Regression, Correlation and Hypothesis Tests

1:: Exponential Models

Recap of Pure Year 1. Using $y = ab^x$ to model an exponential relationship between two variables.

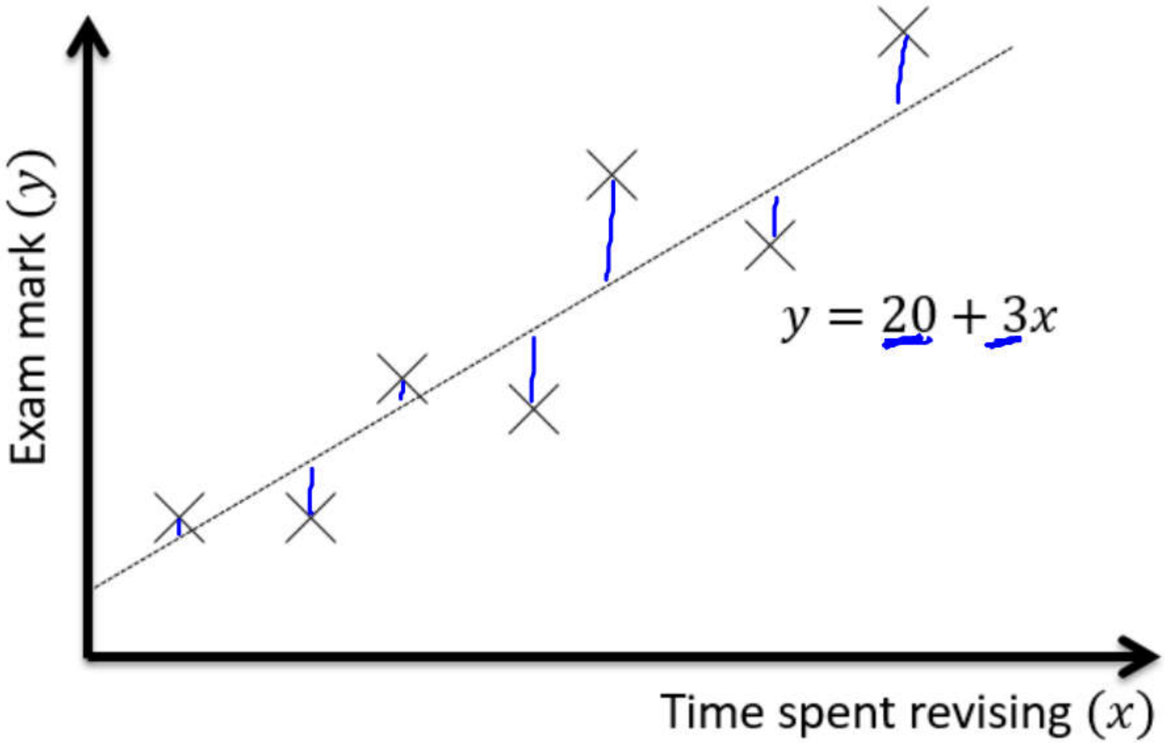
2:: Measuring Correlation

Using the Product Moment Correlation Coefficient (PMCC), r, to measure the strength of correlation between two variables.

3:: Hypothesis Testing for no correlation

We want to test whether two variables have some kind of correlation, or whether any correlation observed just happened by chance.

What is regression?



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?

