# 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.

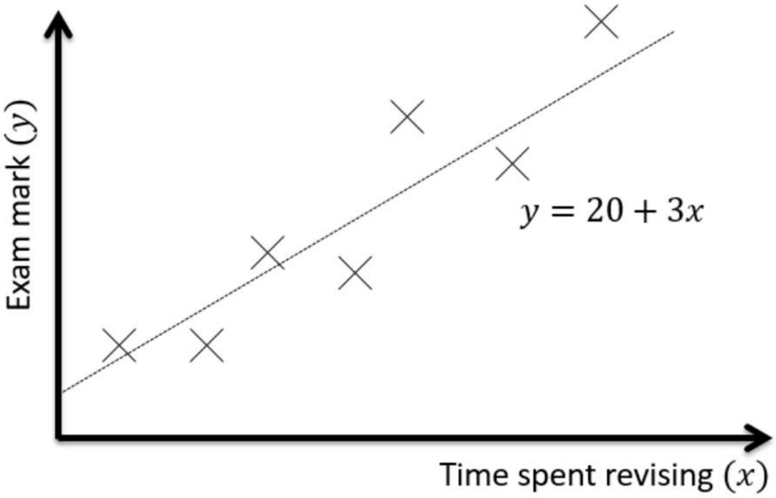
**2**:: Measuring Correlation

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?

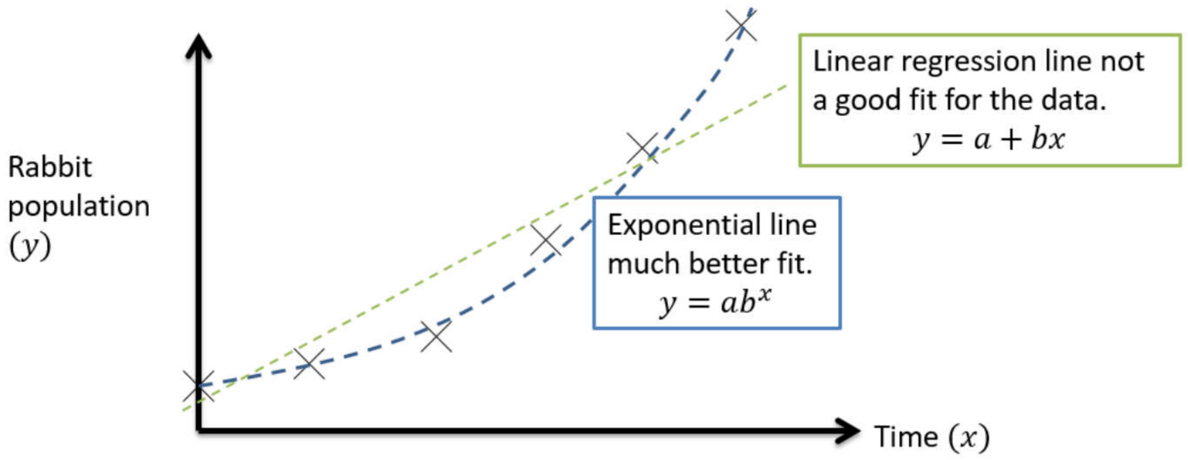
What we've done here is come up with a **model** to explain the data, in this case, a line $y = a + bx$. We've then tried to set $a$ and $b$ such that the resulting $y$ value matches the actual exam marks as closely as possible.

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



Linear regression line not a good fit for the data. $y = a + bx$

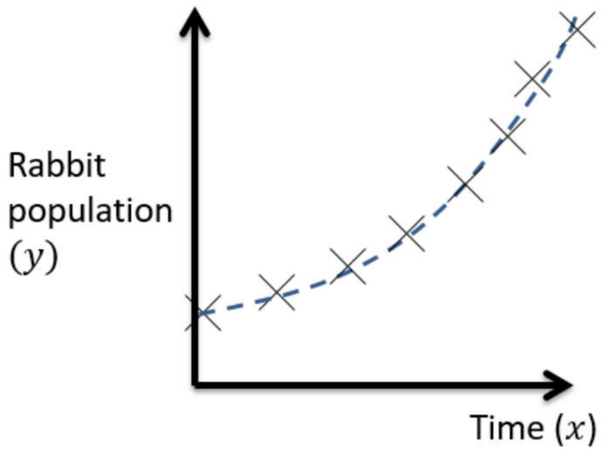
Exponential line much better fit. $y = ab^x$

