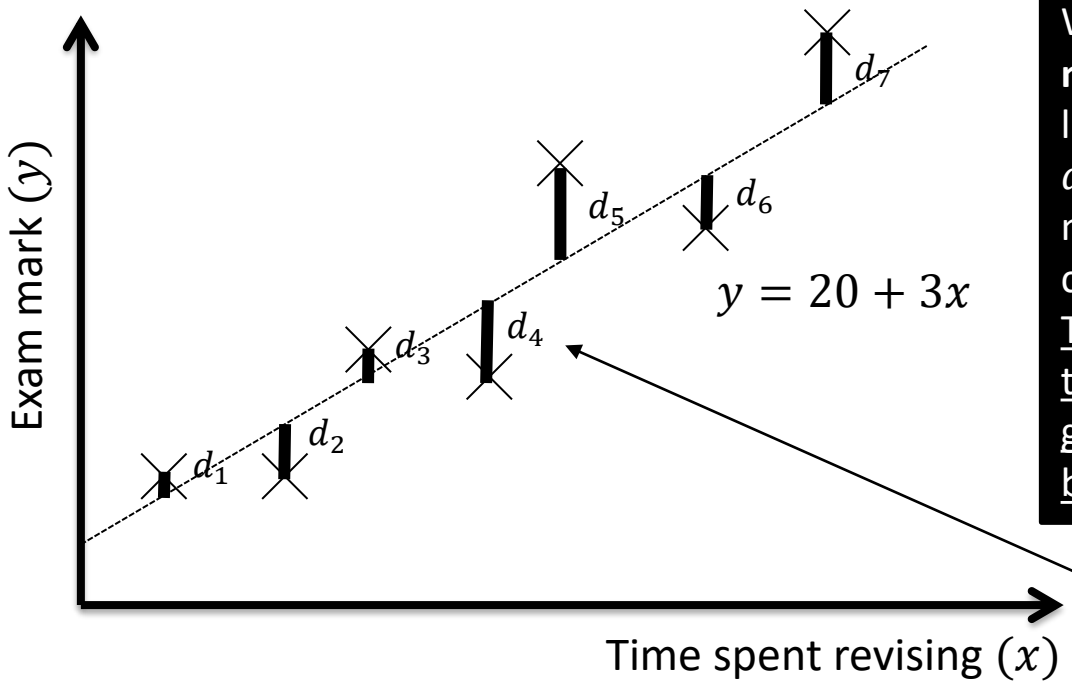
Stats1 Chapter 4: Correlation

Linear Regression

What is regression?



What we've done here is come up with a **model** to explain the data, in this case, a line $y = a + bx$. We've then tried to set a and b such that the resulting y value matches the actual exam marks as closely as possible.
The 'regression' bit is the act of setting the parameters of our model (here the gradient and y-intercept of the line of best fit) to best explain the data.

I record people's exam marks as well as the time they spent revising. I want to predict how well someone will do based on the time they spent revising. How would I do this?

