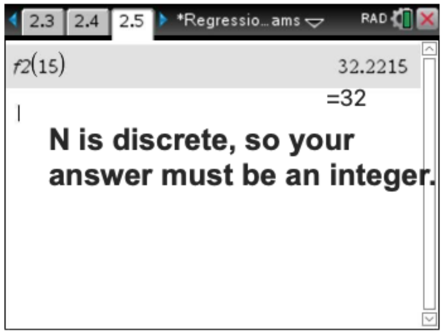
	A temp	B visitors	C	D
=				
1	4	24		
2	5	26		
3	17	36		
4	31	38		
5	29	46		



(b) Use the regression equation to estimate the number of visitors on a day when the maximum temperature is 15 °C. [3]



```
LinRegMx temp,visitor,1: CopyVar stat.RegE▸
    "Title"     "Linear Regression (mx+b)"
  "RegEqn"            "m· x+b"
     "m"               0.667
     "b"               22.2
     "r²"              0.852
     "r"               0.923
   "Resid"            "{...}"
```



```
f2(15)                 32.2215
                    =32
```

N is discrete, so your answer must be an integer.

example

265

An environmental group records the numbers of coyotes and foxes in a wildlife reserve after t years, starting on 1 January 1995.

Let c be the number of coyotes in the reserve after t years. The following table shows the number of coyotes after t years.

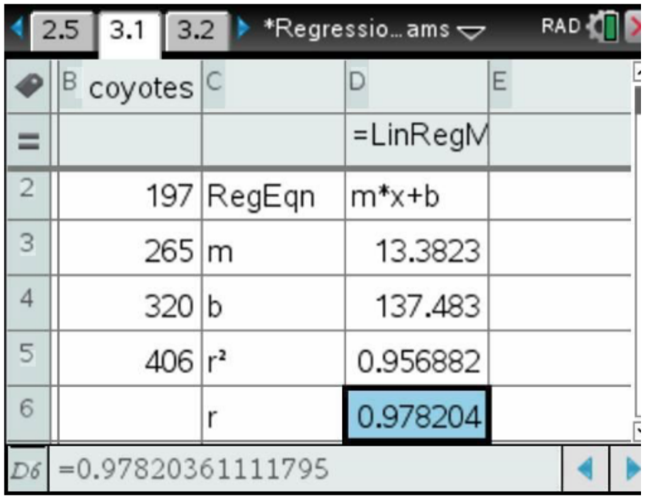
number of years (t)	0	2	10	15	19
number of coyotes (c)	115	197	265	320	406

The relationship between the variables can be modelled by the regression equation $c = at + b$.

(a) Find the value of a and of b. [3]

$a \approx 13.4$

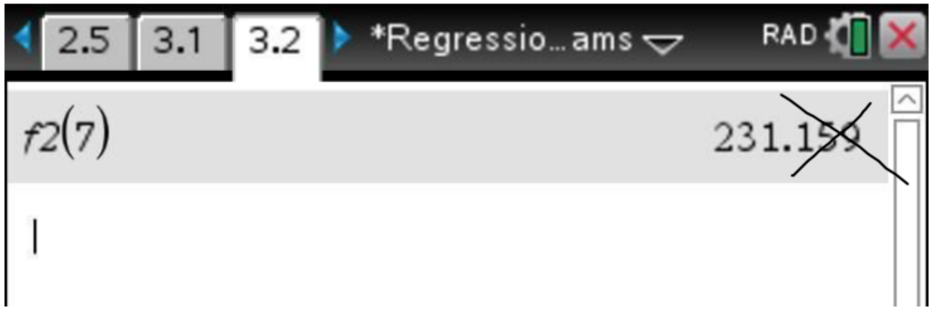
$b \approx 137$



	B coyotes	C	D	E
=			=LinRegM	
2	197	RegEqn	m*x+b	
3	265	m	13.3823	
4	320	b	137.483	
5	406	r^2	0.956882	
6		r	0.978204	

D6 =0.97820361111795

(b) Use the regression equation to estimate the number of coyotes in the reserve when $t = 7$. [3]



f2(7) ~~231.159~~

example 266

The following table shows the average number of hours per day spent watching television by seven mothers and each mother's youngest child.

Hours per day that a mother watches television (x)	2.5	3.0	3.2	3.3	4.0	4.5	5.8
Hours per day that her child watches television (y)	1.8	2.2	2.6	2.5	3.0	3.2	3.5

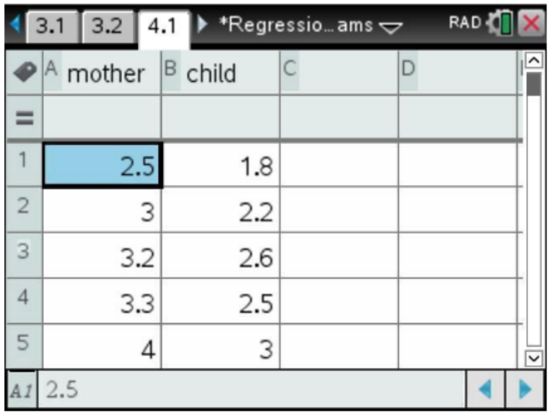
The relationship can be modelled by the regression line with equation $y = ax + b$.

