Regression

Specifications

Correlation and Regression

Calculation of least squares regression lines with one explanatory variable. Scatter diagrams and drawing a regression line thereon.	Where raw data are given, candidates should be encouraged to obtain gradient and intercept values directly from calculators. Where summarised data are given, candidates may be required to use formulae from the booklet provided for the examination. Practical interpretation of values for the gradient and intercept. Use of line for prediction within range of observed values of explanatory variable. Appreciation of the dangers of extrapolation.
Calculation of residuals.	Use of $(\text{residual})_i = y_i - a - bx_i$. Examination of residuals to check plausibility of model and to identify outliers. Appreciation of the possible large influence of outliers on the fitted line.
Linear scaling.	Artificial questions requiring linear scaling will not be set, but candidates should be aware of the effect of linear scaling in correlation and regression.

Given formulae:

For a set of n pairs of values (x_i, y_i)

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

The regression coefficient of y on x is $b = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

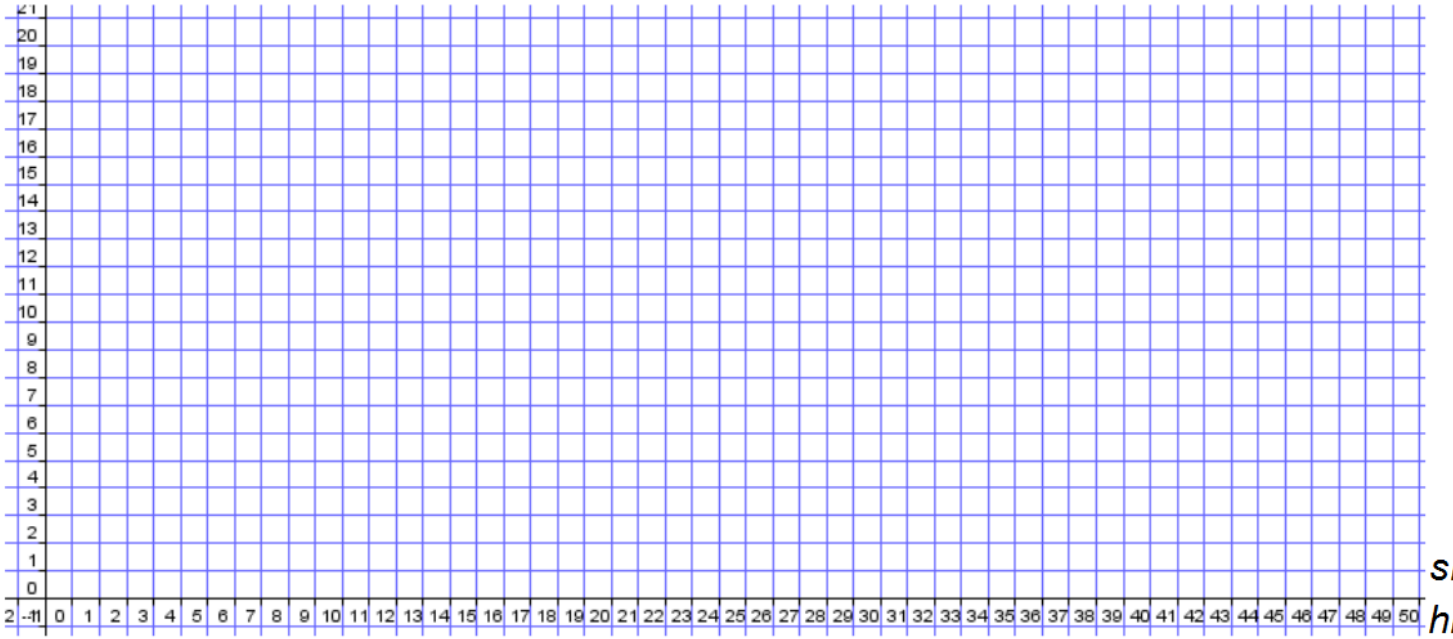
Least squares regression line of y on x is $y = a + bx$, where $a = \bar{y} - b\bar{x}$

Introduction

The data in the table refer to a chain of shops in the London area.
The figures reported are the numbers of sales staff (x)
and the average daily takings in thousands of pounds (y)
for a random sample of shops.

															Mean
x	17	39	32	17	25	43	25	32	48	10	48	42	36	30	19	
y	7	17	10	5	7	15	11	13	19	3	17	15	14	12	8	

1) Plot the data on a scatter diagram and verify that there is
an approximate linear relation between x and y.
using the calculator, work out the correlation coefficient "r".



show
hide

2) Work out the mean $\bar{x}$ and $\bar{y}$. Plot the point $(\bar{x},\bar{y})$

3) Draw "the line of best fit" (going through the mean point).