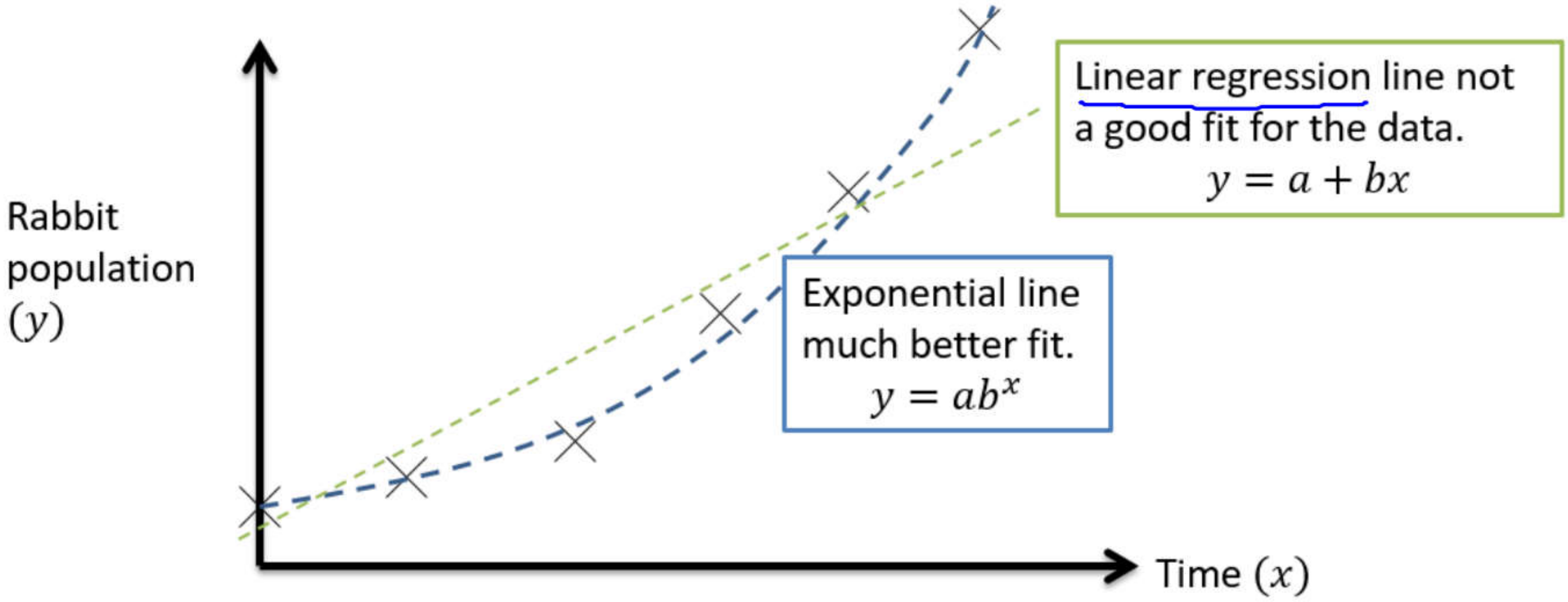
What we've done here is come up with a **model** to explain the data, in this case, a line $y = a + bx$. We've then tried to set a and b such that the resulting y value matches the actual exam marks as closely as possible.

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.

Note from Year 1

Extrapolation: making predictions outside the original data range
Extrapolation is unreliable as the trend may not continue outside the given range.

Exponential Regression



For some variables, e.g. population with time, it may be more appropriate to use an **exponential** equation, i.e. $y = ab^x$, where a and b are constants we need to fix to best match the data.