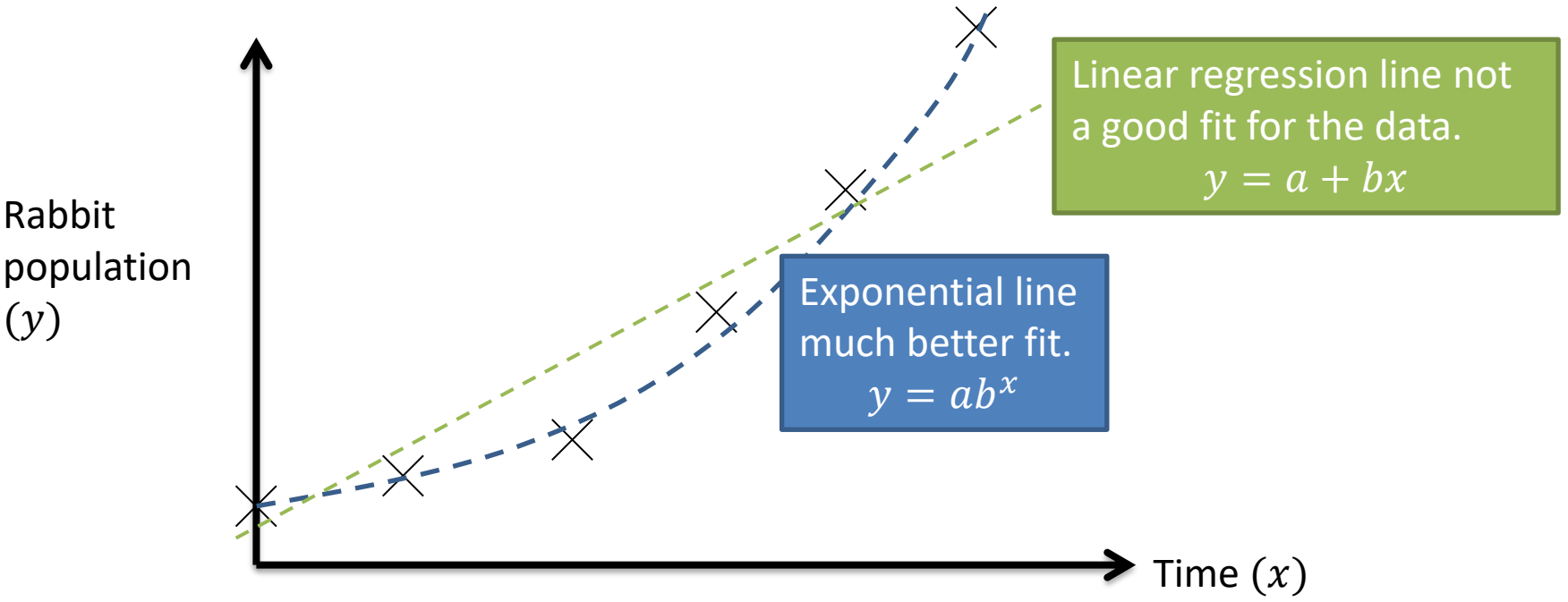
One type of line of best fit is the **least squares regression line**. This minimises the sum of the square of these 'errors', i.e.

$$d_1^2 + d_2^2 + \cdots = \Sigma d_i^2$$

Part of the reason we square these errors is so that each distance is treated as a positive value.

Unlike in the old S1, you are no longer required to work out the equation of the least squares regression line yourself; you will be given the equation.

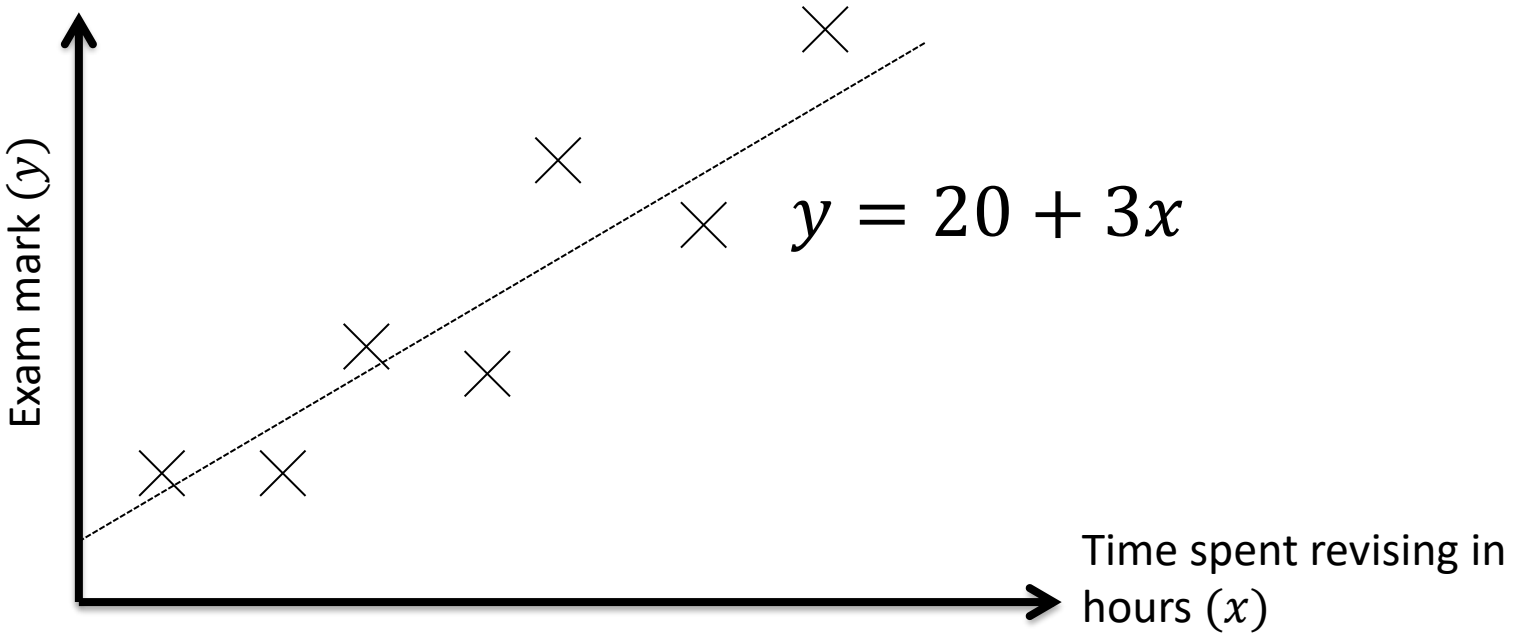
What is regression?



