MATHEMATICAL APPLICATIONS 3

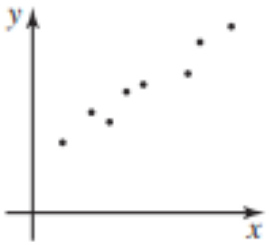
WEEK 4 NOTES & EXERCISES

Fitting straight lines to bivariate data

Fitting a line of 'best fit' to bivariate data (scatterplots) enables us to analyse data and possibly make predictions
Regression analysis is concerned with find these straight lines.
In our previous work when we displayed bivariate data as a scatterplot, the independent variable was placed on the horizontal axis and the dependent variable was placed on the vertical axis. When the relationship between two variables (x and y) is described in equation form, such as $y = mx + c$, the subject, y, is the dependent variable and x is the independent variable.

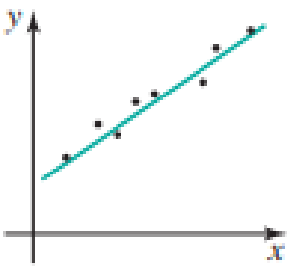
Fitting a straight line by eye

Consider the set of bivariate data points below. In this case the x-values could be heights of married women, while y-values could be the heights of their husbands. We wish to determine a linear relationship between these two random variables.



Of course, there is no single straight line which would go through all the points, so we can only estimate such a line. Furthermore, the more closely the points appear to be on or near a straight line, the more confident we are that such a linear relationship may exist and the more accurate our fitted line should be.

Consider the estimate, drawn 'by eye' in the figure below. It is clear that most of the points are on or very close to this straight line. This line was easily drawn since the points are very much part of an apparent linear relationship. However, note that some points are below the line and some are above it. Furthermore, if x is the height of wives and y is the height of husbands, it seems that husbands are generally taller than their wives. Regression analysis is concerned with finding these straight lines using various methods so that the number of points above and below the lines are 'balanced'.

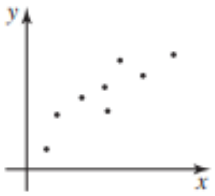


Method of fitting lines by eye

There should be an equal number of points above and below the line. For example, if there are 12 points in the data set, 6 should be above the line and 6 below it. This may appear logical or even obvious, but fitting by eye involves a considerable margin of error.

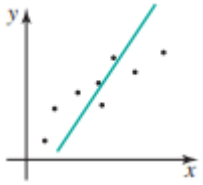
Example

Fit a straight line to the data in the figure at right using the equal-number-of-points method.

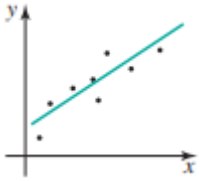


Note that the number of points (n) is 8.

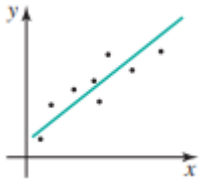
Fit a line where 4 points are above the line.
Using a clear plastic ruler, try to fit the best line.



The first attempt has only 3 points below the line where there should be 4. Make refinements.



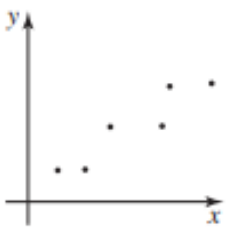
The second attempt is an improvement, but the line is too close to the points above it.
Improve the position of the line until a better 'balance' between upper and lower points is achieved.



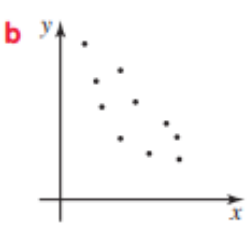
Exercise Set 1

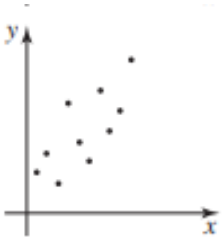
Fit a straight line to the data in the scatterplots using the equal-number-of-points method.