$$y = ab^x$$

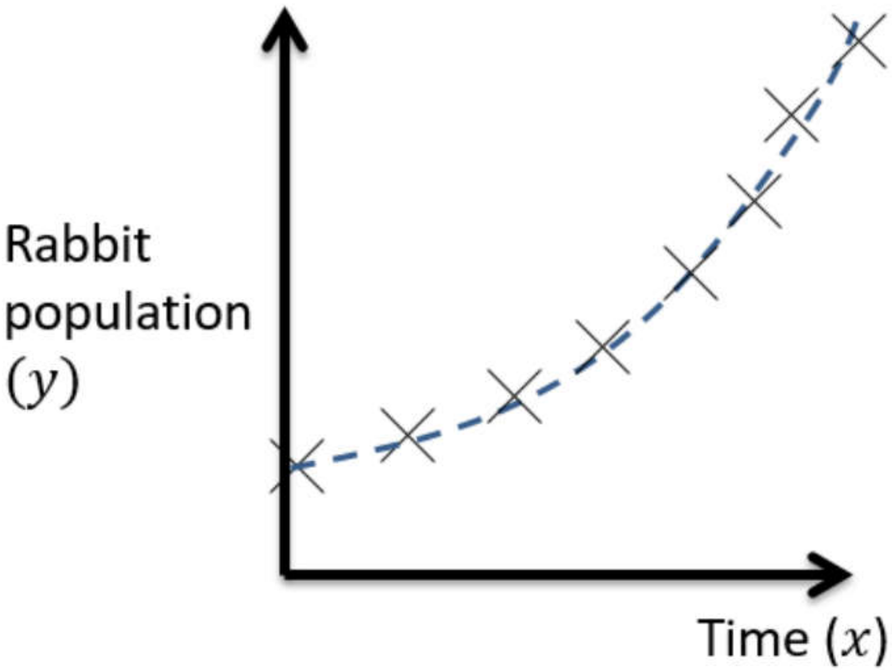
$$\log y = \log ab^x$$

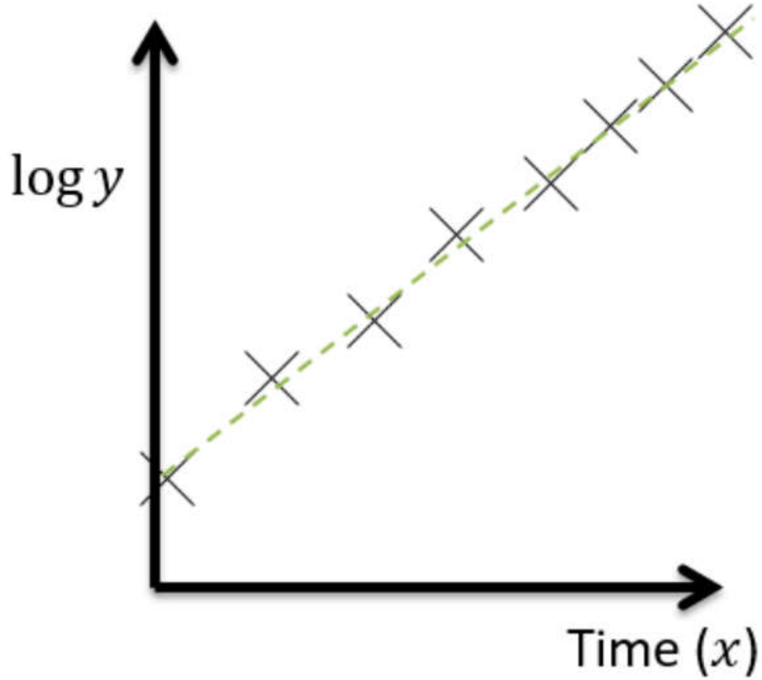
$$\log y = \log a + \log b^x$$

$$\log y = \log a + x \log b$$

✎ If $y = ab^x$ for constants a and b then $\log y = \log a + x \log b$

$$\boxed{\log y = \underbrace{\log a}_{c} + x\,\underbrace{\log b}_{m}}$$





Comparing the equations, we can see that if we log the y values (although leave the x values), the data then forms a straight line, with y-intercept $\log a$ and gradient $\log b$.

The table shows some data collected on the temperature, in °C, of a colony of bacteria (t) and its growth rate (g).

Temperature, $\boldsymbol{t}$ (°C)	3	5	6	8	9	11
Growth rate, g	1.04	1.49	1.79	2.58	3.1	4.46

The data are coded using the changes of variable $x = t$ and $y = \log g$. The regression line of y on x is found to be $y = -0.2215 + 0.0792x$. $y(x)$

a. Mika says that the constant -0.2215 in the regression line means that the colony is shrinking when the temperature is 0°C. Explain why Mika is wrong

b. Given that the data can be modelled by an equation of the form $g = kb^t$ where k and b are constants, find the values of k and b. $g(t)$ ↳ g is the subject.

a) $\boxed{y = -0.2215 + 0.0792x}$ (linear)

$t = 0 = x$

$y = -0.2215$

$\log g = -0.2215$

$g = 10^{-0.2215}$

$g = 0.60$ (2dp) > 0 so there is a growing colony at $t=0$

b) $\log g = -0.2215 + 0.0792t$

$g = 10^{-0.2215+0.0792t}$

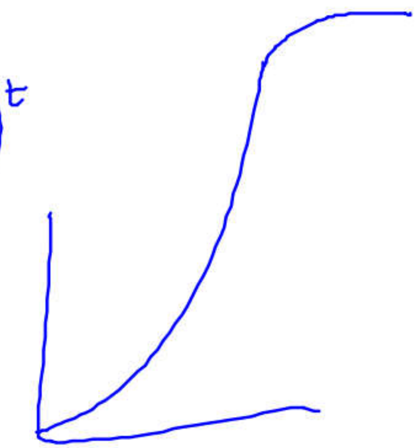
$g = 10^{-0.2215} \times 10^{0.0792t} \rightarrow (10^{0.0792})^t$