In this chapter we only cover **linear regression**, where our chosen model is a straight line.

But in general we could use any model that might best explain the data. Population tends to grow exponentially rather than linearly, so we might make our model $y = a \times b^x$ and then try to use regression to work out the best a and b to use. **You will do exponential regression in Chapter 14 of Pure Year 1.**

Interpreting a and b.



How do we interpret the gradient of 3?

For each extra hour spent revising, 3 marks are gained.
(i.e. the gradient tells you the change in y for each unit increase in x)

How do we interpret the y-intercept of 20?

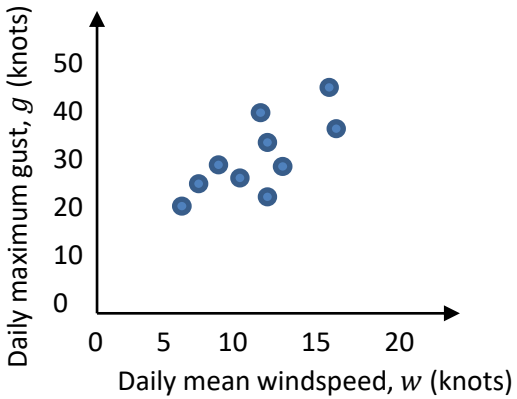
20 marks would be obtained turning up to the exam with no revision.
(i.e. the value of y we get when $x = 0$)

Example

From the large data set, the daily mean windspeed, w knots, and the daily maximum gust, g knots, were recorded for the first 15 days in May in Camborne in 2015.

w	14	13	13	9	18	18	7	15	10	14	11	9	8	10	7
g	33	37	29	23	43	38	17	30	28	29	29	23	21	28	20

The data was plotted on a scatter diagram.



(a) Describe the correlation between daily mean windspeed and daily maximum gust.

The equation of the regression line of g on w for these 15 days is $g = 7.23 + 1.82w$

(b) Give an interpretation of the value of the gradient of this regression line.

(c) Justify the use of a linear regression line in this instance.

a ?

b ?

c ?

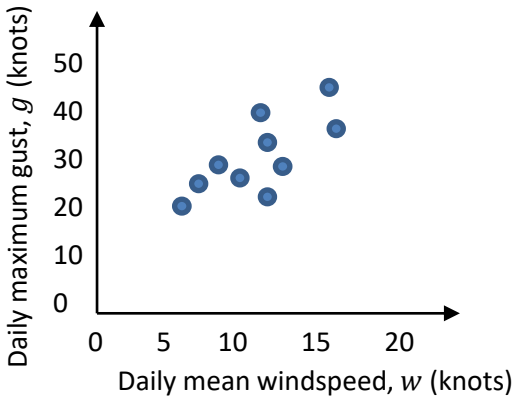
The stronger the (linear) correlation, the more suitable a linear regression line is.

Example

From the large data set, the daily mean windspeed, w knots, and the daily maximum gust, g knots, were recorded for the first 15 days in May in Camborne in 2015.

w	14	13	13	9	18	18	7	15	10	14	11	9	8	10	7
g	33	37	29	23	43	38	17	30	28	29	29	23	21	28	20

The data was plotted on a scatter diagram.



(a) Describe the correlation between daily mean windspeed and daily maximum gust.

The equation of the regression line of g on w for these 15 days is $g = 7.23 + 1.82w$

(b) Give an interpretation of the value of the gradient of this regression line.

(c) Justify the use of a linear regression line in this instance.