(a) (i) Find the correlation coefficient.

(ii) Write down the value of a and of b. [4]

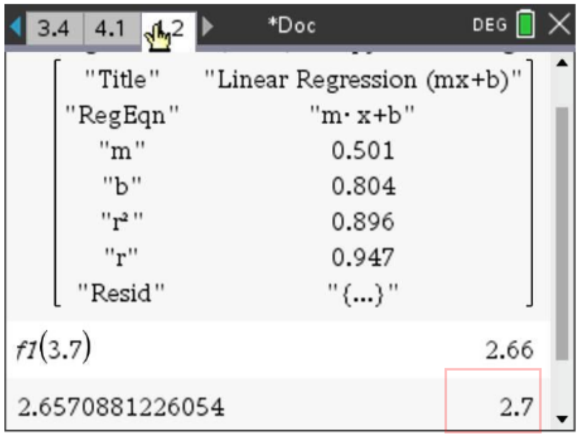
Elizabeth watches television for an average of 3.7 hours per day.

(b) Use your regression line to predict the average number of hours of television watched per day by Elizabeth's youngest child. Give your answer correct to one decimal place. [3]



	A mother	B child	C	D
=				
1	2.5	1.8		
2	3	2.2		
3	3.2	2.6		
4	3.3	2.5		
5	4	3		

A1 2.5



```
"Title"     "Linear Regression (mx+b)"
"RegEqn"    "m·x+b"
"m"         0.501
"b"         0.804
"r²"        0.896
"r"         0.947
"Resid"     "{...}"
f1(3.7)                       2.66
2.6570881226054               2.7
```

example

The price of a used car depends partly on the distance it has travelled. The following table shows the distance and the price for seven cars on 1 January 2010.

Distance, x km	11 500	7500	13 600	10 800	9500	12 200	10 400
Price, y dollars	15 000	21 500	12 000	16 000	19 000	14 500	17 000

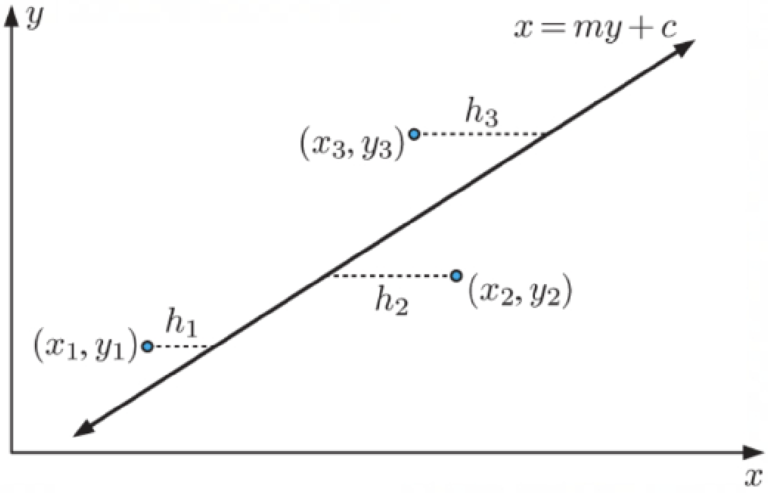
The relationship between x and y can be modelled by the regression equation $y = ax + b$.

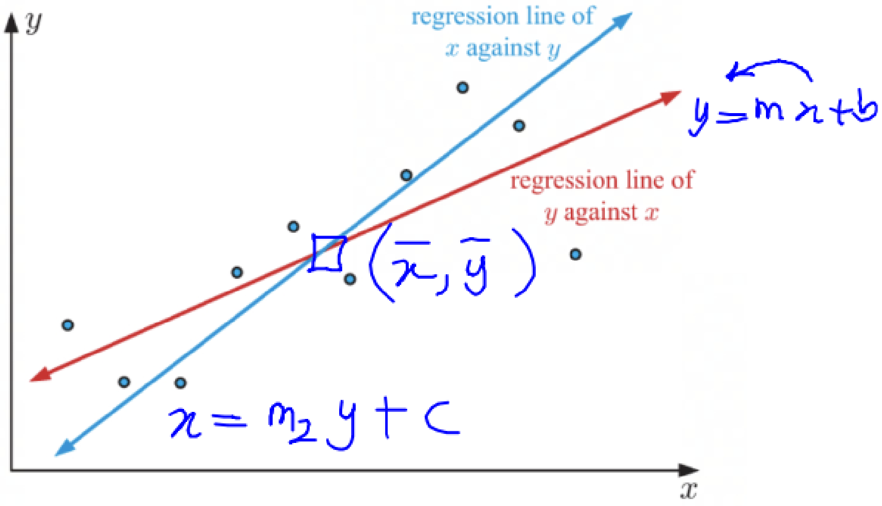
(a) (i) Find the correlation coefficient.