Extension:

Work out

$n =$ $\sum x =$

$\sum x^2 =$ $\sum xy =$

$S_{xy} =$ $S_{xx} =$

The least squares regression line ("y on x")

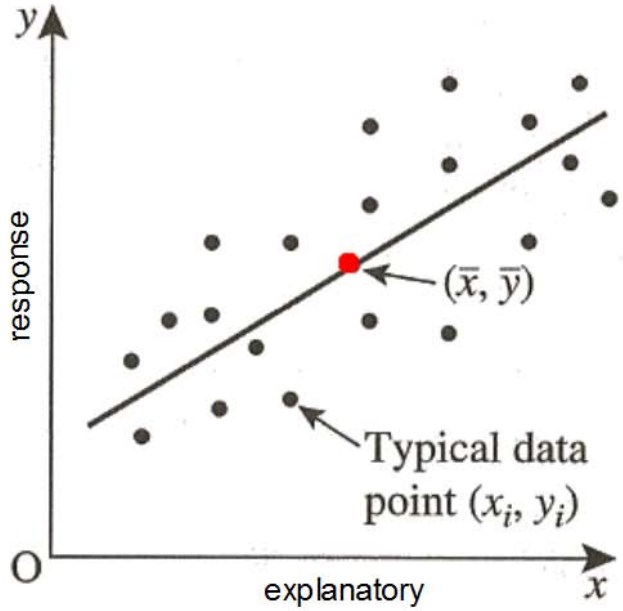
In statistics, a linear relationship between x and y is noted:

$$y = a + bx$$

b is the gradient, its value is $b = \dfrac{S_{xy}}{S_{xx}}$

and

a is the y-intercept, its value is $a = \bar{y} - b\bar{x}$



Method to work out the equation of the least squares regression line:

To find the least squares regression line of y on x:

- Calculate S_{xx} and S_{xy}.
- From these values, work out b.
- Calculate $\bar{x}$ and $\bar{y}$.
- From these values, work out a.

Advice to plot the regression line

The point $(\bar{x}, \bar{y})$ belongs to the regression line.

To complete plotting the line accurately, one or two other points should be plotted.

Any suitable value for x can be chosen but they need to be spread out over the given range.

Explanation:

																Mean
x	17	39	32	17	25	43	25	32	48	10	48	42	36	30	19	*30.9*
y	7	17	10	5	7	15	11	13	19	3	17	15	14	12	8	*11.5*

regression line: y = -0.324 + 0.384x

For a given value of x, we will note

y for the actual value recorded

$\hat{y}$ for the estimate of y, worked out with the regression line equation

Go back to the introduction exercise, work out the equation of the regression line and plot it.
How close were you?

Exercises: (Plot the points on the next page)

1 A random sample of eight pairs of (x, y) values are given in the table.

x	1.2	0.5	0.8	0.1	2.3	1.1	1.8	2.2
y	8.1	4.3	7.1	3.5	12.8	8.4	9.9	11.4

a) Plot a scatter diagram.

b) Find the coordinates of the point $(\bar{x}, \bar{y})$ and mark the point on your scatter diagram.

c) Find the values of a and b for the least-squares regression line $y = a + bx$.

d) Draw the regression line on your diagram, verifying that it passes through the point $(\bar{x}, \bar{y})$.

---

2 A random sample of six pairs of (g, h) values are given in the table.

g	55.7	10.4	67.1	91.2	30.8	72.1
h	21.2	45.9	88.3	11.4	75.4	21.4

a) Plot a scatter diagram, with values of g on the horizontal axis.

b) Find the coordinates of the point $(\bar{g}, \bar{h})$ and mark the point on your scatter diagram.

c) Using your calculator, find the equation of the regression line of h on g.

d) Draw the regression line on your diagram and verify that it passes through the point $(\bar{g}, \bar{h})$.