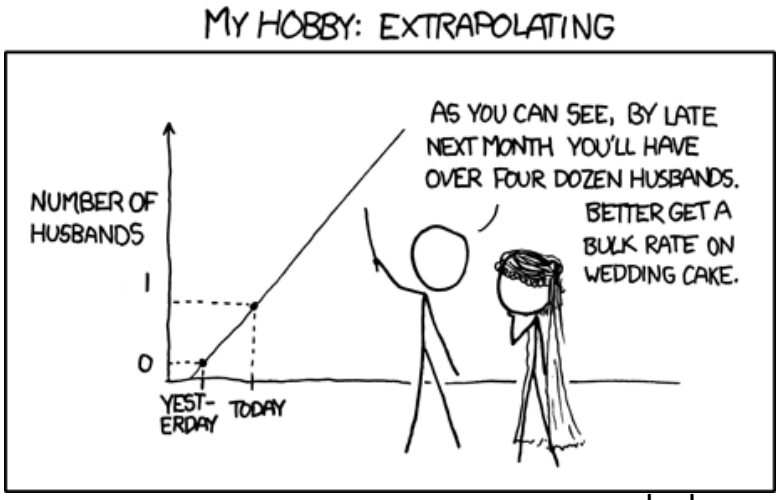
a There is a strong positive correlation between daily mean windspeed and daily maximum gust.

b If the daily mean windspeed increases by 10 knots the daily maximum gust increases by approximately 18 knots.

c The correlation suggests that there is a linear relationship between g and w so a linear regression line is a suitable model.

The stronger the (linear) correlation, the more suitable a linear regression line is.

Interpolating and Extrapolating

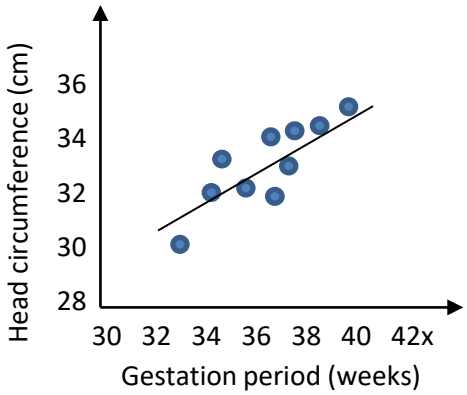


xkcd.com

You should only use the regression line to make predictions for values **of the dependent variable** that are **within the range of the given data**.

Estimating a value inside the data range is known as **interpolating**.
Estimating a value outside the data range is known as **extrapolating**
(as per the cartoon on the left!)

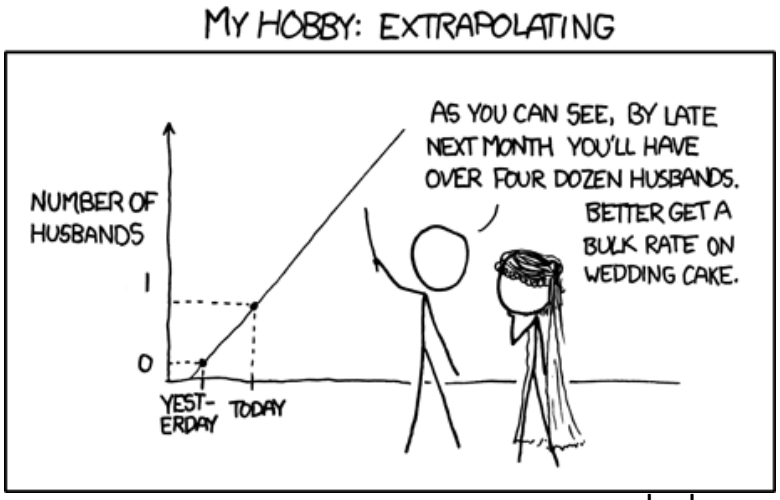
[Textbook] The head circumference, y cm, and gestation period, x weeks, for a random sample of eight newborn babies at a clinic are recorded.
The scatter graph shows the results.
The equation of the regression line of y on x is $y = 8.91 + 0.624x$. The regression equation is used to estimate the head circumference of a baby born at 39 weeks and a baby born at 30 weeks.
(a) Comment on the reliability of these estimates.
A nurse wants to estimate the gestation period for a baby born with a head circumference of 31.6cm.
(b) Explain why the regression equation given above is not suitable for this estimate.



a ?

b ?

