Stats1 Chapter 4 :: Correlation

jfrost@tiffin.kingston.sch.uk

www.drfrostmaths.com

@DrFrostMaths

Last modified: 3rd March 2018

Experimental

i.e. Dealing with collected data.

Chp1: Data Collection

Methods of sampling, types of data, and populations vs samples.

Chp2: Measures of Location/Spread

Statistics used to summarise data, including mean, standard deviation, quartiles, percentiles. Use of linear interpolation for estimating medians/quartiles.

Chp3: Representation of Data

Producing and interpreting visual representations of data, including box plots and histograms.

Chp4: Correlation

Measuring how related two variables are, and using linear regression to predict values.

Theoretical

Deal with probabilities and modelling to make inferences about what we 'expect' to see or make predictions, often using this to reason about/contrast with experimentally collected data.

Chp5: Probability

Venn Diagrams, mutually exclusive + independent events, tree diagrams.

Chp6: Statistical Distributions

Common distributions used to easily find probabilities under certain modelling conditions, e.g. binomial distribution.

Chp7: Hypothesis Testing

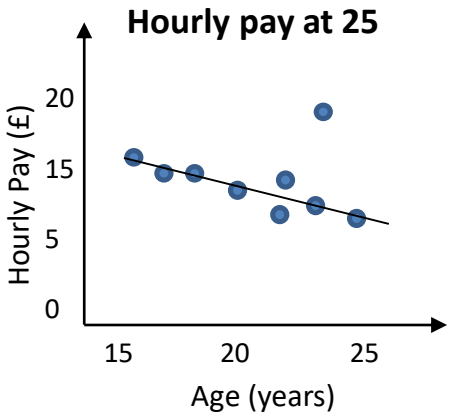
Determining how likely observed data would have happened 'by chance', and making subsequent deductions.

This Chapter Overview

Previously we have only considered one variable at a time. When we introduce a second variable (e.g. height with age), **we might want to consider the relationship between them**.

This is a short chapter!

"Describe the type of correlation."



"The daily mean windspeed, w knots, and the daily maximum gust, g knots, were recorded. The equation of the regression line of g on w for these 15 days is $g = 7.23 + 1.82w$.

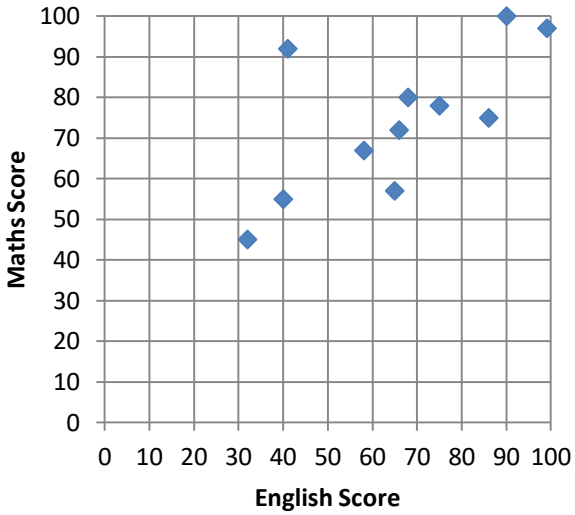
(a) Given an interpretation of the value of the gradient of this regression line.

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

Changes since the old 'S1' syllabus:
This chapter has been scaled back significantly since the S1 'Correlation' and 'Regression' chapters. You no longer need to determine the equation of the line of best fit (the regression line), or calculate measures of correlation, but merely have to interpret an equation already given or the limitations of estimates made or comment on the suitability of a linear regression model.

Recap of correlation

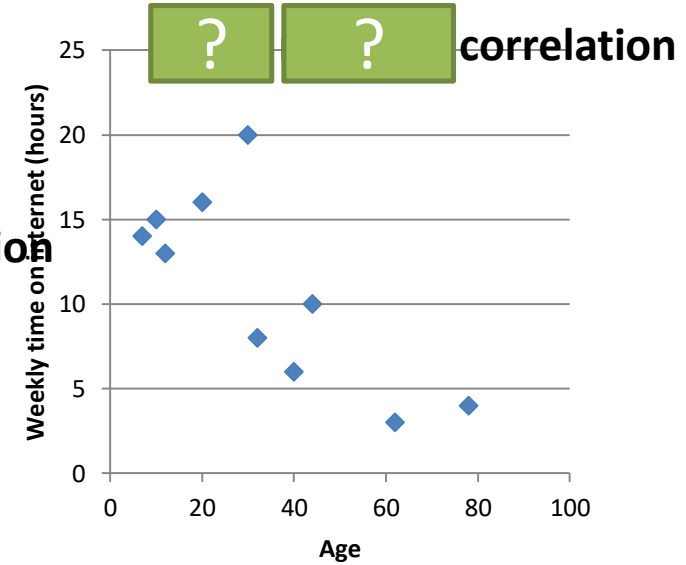
Correlation gives the **strength of the relationship** (and the type of relationship) between two variables. Data with two variables is known as **bivariate data**.