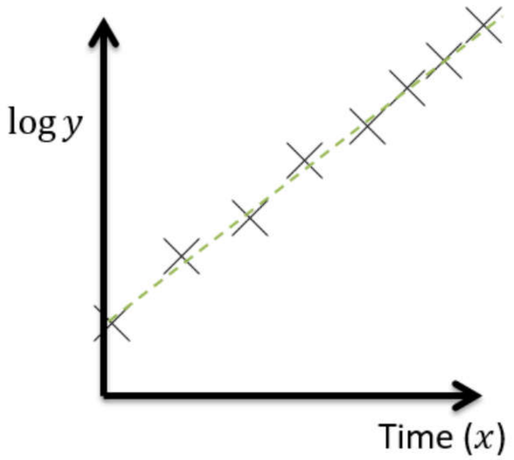
For some variables, e.g. population with time, it may be more appropriate to use an **exponential** equation, i.e. $y = ab^x$, where $a$ and $b$ are constants we need to fix to best match the data.

$$y = ab^x$$

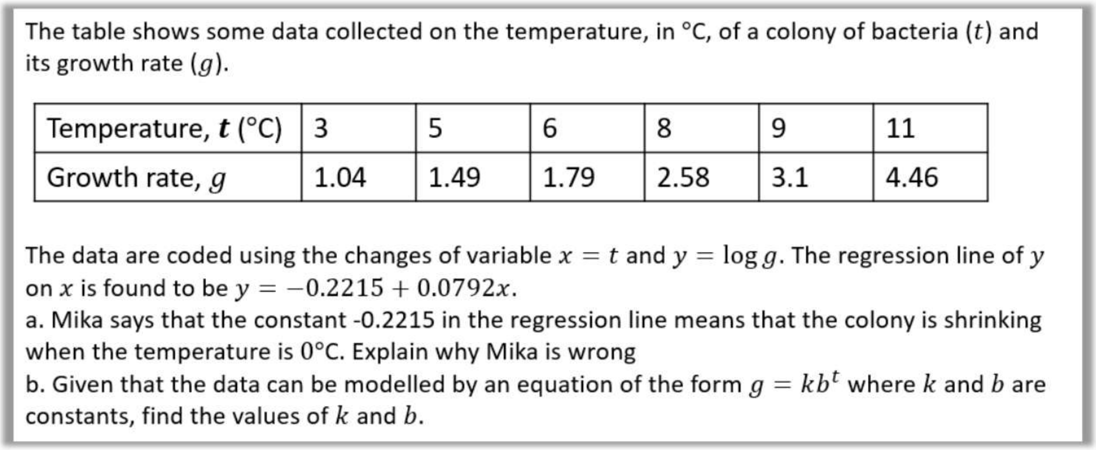
✎ If $y = ab^x$ for constants $a$ and $b$ then $\log y = \log a + x \log b$

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





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 $\log b$.



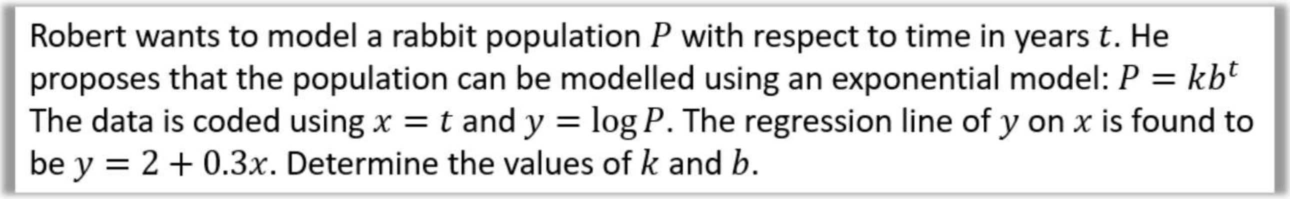
The table shows some data collected on the temperature, in °C, of a colony of bacteria ($t$) and its growth rate ($g$).

| Temperature, ***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 $y = -0.2215 + 0.0792x$.

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 $b$.



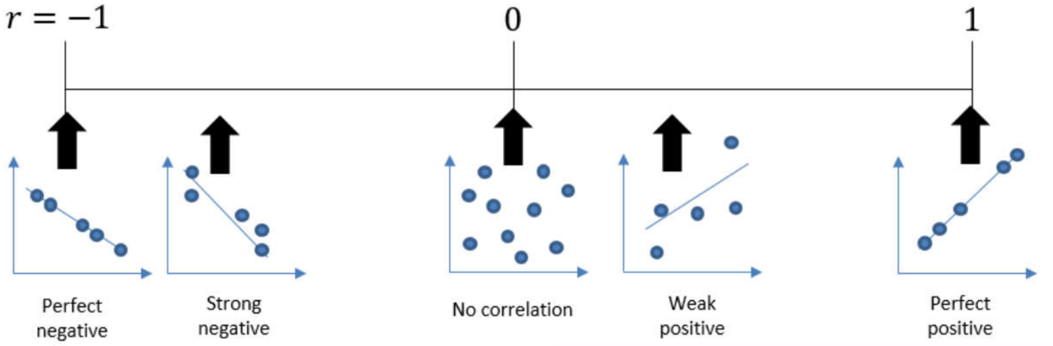
Robert wants to model a rabbit population $P$ with respect to time in years $t$. He proposes that the population can be modelled using an exponential model: $P = kb^t$ The data is coded using $x = t$ and $y = \log P$. The regression line of $y$ on $x$ is found to be $y = 2 + 0.3x$. Determine the values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