---

3 A random sample of five pairs of (w, z) values are given in the table.

w	357.2	284.3	435.8	571.9	101.2
z	0.0149	0.0375	−0.0172	−0.0345	0.0651

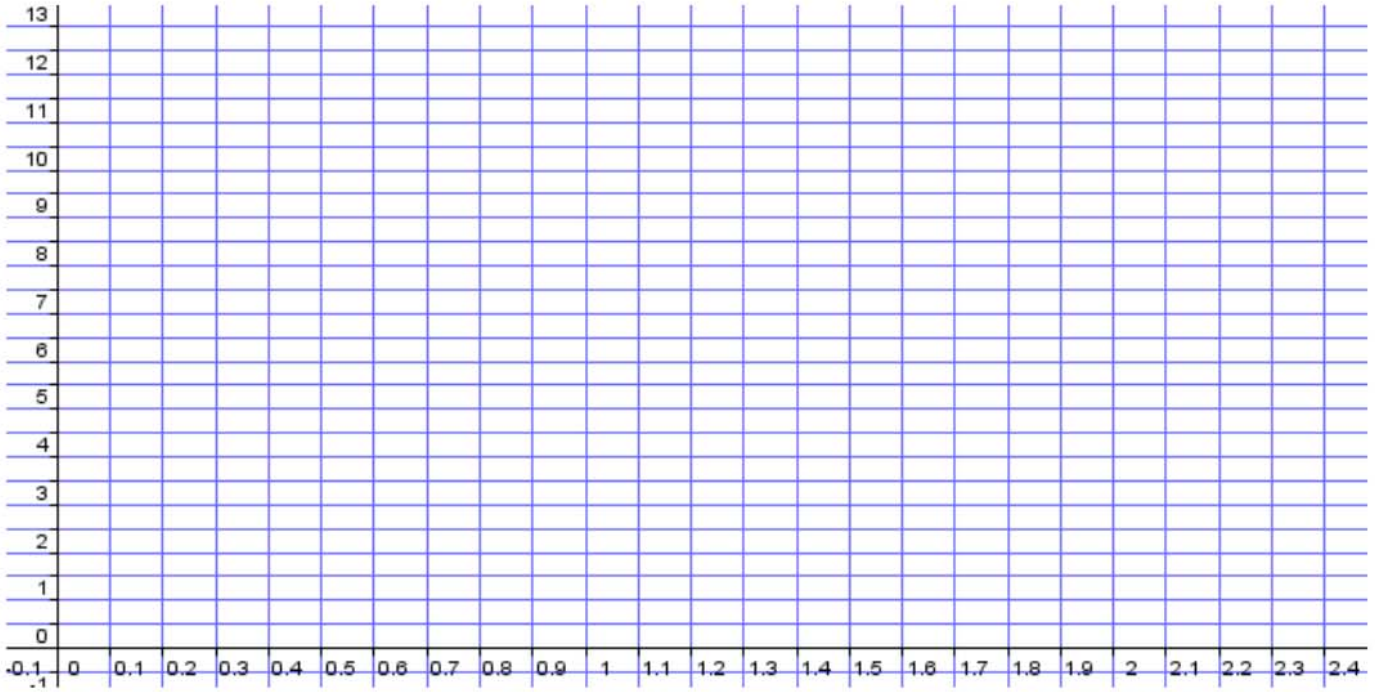
a) Plot a scatter diagram, with values of z on the horizontal axis.

b) Find the coordinates of the point $(\bar{w}, \bar{z})$ and mark the point on your scatter diagram.

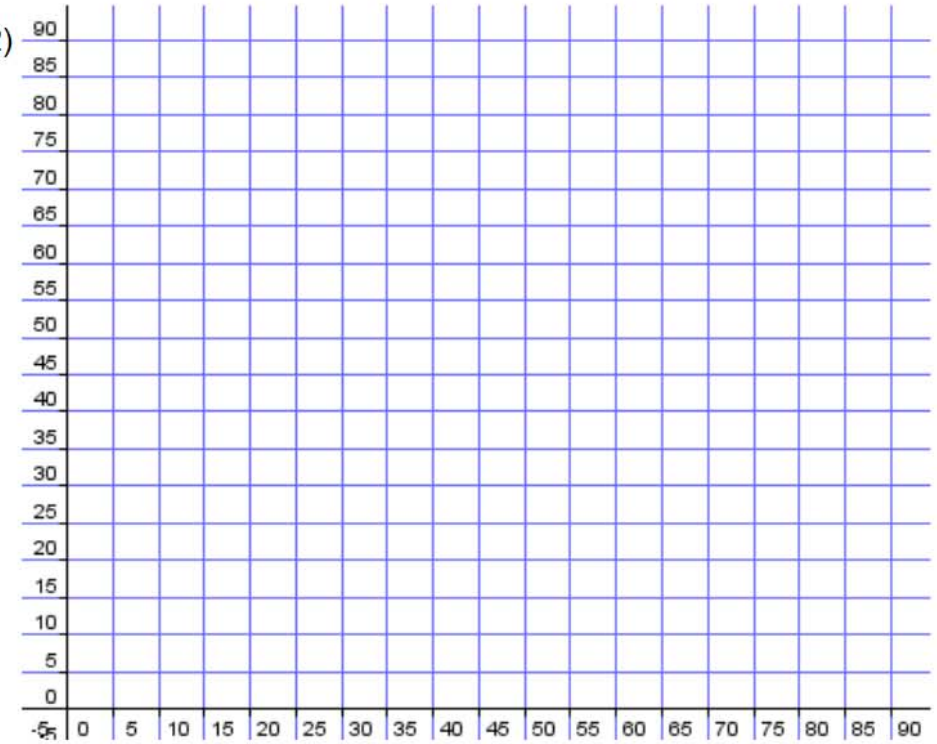
c) Using your calculator, find the equation of the regression line of w on z.

d) Draw the regression line on your diagram and verify that it passes through the point $(\bar{w}, \bar{z})$.