(ii) Write down the value of a and of b. [4]

On 1 January 2010, Lina buys a car which has travelled 11 000 km.

(b) Use the regression equation to estimate the price of Lina's car, giving your answer to the nearest 100 dollars. [3]


y
x = my + c
h_3
(x_3, y_3)
h_2
(x_2, y_2)
h_1
(x_1, y_1)
x


y
regression line of
x against y
y = mx + b
regression line of
y against x
(x̄, ȳ)
x = m_2 y + c
x

example 267

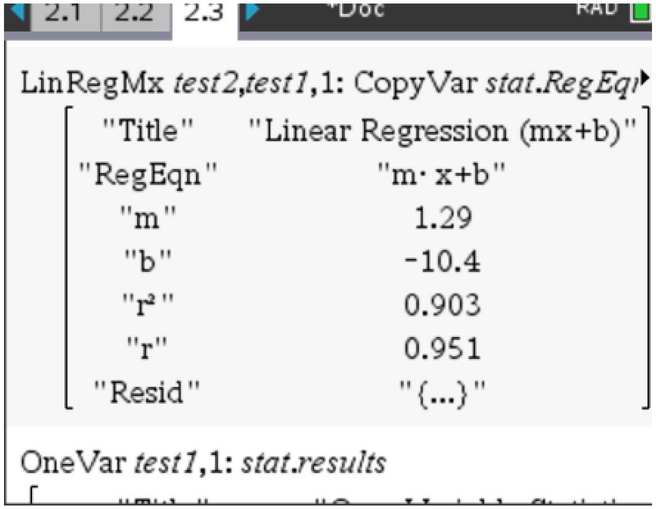
[Maximum mark: 5]

The following table below shows the marks scored by seven students on two different mathematics tests.

Test 1 (x)	15	23	25	30	34	34	40
Test 2 (y)	20	26	27	32	35	37	35

Let L_1 be the regression line of x on y. The equation of the line L_1 can be written in the form $x = ay + b$.

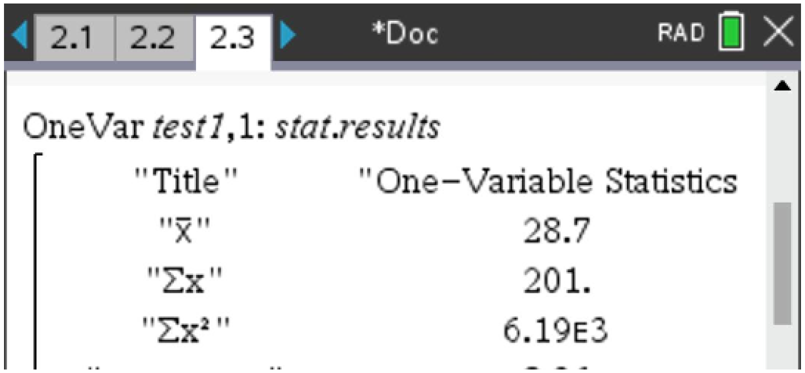
(a) Find the value of a and the value of b. [2]



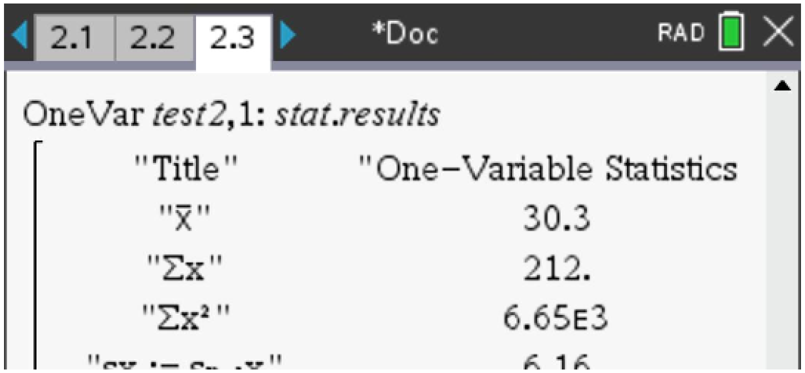
```
LinRegMx test2,test1,1: CopyVar stat.RegEq
   "Title"     "Linear Regression (mx+b)"
  "RegEqn"          "m·x+b"
    "m"               1.29
    "b"              -10.4
    "r²"              0.903
    "r"               0.951
   "Resid"           "{...}"

OneVar test1,1: stat.results
```

Let L_2 be the regression line of y on x. The lines L_1 and L_2 pass through the same point with coordinates (p, q).

(b) Find the value of p and the value of q. [3]



```
OneVar test1,1: stat.results
   "Title"     "One-Variable Statistics
    "x̄"              28.7
    "Σx"             201.
   "Σx²"             6.19E3
```



```
OneVar test2,1: stat.results
   "Title"     "One-Variable Statistics
    "x̄"              30.3
    "Σx"             212.
   "Σx²"             6.65E3
 "sx := Sn-1x"       6.16
```