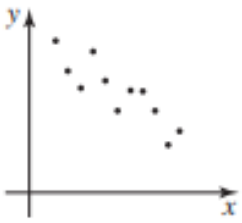
a)



b)

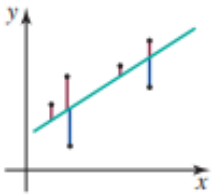


c) 

d) 

Fitting a straight line — least-squares regression

This is used when the data shows a linear relationship and there are no obvious outliers. (or these are discarded). To understand the underlying theory behind least-squares, consider the regression line shown below.



We wish to minimise the total of the vertical lines, or 'errors' in some way. For example, balancing the errors above and below the line. This is reasonable, but for sophisticated mathematical reasons it is preferable to minimise the sum of the squares of each of these errors. This is the essential mathematics of least-squares regression.

The formula for a straight line when using the x and y axis is $y = mx + c$

Where m = gradient of the line $(\frac{rise}{run})$
and c = y intercept ie the value of y when $x = 0$

Fortunately we can use an on-line calculator to find the equation of the line. This calculator is found at

https://www.socscistatistics.com/tests/regression/Default.aspx

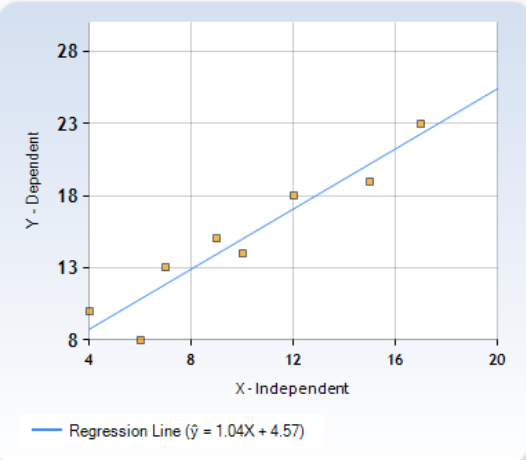
Example

x	4	6	7	9	10	12	15	17
y	10	8	13	15	14	18	19	23

Using the calculator above gives the equation of the line of best fit as

$y = 1.04x + 4.57$

The calculator also shows the scatterplot and shows the line of best fit. It would be better if it had started the graph at (0, 0) as then we could see that the y intercept was indeed about 4.5.



Exercise Set 2

Find the equation of the linear regression line for the following data set using the least-squares method utilizing the on-line calculator.

a)

x	4	6	7	9	10	12	15	17
y	10	8	13	15	14	18	19	23

b)

x	1	2	3	4	5	6	7	8	9
y	35	28	22	16	19	14	9	7	2

Interpretation, interpolation and extrapolation

Once you have a linear regression line, the slope and intercept can give important information about the data set. The slope (m) indicates the rate at which the data are increasing or decreasing. The y-intercept indicates the approximate value of the data when x = 0.

The line can be used to make predictions ie to find a value for y for a given value of x – where x is not in the given data.

The line is always of the form $y = (\text{gradient}) \times x + (y\text{-intercept})$

$y = \quad m \quad x + \quad c$

This line can be used to 'predict' data values for a given value of x. Of course, these are only approximations, since the regression line itself is only an estimate of the 'true' relationship between the bivariate data. However,

they can still be used, in some cases, to provide additional information about the data set (that is, make predictions).

Interpolation

Interpolation is the use of the regression line to predict values in between two values already in the data set. If the data are highly linear (r near +1 or -1) then we can be confident that our interpolated value is quite accurate. If the data are not highly linear (r near 0) then our confidence is duly reduced.

Extrapolation

Extrapolation is the use of the regression line to predict values smaller than the smallest value already in the data set or larger than the largest value.