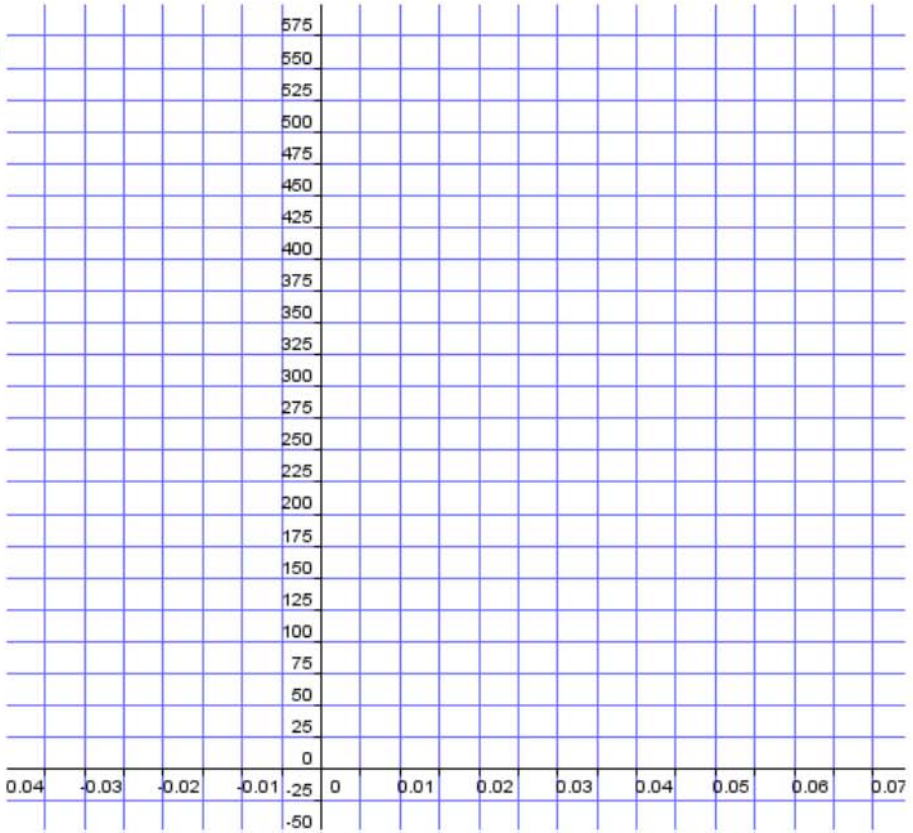
1)