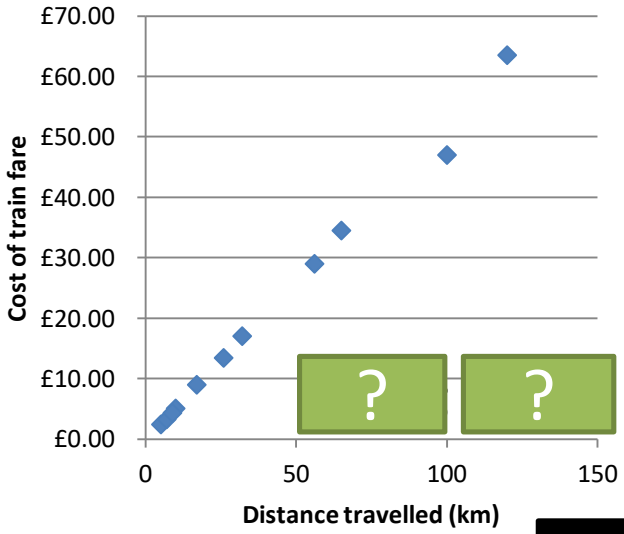


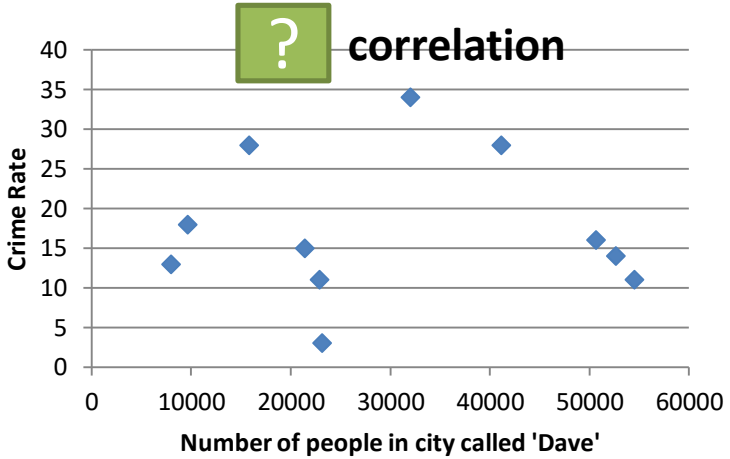
Type of correlation:

? ? correlation

strength type





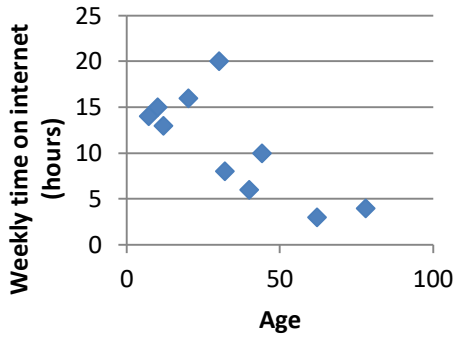


The vertical-axis variable usually **depends** on the horizontal-axis value. For this reason distance would be the **independent/explanatory variable** and cost the **dependent/response variable**.

Important correlation concepts

Important Point 1

To **interpret** the correlation between two variables is to give a worded description in the context of the problem.



a) State the correlation shown.
b) Describe/interpret the relationship between age and weekly time on the internet.

