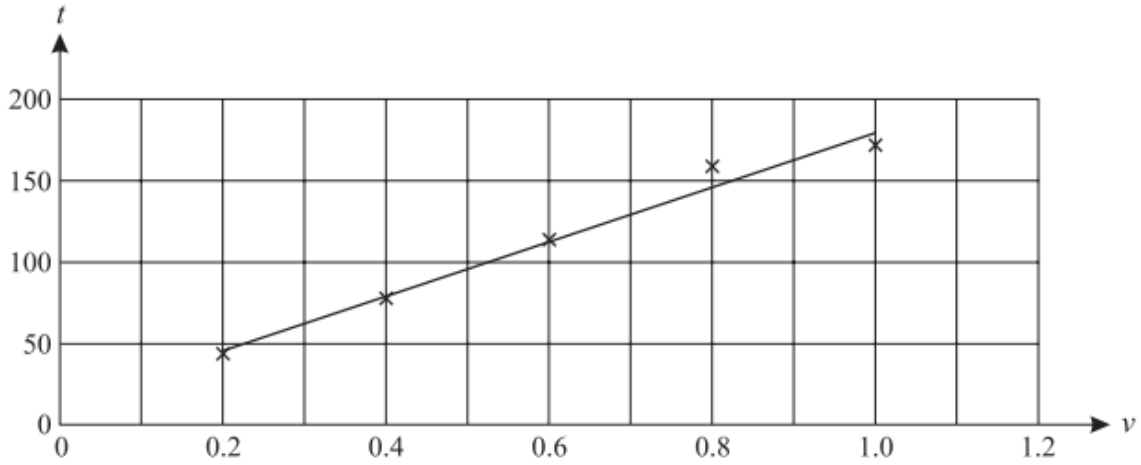
Least Squares Regression Line

Q1, (OCR 4767, Jan 2007, Q1i-iii)

In a science investigation into energy conservation in the home, a student is collecting data on the time taken for an electric kettle to boil as the volume of water in the kettle is varied. The student's data are shown in the table below, where v litres is the volume of water in the kettle and t seconds is the time taken for the kettle to boil (starting with the water at room temperature in each case). Also shown are summary statistics and a scatter diagram on which the regression line of t on v is drawn.

v	0.2	0.4	0.6	0.8	1.0
t	44	78	114	156	172

$n = 5,\ \Sigma v = 3.0,\ \Sigma t = 564,\ \Sigma v^2 = 2.20,\ \Sigma vt = 405.2.$



(i) Calculate the equation of the regression line of t on v, giving your answer in the form $t = a + bv$. [5]

(ii) Use this equation to predict the time taken for the kettle to boil when the amount of water which it contains is

(A) 0.5 litres,

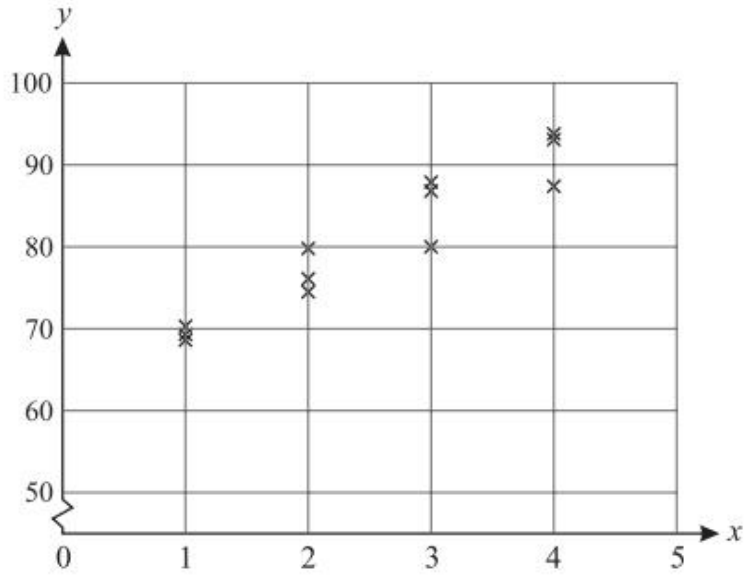
(B) 1.5 litres.

Comment on the reliability of each of these predictions. [4]

(iii) In the equation of the regression line found in part **(i)**, explain the role of the coefficient of v in the relationship between time taken and volume of water. [2]

Q2, (Jan 2008, Q1i-iii)

A biology student is carrying out an experiment to study the effect of a hormone on the growth of plant shoots. The student applies the hormone at various concentrations to a random sample of twelve shoots and measures the growth of each shoot. The data are illustrated on the scatter diagram below, together with the summary statistics for these data. The variables x and y, measured in suitable units, represent concentration and growth respectively.



$$n = 12,\ \Sigma x = 30,\ \Sigma y = 967.6,\ \Sigma x^2 = 90,\ \Sigma y^2 = 78\,926,\ \Sigma xy = 2530.3.$$

(i) State which of the two variables x and y is the independent variable and which is the dependent variable. Briefly explain your answers. [3]

(ii) Calculate the equation of the regression line of y on x. [5]

(iii) Use the equation of the regression line to calculate estimates of shoot growth for concentrations of

(*A*) 1.2,

(*B*) 4.3.