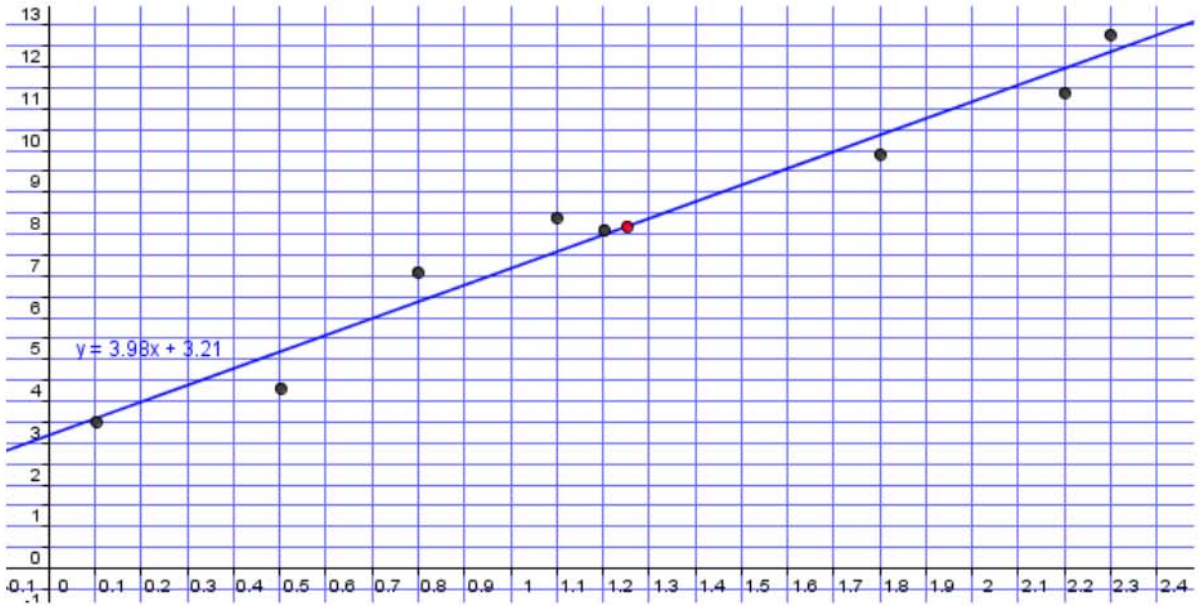

13
12
11
10
9
8
7
6
5
4
3
2
1
0
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2
2.1
2.2
2.3
2.4

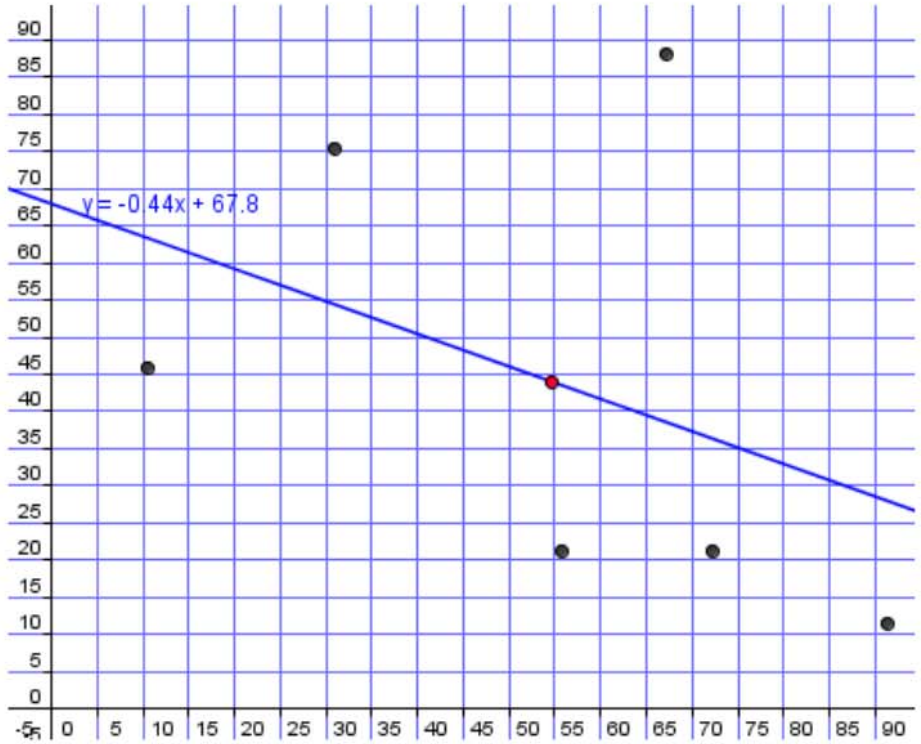
2)