Interpolating and Extrapolating

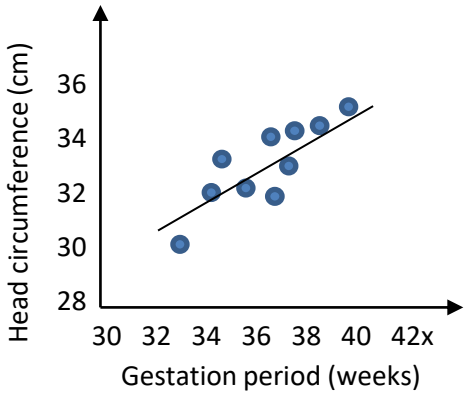


xkcd.com

You should only use the regression line to make predictions for values **of the dependent variable** that are **within the range of the given data**.

Estimating a value inside the data range is known as **interpolating**.
Estimating a value outside the data range is known as **extrapolating**
(as per the cartoon on the left!)

[Textbook] The head circumference, y cm, and gestation period, x weeks, for a random sample of eight newborn babies at a clinic are recorded.
The scatter graph shows the results.
The equation of the regression line of y on x is $y = 8.91 + 0.624x$. The regression equation is used to estimate the head circumference of a baby born at 39 weeks and a baby born at 30 weeks.
(a) Comment on the reliability of these estimates.
A nurse wants to estimate the gestation period for a baby born with a head circumference of 31.6cm.
(b) Explain why the regression equation given above is not suitable for this estimate.



a The prediction for 39 weeks is within the range of the data so is more likely to be correct.
The prediction for 30 weeks is outside the range of the data so is less likely to be accurate.

b The independent variable in this model is the gestation period, x. You should not use this model to predict a value of x for a given value of y.

Exercise 4.2

Pearson Statistics/Mechanics Year 1/AS
Pages 28-29

Homework Exercise

1 An accountant monitors the number of items produced per month by a company together with the total production costs. The table shows this data.

Number of items, n (1000s)	21	39	48	24	72	75	15	35	62	81	12	56
Production costs, p (£1000s)	40	58	67	45	89	96	37	53	83	102	35	75

a Draw a scatter diagram to represent this data.

The equation of the regression line of p on n is $p = 21.0 + 0.98n$.

b Draw the regression line on your scatter diagram.

c Interpret the meaning of the figures 21.0 and 0.98.

The company expects to produce 74 000 items in June, and 95 000 items in July.

d Comment on the suitability of this regression line equation to predict the production costs in each of these months.

2 The relationship between the number of coats of paint applied to a boat and the resulting weather resistance was tested in a laboratory. The data collected is shown in the table.

Coats of paint (x)	**Protection (years) (y)**
1	4.4
2	5.9
3	7.1
4	8.8
5	10.2

a Draw a scatter diagram to represent this data.

The equation of the regression line is $y = 2.93 + 1.45x$.

Helen says that a gradient of 1.45 means that if 10 coats of paint are applied the protection will last 14.5 years.

b Comment on Helen's statement.

3 The table shows the ages of some chickens and the number of eggs that they laid in a month.

Age of chicken, a (months)	18	32	44	60	71	79	99	109	118	140
Number of eggs laid in a month, n	16	18	13	7	12	7	11	13	6	9

a Draw a scatter diagram to show this information.

Robin calculates the regression line of n on a as $n = 16.1 + 0.063a$.

b Without further calculation, explain why Robin's regression equation is incorrect.

Homework Exercise

4 Aisha collected data on the numbers of bedrooms, x, and the values, y (£1000s), of the houses in her village. She calculates the regression equation of y on x to be $y = 190 + 50x$.

She states that the value of the constant in her regression equation means that a house with no bedrooms in her village would be worth £190 000. Explain why this is not a reasonable statement.

5 The table shows the daily maximum relative humidity, h (%), and the daily mean visibility, v decametres (Dm), in Heathrow for the first two weeks in September 2015, from the large data set.

h	94	95	92	80	97	94	93	90	87	95	93	92	91	98
v	2600	2900	3900	4300	2800	2400	2700	3500	3000	2200	2200	3300	2800	2200

The equation of the regression line of v on h is $v = 12\,700 - 106h$

a Give an interpretation of the value of the gradient of the regression line. **(1 mark)**

b Use your knowledge of the large data set to explain whether there is likely to be a causal relationship between humidity and visibility. **(2 marks)**

c Give reasons why it would not be reliable to use this regression equation to predict:

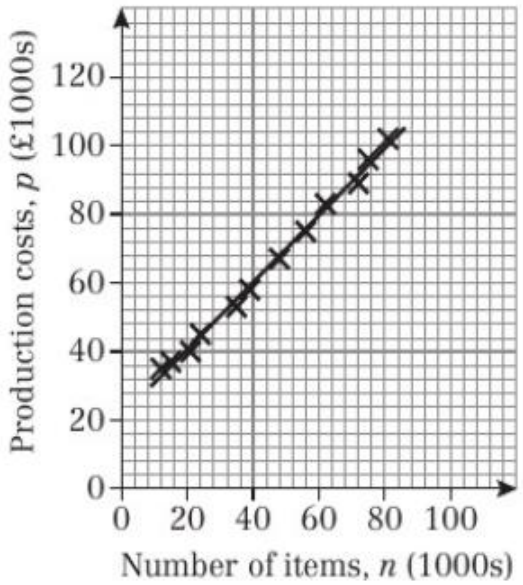
i the mean visibility on a day with 100% humidity **(2 marks)**

ii the humidity on a day with visibility of 3000 dm. **(2 marks)**

d State two ways in which better use could be made of the large data set to produce a model describing the relationship between humidity and visibility. **(2 marks)**

Homework Answers

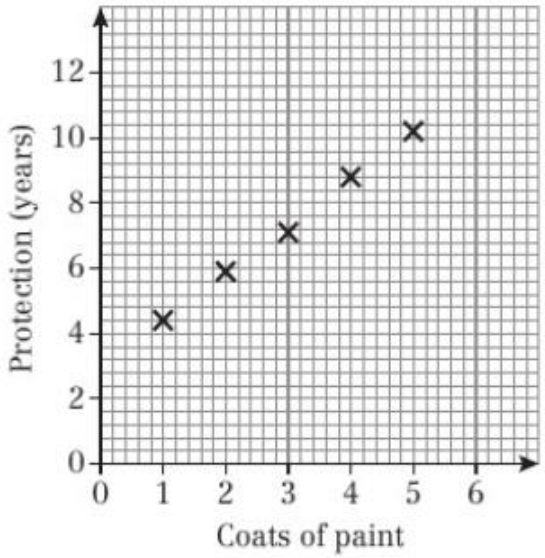
1 a, b



c If the number of items produced per month is zero, the production costs will be approximately £21 000. If the number of items per month increases by 1000 items, the production costs increase by approximately £980.

d The prediction for 74 000 is within the range of the data (interpolation) so is more likely to be accurate. The prediction for 95 000 is outside the range of the data (extrapolation) so is less likely to be accurate.

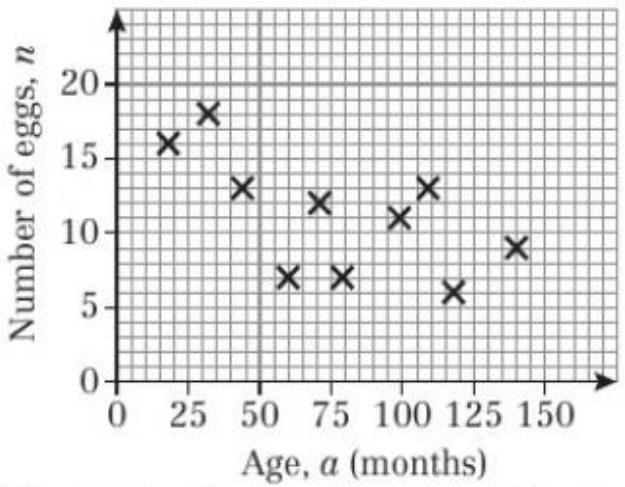
2 a



b A gradient of 1.45 means that for every extra coat of paint, the protection will increase by 1.45 years, therefore if 10 coats of paint are applied, the protection will be 14.5 years longer than if zero coats of paint were applied. After 10 coats of paint, the protection will last 2.93 + 14.5 = 17.43 years.

Homework Answers

3 a



b The scatter diagram shows negative correlation, therefore the gradient in the regression equation should be negative.

4 This is not sensible as there are unlikely to be any houses with no bedrooms.

5 a For each percent increase in daily maximum relative humidity there is a decrease of 106 Dm in daily mean visibility.

b High levels of relative humidity cause mist or fog which will decrease visibility. Hence there is likely to be a causal relationship.

c i The prediction for 100% is outside the range of the data (extrapolation) so is less likely to be accurate.

ii The regression equation should only be used to predict a value for v given h.

d Data is only useful for analysing the first two weeks of September. Random values throughout September should be used and analysis made of the whole month. The sample size could also be increased across multiple months as data between May and October is available.