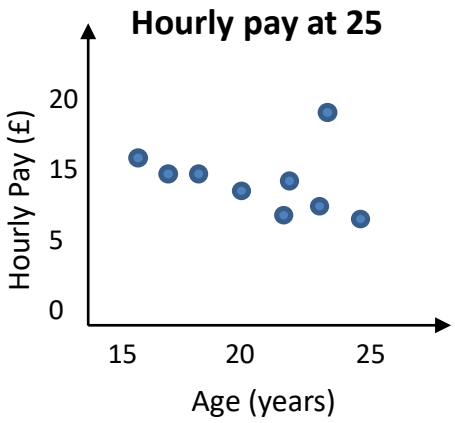
a) ?
b) ?

Important Point 2

[Textbook] Two variables have a **causal relationship** if a change in one variable directly causes a change in the other. Just because two variables show correlation it does not necessarily mean that they have a causal relationship.

Hideko was interested to see if there was a relationship between what people earn and the age which they left education or training. She says her data supports the conclusion that more education causes people to earn a lower hourly rate of pay. Give one reason why Hideko's conclusion might not be valid.

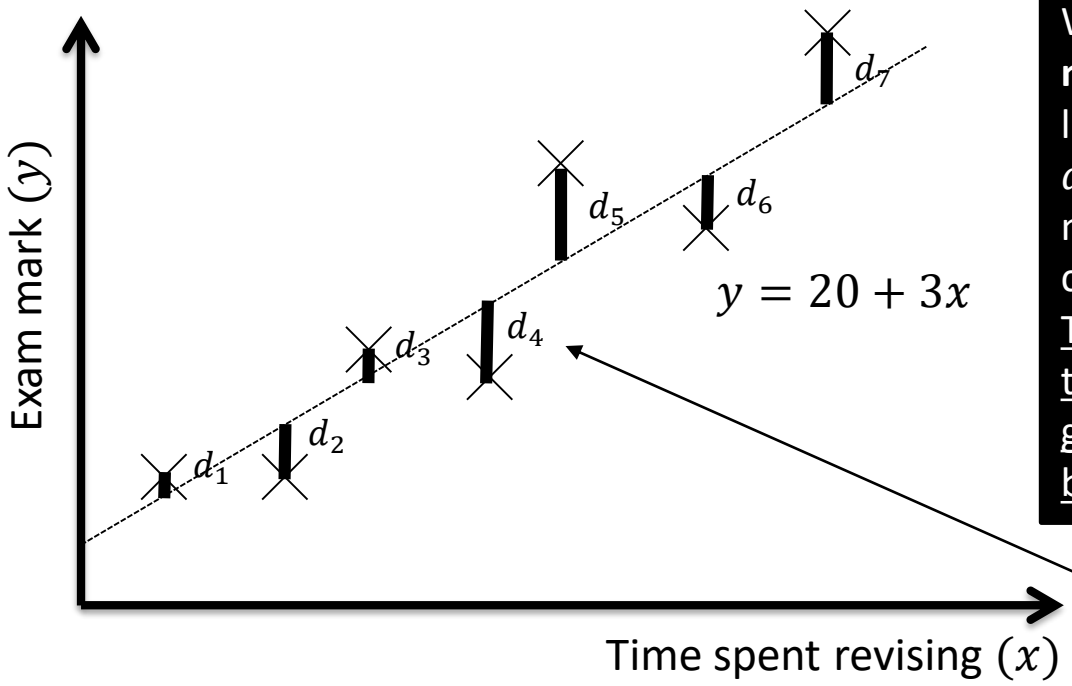
?



Exercise 4A

Pearson Statistics/Mechanics Year 1/AS
Pages 61-62

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.

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.

