Lesson 2

A-level Applied Mathematics
Statistics : Correlation I : Year 1

2.1 The Regression Line Of *y* On *x*

In pure mathematics, straight lines are presented as being of the form

$$y = mx + c$$

where m is the gradient, the amount of slope
and c is where the line passes through the y-axis, the y-axis intercept

In statistics we write the (least squares) regression line of y on x in the form

$$y = a + bx$$

where b is the gradient, the amount of slope
and a is where the line passes through the y-axis, the y-axis intercept

In this equation y is the dependent variable
and x is the independent variable

The regression line should only be used to make predictions about y (the dependent variable) from a value of x, (the independent variable) and then only if x is within the range of "known independent data".

The value of the gradient, b tells you the change in y for each unit change in x
If the data is positively correlated, b will be positive.
If the data is negatively correlated, b will be negative.

Example #1

The mass, m kg and gestation period, t weeks for a random sample of eighty newborn babies at a hospital were recorded.
State which is the dependent variable and give a reason for you answer.

Example #2

The temperature, t °C and the number of ice creams, N, sold by a street trader on eight weekends one summer were recorded.
Strong correlation was noticed and the regression line of N on t was then found to be

$$N = 43 + 27t$$

(i) State which is the independent variable and give a reason for your answer.

(ii) Interpret the meaning of the figure 27 in the regression line's equation.

2.2 Exercise

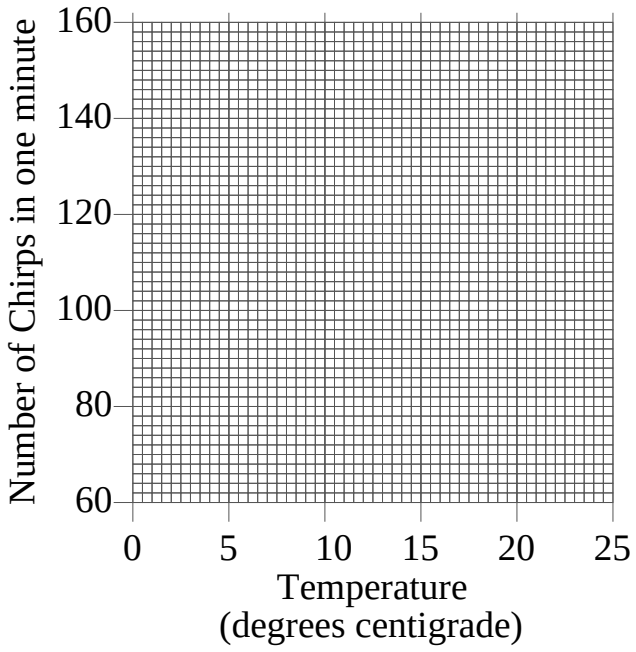
Question 1

A biologist is researching correlation between temperature and cricket chirps.
At 3 pm on eight randomly selected sunny, windfree afternoons, the biologist records the number of cricket chirps in one minute, N, and the air temperature, t °C

Temperature, t °C	20	15	19	25	13	21	23	18
Number of Chirps in one minute	110	80	90	150	70	120	140	90

(i) Explain why temperature is the independent variable

(ii) Plot a scatter graph to represent the data presented in the table above



(iii) Rebecca works out that the regression line of N on t is

$$N = -28.7 + 7t$$

Walter works out that the regression line of N on t is

$$N = 240 - 7t$$

One of the two suggested regression lines is correct.
State, with a reason, which is the correct equation for the regression line.

(iv) Add the correct regression line to the scatter graph.

(v) State the value of the gradient of the regression line, and give an interpretation of what this value means in terms of temperature and the number of chirps per minute.

(vi) A friend of the biologist wonders at what temperature the chirps stop. Give **TWO** reasons why this temperature can not be predicted from the biologist's research.

Question 2

A field was divided into 12 plots of equal area.
Each plot was fertilised with a different amount of fertilizer, f, litres.
The yield of grain, g kg, was measured for each plot.

Strong positive correlation is observed.

The regression line of g on f is

$$g = 2.4 + 3.6f$$

(i) State, with a reason, which is the dependent variable.

(ii) Interpret the meaning of the number 2.4 in the equation for the regression line.

(iii) Interpret the meaning of the number 3.6 in the equation for the regression line.