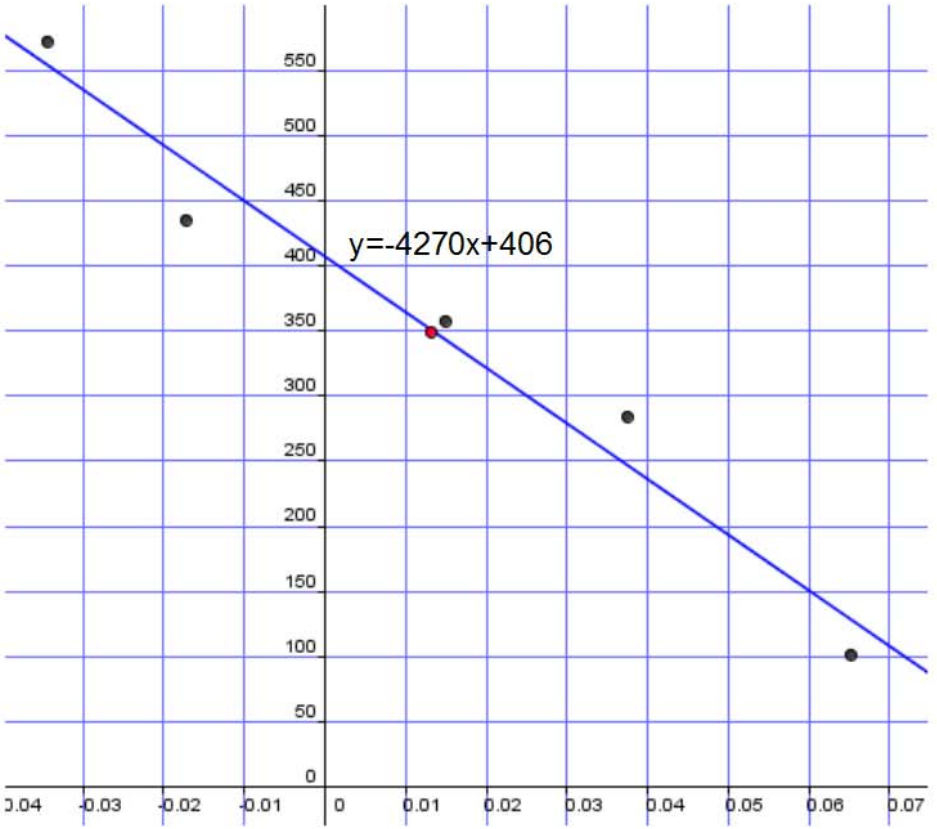

90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90

3)


575
550
525
500
475
450
425
400
375
350
325
300
275
250
225
200
175
150
125
100
75
50
25
0
-25
-50
0.04
-0.03
-0.02
-0.01
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07

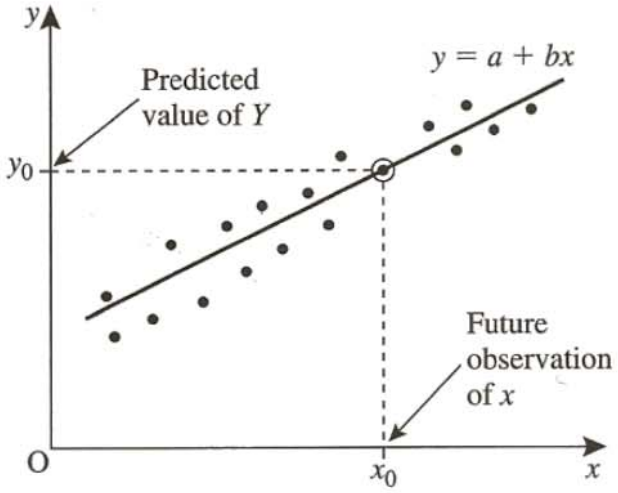

y = 3.98x + 3.21


y = -0.44x + 67.8


y=-4270x+406

Interpolation - Extrapolation

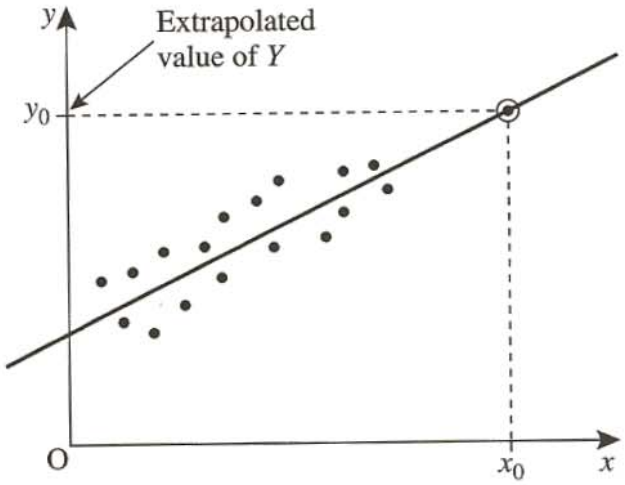
Interpolation or prediction



For this estimate to be a good one, the following criteria should apply.

- The least squares regression line is a good fit to the data.
- The future observation will be taken on the population from which the current data were obtained.
- The value of x for the future observation should lie in (or close to) the range of values of x used in the calculation of the least squares regression line.

Extrapolation



If the third criterion is not satisfied, you are said to be extrapolating. **Extrapolation** can lead to seriously incorrect estimates of y if, in fact, the relation between y and x is approximately linear only in the range of the original x-values.

Example:

A chemist conducts a series of experiments, using the same quantities of reactants in each experiment. Each experiment lasts exactly 30 minutes, at the end of which the amount of compound formed is measured. Each experiment takes place at a different temperature. The results are as follows.