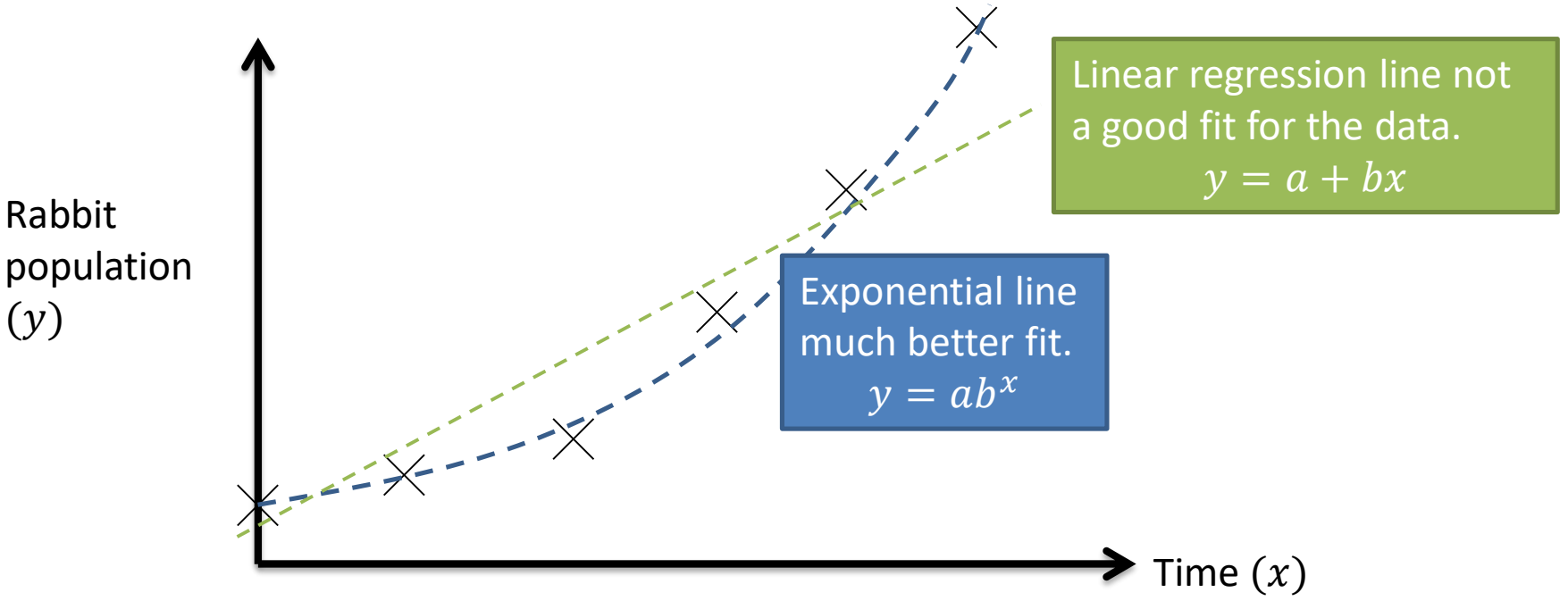
$$d_1^2 + d_2^2 + \cdots = \Sigma d_i^2$$

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.

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?

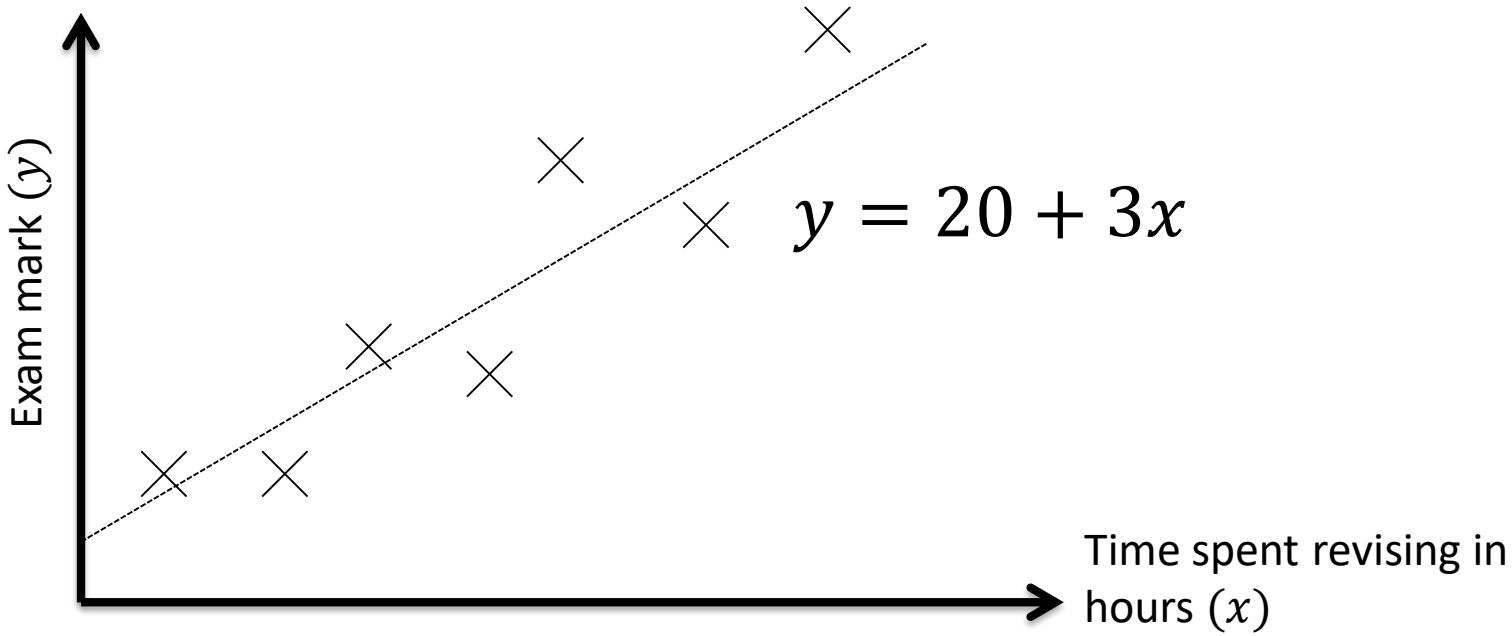
What is regression?