Two problems may arise in attempting to extrapolate from a data set. Firstly, it may not be reasonable to extrapolate too far away from the given data values. For example, suppose there is a weather data set for 5 days. Even if it is highly linear (r near +1 or -1) a regression line used to predict the same data 15 days in the future is highly risky. Weather has a habit of randomly fluctuating and patterns rarely stay stable for very long.

Secondly, the data may be highly linear in a narrow band of the given data set. For example, there may be data on stopping distances for a train at speeds of between 30 and 60 km/h. Even if they are highly linear in this range, it is unlikely that things are similar at very low speeds (0–15 km/h) or high speeds (over 100 km/h).

Generally, one should feel more confident about the accuracy of a prediction derived from interpolation than one derived from extrapolation. Of course, it still depends upon the correlation coefficient (r). The closer to linearity the data are, the more confident our predictions in all cases.

Example

Age (years)	1	3	5	7	9	11
Height (cm)	60	76	115	126	145	148

Using the on-line calculator gives the equation $y = 9.4x + 55.3$

The gradient is 5.4 thus the average growth is 9.4 cm per year. b = 55.3 thus at birth ie $x = 0$ the height of the baby is 55.3 cm.

- Interpolation – find the height of an 8 year old child.
  $y = mx + c$
  $= 9.4(8) + 55.3$
  $= 129.5$ cm

- Extrapolation – predict the height of the child at 15
  $y = mx + c$
  $= 9.4(15) + 55.3$
  $= 194.08$ cm

This prediction is clearly unreliable as it assumes a constant rate of growth throughout life. Thus, even though r = 0.96 in this case it is only useful in the range the data provides.

Exercise Set 3

Q1. A drug company wishes to test the effectiveness of a drug to increase red blood cell counts in people who have a low count. The following data are collected.

Day of experiment	4	5	6	7	8	9
Red blood cell count	210	240	230	260	260	290

a) What are the independent and dependent variables in this case?

b) Find the relationship in the form $y = mx + c$ using the on-line calculator.

c) What is the rate at which the red blood cell count is changing?

d) What was the blood cell count at the beginning of the experiment (ie on day 0)

Q2. A wildlife exhibition is held over 6 weekends and features still and live displays. The number of live animals that are being exhibited varies each weekend. The number of animals participating, together with the number of visitors to the exhibition each weekend, is shown below.

Number of animals	6	4	8	5	7	6
Number of visitors	311	220	413	280	379	334

Find

a) the rate of increase of visitors as the number of live animals is increased

b) the predicted number of visitors if there are no live animals.

Q3. An electrical goods warehouse produces the following data showing the selling price of electrical goods to retailers and the volume of those sales.

Selling price ($)	60	80	100	120	140	160	200	220	240	260
Sales volume (× 1000)	400	300	275	250	210	190	150	100	50	0

Perform a least-squares regression analysis and discuss the meaning of the gradient and y-intercept.

Q4. The following table represents the costs for shipping a consignment of shoes from Melbourne factories. The cost is given in terms of distance from Melbourne. There are two factories that can be used. The data are summarised below.

Distance from Melbourne (km)	10	20	30	40	50	60	70	80
Factory 1 cost ($)	70	70	90	100	110	120	150	180
Factory 2 cost ($)	70	75	80	100	100	115	125	135

a) Find the least-squares regression equation for each factory.

b) Which factory is likely to have the lowest cost to ship to a shop in Melbourne? (use numbers)

c) Which factory is likely to have the lowest cost to ship to Mytown, 115 kilometres from Melbourne? (use maths)

Q5. A factory produces calculators. The least-squares regression line for cost of production (C) as a function of numbers of calculators (n) produced is given by:

$$C = 7.76n + 660$$

Furthermore, this function is deemed accurate when producing between 100 and 1000 calculators.

a) Find the cost to produce 200 calculators.

b) How many calculators can be produced for $2000?

c) Find the cost to produce 10 000 calculators.

d) What are the 'fixed' costs for this production?

e) Which of a) to c) above is an interpolation?