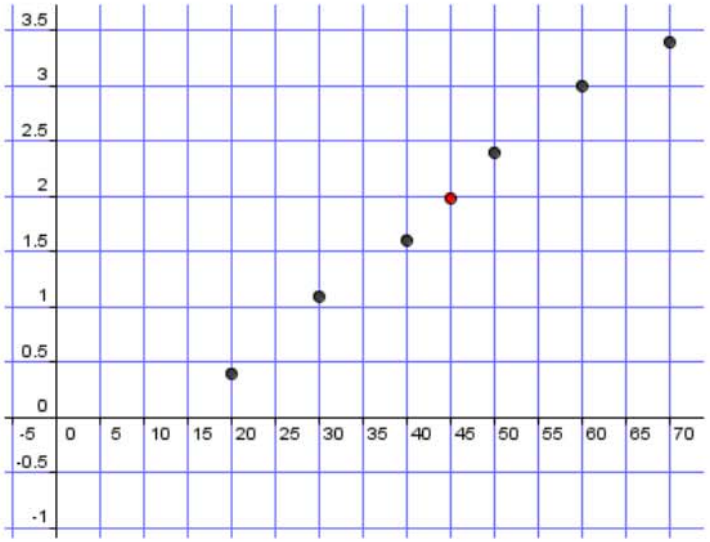
Temperature, x (°C)	20	30	40	50	60	70
Amount formed, y (grams)	0.4	1.1	1.6	2.4	3.0	3.4

a) Plot the data on a scatter diagram and comment thereon.

b) Calculate the equation of the least squares regression line.

c) Use the regression line to estimate the amount of compound that would be formed in 30 minutes at a temperature of 35°C.

d) Use the regression line to estimate the amount of compound that would be formed in 30 minutes at a temperature of 10°C. Comment on your result.



y = - 0.781 + 0.0614x

$x = 35$, $\hat{y} = -0.781 + 0.0614 \times 35 = 1.37$ $(to\ 3sf)$

$x = 10$, $\hat{y} = -0.781 + 0.0614 \times 10 = -0.167$ $(to\ 3sf)$

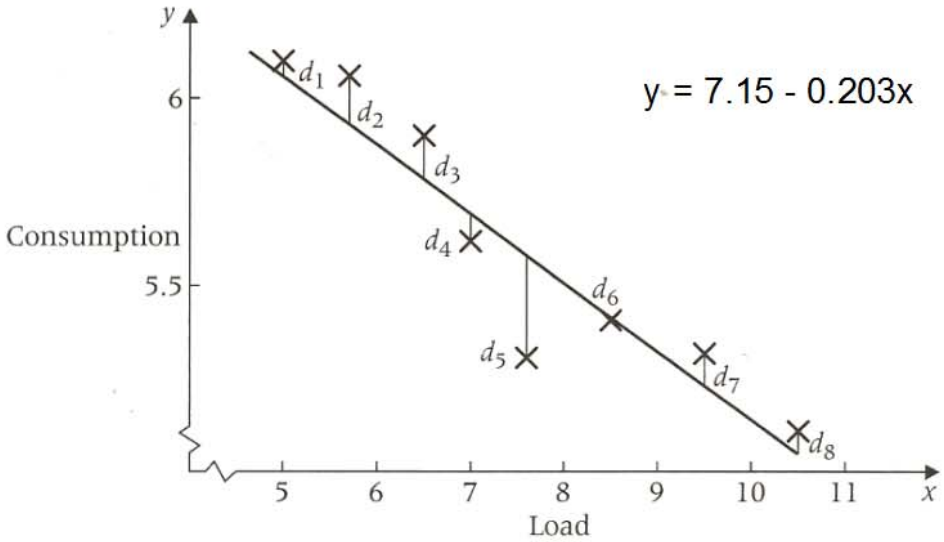
this is impossible, it is a case of extrapolation

(the value of x chosen is outside the given range)

Calculation of residuals

How close to the regression line are the points?

x lorry load (000s kg)	5	5.7	6.5	7	7.6	8.5	9.5	10.5
y fuel consumption (km l^{-1})	6.21	6.12	5.90	5.62	5.25	5.41	5.32	5.11



For a given value of x, the difference between the response value y and the estimate/prediction $\hat{y}$ (using the equation of the regression line) is called the RESIDUAL:

$$d = y - \hat{y}$$

$$d = y - (a + bx)$$

The residual associated with the data point (x_i, y_i) is d_i, given by:

$$d_i = y_i - (a + bx_i) = y_i - a - bx_i$$

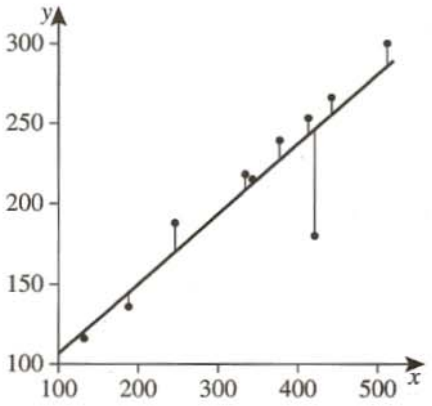
A large (positive or negative) value of d_i therefore indicates a data point which is not well 'explained' by the regression line.