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?

?

How do we interpret the y-intercept of 20?

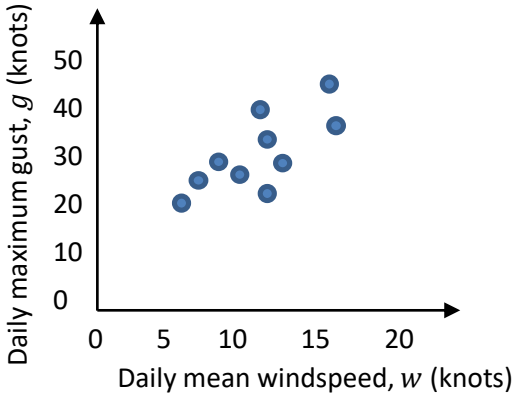
?

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.

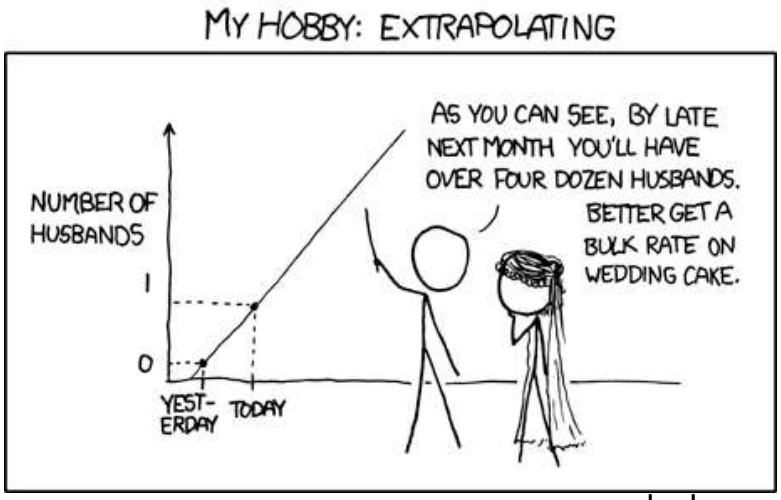
a ?

b ?

c ?

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

Interpolating and Extrapolating

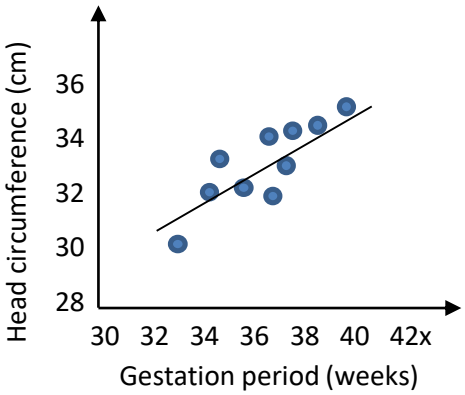


xkcd.com

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 ?

Exercise 4B

Pearson Statistics/Mechanics Year 1/AS
Pages 65-66

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

"Use of Technology" Monkey says:
Dr Frost's 'Interactive Guide to the Classwiz' has a guide to determine the equation of the regression line (even though you are not required to do so for the exam). Try verifying the regression line of g on w for the data above has equation $g = 7.23 + 1.82w$.
http://www.drfrostmaths.com/resources/resource.php?rid=262