Question 3

The National Health service has investigated the correlation between the response time, t minutes, that it takes for an ambulance to get to the victims of heart-attacks and the survival rate, s %, of the victims.
Strong correlation is observed.
The regression line of s on t is

$$s = 92 - 0.8t$$

(i) Explain why the survival rate is the dependent variable.

(ii) Is the "strong correlation" positive or negative ?

(iii) The official target response time for a life threatening situation is 7 minutes.
What is a heart-attack victim's survival rate if the ambulance arrives exactly on the target time ?

(iv) If the victim is in hospital when the heart-attack occurred, what is the victims survival rate ?
Explain any assumptions made.

(v) The 90^{th} centile response time for a life threatening situation is 13 minutes, 15 seconds.
What is a 90^{th} centile victim's survival rate ?

(vi) Explain why the regression line should not be used to predict the survival rate for a victim who has to wait an hour for an ambulance.

(vii) Interpret the meaning of the number -0.8 in the regression line equation.

Question 4

A software company is trying to debug an Artificial Intelligence (AI) product intended to make deductions about the causes of allergies and poisonings.

It issues the following two bullet points followed by a conclusion;

- Today's tomatoes are not poisonous.
- 99 % of people who ate tomatoes before 1910 are dead

$\therefore$ Tomatoes grown in the early 20th century were poisonous.

What has the AI software not been made aware of ?
In your answer use the word ***correlation*** and the word ***causation***.

Question 5

Energy consumption is claimed to be a good predictor of Gross National Product. Gross National Product, GNP, is the total value of goods produced and services provided by a country during one year. An economist recorded the energy consumption, x, and the Gross national product, y, for eight countries.
The data is shown in the table;

Energy consumption, x	3.4	7.7	12.0	75	58	67	113	131
Gross National Product, y	55	240	390	1100	1390	1330	1400	1900

The equation of the regression line of y on x is

$$y = 225 + 12.9x$$

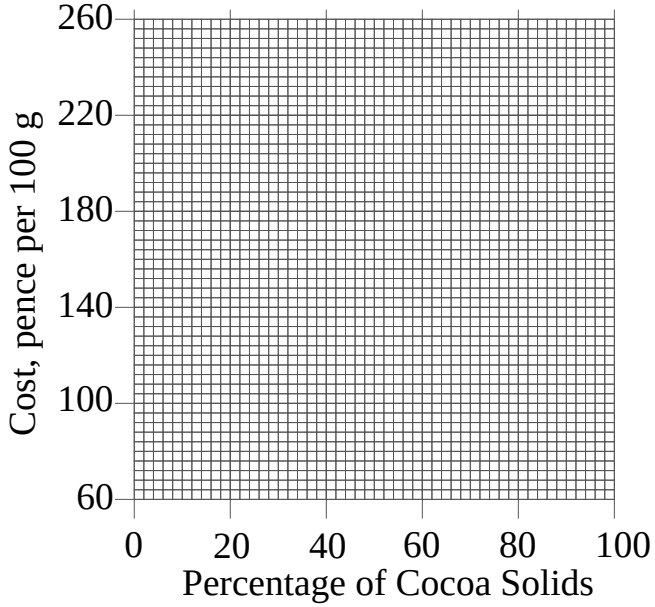
The economist uses this regression equation to estimate the energy consumption of a country with a Gross National Product of 3500.
Give two reasons why this may not be a valid estimate.

Question 6

A student is investigating the relationship between the price, y pence of 100 g of chocolate, and the percentage, x %, of cocoa solids in the chocolate.
The data obtained is shown in the table;

Brand	A	B	C	D	E	F	G	H
Solids, x %	10	20	30	35	40	50	60	70
Cost, y pence per 100 g	70	110	80	200	120	180	220	260

(i) Draw a scatter graph to represent this data.



(ii) The equation of the regression line of y on x is

$$y = 34 + 3.1x$$

Draw the regression line on your diagram.

(iii) Explain what the number 34 in the regression line's equation represents.

The student believes that one brand of chocolate is overpriced and uses the regression line to suggest a fair price for this brand.

(iv) Suggest, with a reason, which brand is overpriced.

(v) What will be the student's suggested fair price ?

This document is a part of a **Mathematics Community Outreach Project** initiated by Shrewsbury School
It may be freely duplicated and distributed, unaltered, for non-profit educational use
In October 2020, Shrewsbury School was voted "**Independent School of the Year 2020**"
© 2025 Number Wonder
Teachers may obtain detailed worked solutions to the exercises by email from MHHShrewsbury@Gmail.com