When the magnitude of the residual associated with one data point is much larger than that for any of the others, the data point is described as being an **outlier**.

An **outlier** is a data point having a relatively large residual.

Residuals can be positive (points **above** line) like d_2, d_3, d_7 or negative (points **below** line) like d_4, d_5.

Occasionally, a point might lie exactly on the line.



More practice:

1 The heart and body mass of 14 10-month-old male mice are given in the following table.

Body mass, x (g)	Heart mass, y (mg)
27	118
30	136
37	156
38	150
32	140
36	155
32	157
32	114
38	144
42	159
36	149
44	170
33	131
38	160

(a) Draw a scatter diagram of these data.

(b) Calculate the line of regression of heart mass on body mass (y on x).

2 The systolic blood pressure of ten men of various ages are given in the following table.

Age, x (years)	Systolic blood pressure, y (mm Hg)
37	110
35	117
41	125
43	130
42	138
50	146
49	148
54	150
60	154
65	160

(a) Draw a scatter diagram.

(b) Find the line of regression of systolic blood pressure on age.

(c) Use your line to predict the systolic blood pressure for a man who is:
(i) 20 years old,
(ii) 45 years old.

(d) Comment on the likely accuracy of your predictions in **(i)** and **(ii)**.

5 The given data relate to the price and engine capacity of new cars in January 1982.

Car model	Price (£) y	Capacity (cc) x
A	3900	1000
B	4200	1270
C	5160	1750
D	6980	2230
E	6930	1990
F	2190	600
G	2190	650
H	4160	1500
J	3050	1450
K	6150	1650

(a) Plot a scatter diagram of the data.

(b) Calculate the line of regression of y on x.

(c) Draw the line of regression on the scatter diagram.

(d) A particular customer regards large engine capacity and a low price as the two most important factors in choosing a car. Examine your scatter diagram and the regression line to suggest to him one model which, in January 1982, gave good value for money. Also suggest three models which you would advise the customer not to buy. [A]

6 A small firm tries a new approach to negotiating the annual pay rise with each of its 12 employees. In an attempt to simplify the process, it is suggested that each employee should be assigned a score, x, based on his/her level of responsibility. The annual salary will be £$(a + bx)$ and negotiations will only involve the values of a and b.

The following table gives last year's salaries (which were generally regarded as fair) and the proposed scores.

Employee	x	Annual salary (£) y
A	10	5750
B	55	17 300
C	46	14 750
D	27	8200
E	17	6350
F	12	6150
G	85	18 800
H	64	14 850
I	36	9900
J	40	11 000
K	30	9150
L	37	10 400

(a) Plot the data on a scatter diagram.

(b) Estimate the values that could have been used for a and b last year by finding the line of regression of y on x.

(c) Comment on whether the suggested method is likely to prove reasonably satisfactory in practice.

(d) Two employees, B and C, had to work away from home for a large part of the year. In the light of this additional information, suggest an improvement to the model. [A]

Answers

1) b) $y = 48.4 + 2.75x$
2) b) $y = 62.8 + 1.58x$
c) i) $y = 94.3$
ii) $y = 133.7$
d) i) *Extrapolation, not accurate*
ii) *Interpolation, likely to be reasonably accurate*
5) b) $y = 237 + 3.02x$
d) *Model J is recommended*
Discourage models A, E and K
6) b) $y = 3713 + 192x$
c) points are close to the line, (apart B and C).
Model should be resonably satisfactory
d) Salary $= a + bx + t$ where t is an additional payment from employees who have to work away from home.

Key point summary

1 A scatter diagram should be drawn to judge whether linear regression analysis is a sensible option.

2 The nature of the data should be considered to determine which is the *independent* or *explanatory* variable (x) and which is the *dependent* or *response* variable (y).

3 The regression line is found using the *method of least squares* in the form

$$y = \mathbf{a} + \mathbf{b}x$$

This is the regression line of y on x and may be used to predict a value for y from a given value of x.

The equations can be found directly using a calculator with a linear regression mode.

Be careful to note the form in which your calculator presents the equation – it may be as $y = \mathbf{a}x + \mathbf{b}$.

4 Using $y = a + bx$

a estimates the value of y when x is zero.
b estimates the rate of change of y with x.

5 Be very careful when predicting from your line. Watch out for extrapolation when predictions can be wildly inaccurate.

Look back to section 8.8.

Never assume a linear model will keep on going forever.