$g = 10$

$\boxed{g = 0.6 \times 1.2^t}$

1.15^t

$k = 0.6 \quad b = 1.2$



Robert wants to model a rabbit population P with respect to time in years t. He proposes that the population can be modelled using an exponential model: $P = kb^t$ The data is coded using $x = t$ and $y = \log P$. The regression line of y on x is found to be $y = 2 + 0.3x$. Determine the values of k and b.

$y = 2 + 0.3x$

$\log P = 2 + 0.3t$

$P = 10^{2+0.3t}$

$P = 10^2 \times 10^{0.3t}$

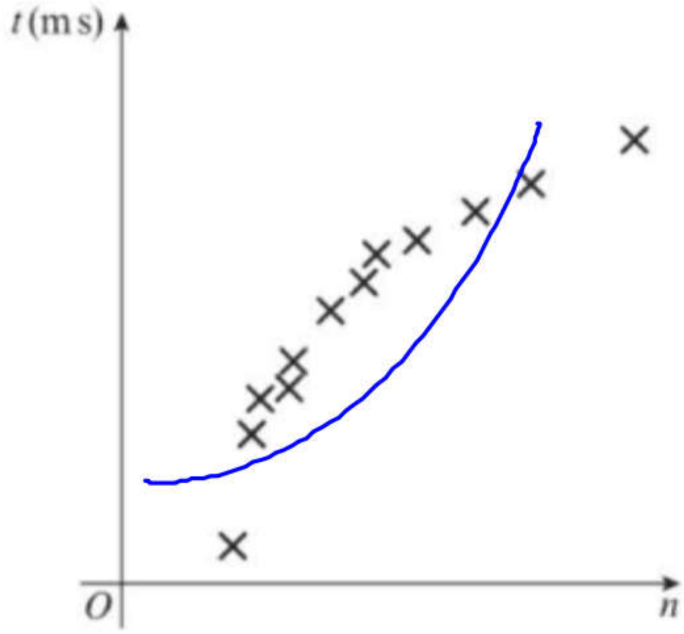
$P = 100 \times 1.995^t$

$k = 100 \quad b = 1.995$

Ex 1A

$\boxed{\text{Q5, 6}}$

5 The time, t ms, needed for a computer algorithm to determine whether a number, n, is prime is recorded for different values of n. A scatter graph of t against n is drawn.



a Explain why a model of the form $t = a + bn$ is unlikely to fit these data.

linear

The data are coded using the changes of variable $y = \log t$ and $x = \log n$. The regression line of y on x is found to be $y = -0.301 + 0.6x$.

b Find an equation for t in terms of n, giving your answer in the form $t = an^k$, where a and k are constants to be found.