Comment on the reliability of each of these estimates. [4]

Q3, (OCR 4732, Jan 2005, Q9)

Five observations of bivariate data produce the following results, denoted as (x_i, y_i) for $i = 1, 2, 3, 4, 5$.

(13, 2.7) (13, 4.0) (18, 2.8) (23, 3.3) (23, 2.2)

$[\Sigma x = 90,\ \Sigma y = 15.0,\ \Sigma x^2 = 1720,\ \Sigma y^2 = 46.86,\ \Sigma xy = 264.0.]$

(i) Show that the regression line of y on x has gradient -0.06, and find its equation in the form $y = a + bx$. [4]

(ii) The regression line is used to estimate the value of y corresponding to $x = 20$, but the value $x = 20$ is accurate only to the nearest whole number. Calculate the difference between the largest and the smallest values that the estimated value of y could take. [3]

The numbers e_1, e_2, e_3, e_4, e_5 are defined by

$$e_i = a + bx_i - y_i \quad \text{for } i = 1, 2, 3, 4, 5.$$

(iii) The values of e_1, e_2 and e_3 are 0.6, -0.7 and 0.2 respectively. Calculate the values of e_4 and e_5. [2]

(iv) Calculate the value of $e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2$ and explain the relevance of this quantity to the regression line found in part **(i)**. [2]

(v) Find the mean and the variance of e_1, e_2, e_3, e_4, e_5. [4]

Q4, (OCR 4732, Jan 2009, Q2)

The table shows the age, x years, and the mean diameter, y cm, of the trunk of each of seven randomly selected trees of a certain species.

Age (x years)	11	12	20	28	35	45	51
Mean trunk diameter (y cm)	12.2	16.0	26.4	39.2	39.6	51.3	60.6

$[n = 7,\ \Sigma x = 202,\ \Sigma y = 245.3,\ \Sigma x^2 = 7300,\ \Sigma y^2 = 10\,510.65,\ \Sigma xy = 8736.9.]$

(i) (a) Use an appropriate formula to show that the gradient of the regression line of y on x is 1.13, correct to 2 decimal places. **[2]**

(b) Find the equation of the regression line of y on x. **[2]**

(ii) Use your equation to estimate the mean trunk diameter of a tree of this species with age

(a) 30 years, **[1]**

(b) 100 years. **[1]**

It is given that the value of the product moment correlation coefficient for the data in the table is 0.988, correct to 3 decimal places.

(iii) Comment on the reliability of each of your two estimates. **[2]**

Q5, (OCR 4767, Jan 2013, Q1)

A manufacturer of playground safety tiles is testing a new type of tile. Tiles of various thicknesses are tested to estimate the maximum height at which people would be unlikely to sustain injury if they fell onto a tile. The results of the test are as follows.

Thickness (t mm)	20	40	60	80	100
Maximum height (h m)	0.72	1.09	1.62	1.97	2.34

(i) Draw a scatter diagram to illustrate these data. **[3]**

(ii) State which of the two variables is the independent variable, giving a reason for your answer. **[1]**

(iii) Calculate the equation of the regression line of maximum height on thickness. **[5]**

(iv) Use the equation of the regression line to calculate estimates of the maximum height for thicknesses of

(A) 70 mm,

(B) 120 mm.

Comment on the reliability of each of these estimates. **[4]**

Q6, (OCR 4767, Jun 2015, Q1i,ii,iv,v)

A random sample of wheat seedlings is planted and their growth is measured. The table shows their average growth, y mm, at half-day intervals.

Time t days	0	0.5	1	1.5	2	2.5	3
Average growth y mm	0	7	21	33	45	56	62

(i) Draw a scatter diagram to illustrate these data. **[3]**

(ii) Calculate the equation of the regression line of y on t. **[5]**

(iv) Use the equation of the regression line to calculate an estimate of the average growth after 5 days for wheat seedlings. Comment on the reliability of this estimate. **[2]**

It is suggested that it would be better to replace the regression line by a line which passes through the origin.

You are given that the equation of such a line is $y = at$, where $a = \dfrac{\Sigma yt}{\Sigma t^2}$.

(v) Find the equation of this line and plot the line on your scatter diagram. **[4]**

Q7, (OCR 4732, Jun 2016, Q2ii)

The table shows the amount, x, in hundreds of pounds, spent on heating and the number of absences, y, at a factory during each month in 2014.

Amount, x, spent on heating (£ hundreds)	21	23	19	15	14	5	2	10	9	20	18	23
Number of absences, y	23	25	18	18	12	10	4	9	11	15	20	26

$n = 12$ $\Sigma x = 179$ $\Sigma x^2 = 3215$ $\Sigma y = 191$ $\Sigma y^2 = 3565$ $\Sigma xy = 3343$

(ii) The months in 2014 were numbered 1, 2, 3, …, 12. The output, z, in suitable units was recorded along with the month number, n, for each month in 2014. The equation of the regression line of z on n was found to be $z = 0.6n + 17$.

(a) Use this equation to explain whether output generally increased or decreased over these months. **[1]**

(b) Find the mean of n and use the equation of the regression line to calculate the mean of z. **[3]**

(c) Hence calculate the total output in 2014. **[2]**