$y = -0.301 + 0.6x$

$\log t = -0.301 + 0.6 \log n$

$\log t = -0.301 + \log n^{0.6}$

$\log t - \log n^{0.6} = -0.301$

$\log \frac{t}{n^{0.6}} = -0.301$

$\frac{t}{n^{0.6}} = 10^{-0.301}$

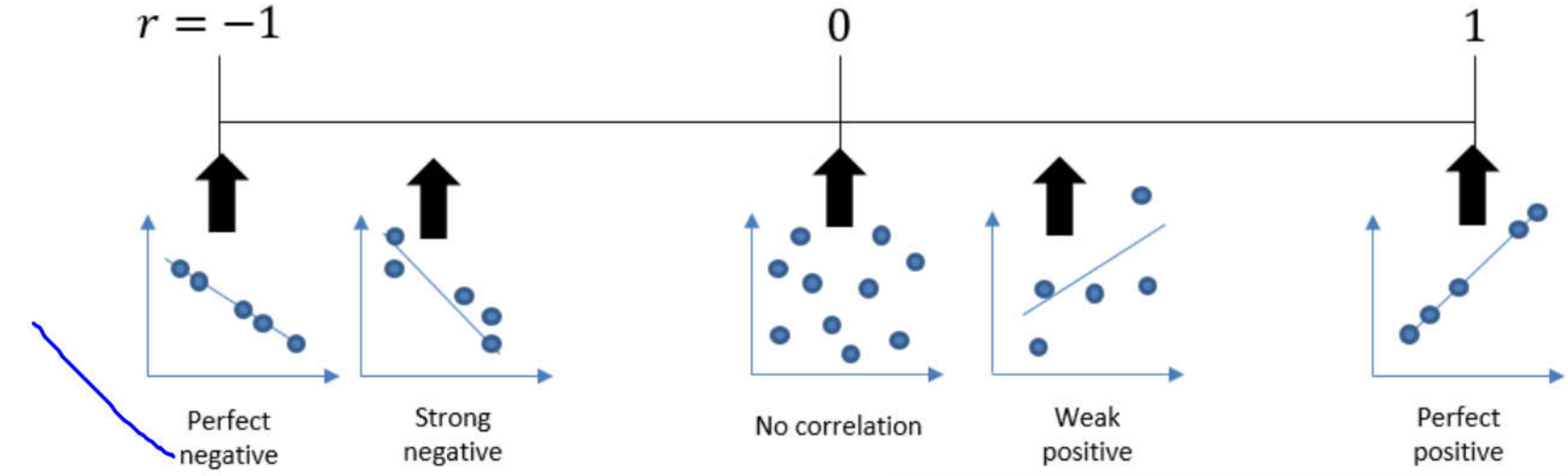
$t = 0.5n^{0.6}$

Measuring Correlation

You're used to using qualitative terms such as "positive correlation" and "negative correlation" and "no correlation" to describe the **type** of correlation, and terms such as "perfect", "strong" and "weak" to describe the **strength**.
The **Product Moment Correlation Coefficient** is one way to quantify this:

PMCC

✎ The product moment correlation coefficient (PMCC), denoted by r, describes the linear correlation between two variables. It can take values between -1 and 1.



Rule of thumb: $r < -0.7$ or $r > 0.7$ is considered to be 'strong' correlation.

Note that PMCC is only applicable for a <u>linear</u> correlation, i.e. closeness of fit to a linear regression line (i.e. a <u>straight</u> 'line of best fit'). It may be the data exhibits strong correlation with respect to a different model (e.g. exponential) even when the PMCC is low.

Calculating r on your calculator

x	y
1	3
2	6
3	5
4	8

6: Statistics

The following instructions are for the Casio Class Wiz.
Press MODE then select 'Statistics'.

$y = a + bx$

We want to measure **linear** correlation, so select $y = a + bx$

Data Entry

Enter each of the x values in the table on the left, press = after each input. Use the arrow keys to get to the top of the y column.

PMCC

While entering data, press OPTN then choose "Regression Calc" to obtain r (i.e. the coefficients of your line of best fit and the PMCC). a and b would give you the y-intercept and gradient of the regression line (but not required in this chapter).

Pressing AC allows you to construct a statistical calculation yourself. In OPTN, there is an additional 'Regression' menu allowing you to insert r into your calculation.

You should obtain $r = 0.868$

From the large data set, the daily mean windspeed, w knots, and the daily maximum gust, g knots, were recorded for the first 10 days in September in Hurn in 1987.

Day of month	1	2	3	4	5	6	7	8	9	10	
w	4	4	8	7	12	12	3	~~4~~	~~7~~	10	15
g	13	12	19	23	33	37	10	~~n/a~~	~~n/a~~	23	?

a. State the meaning of n/a in the table above.

b. Calculate the product moment correlation coefficient for the remaining 8 days.

c. With reference to your answer to part b, comment on the suitability of a linear regression model for these data.

a) No available data.

b) $r = \underline{\underline{0.9533}}$ (4sf)

c) Because r is close to 1, correlation is very strong, so points lie close to a straight line – hence linear model is suitable.

Ex 1B

Q 2, 3, 4