# 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:

> 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 linear correlation, i.e. closeness of fit to a linear regression line (i.e. a straight '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

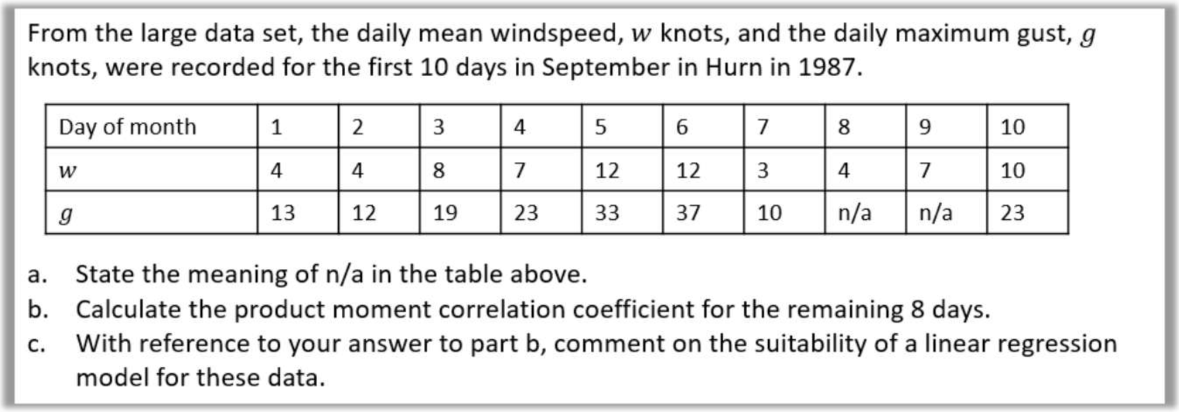
Enter each of the $x$ values in the table on the left, press = after each input. Use the arrow keys to get to the top of the $y$ column.

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 |
| $g$ | 13 | 12 | 19 | 23 | 33 | 37 | 10 | n/a | n/a | 23 |

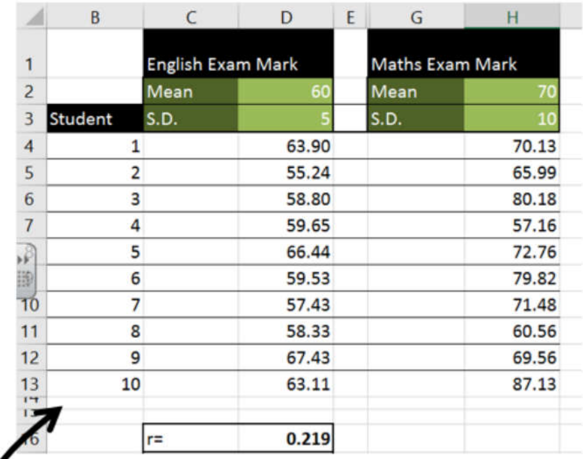
a. State the meaning of n/a in the table above.

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

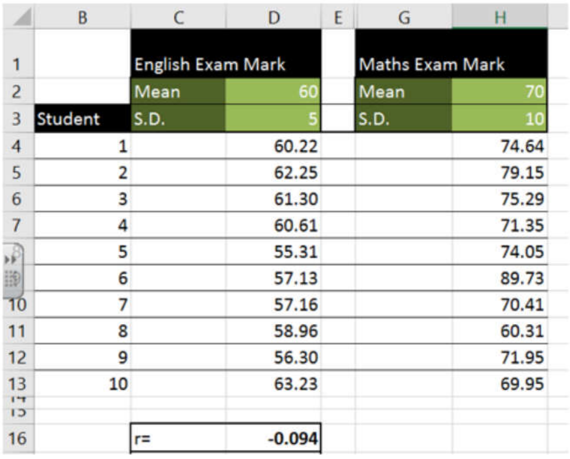
c. With reference to your answer to part b, comment on the suitability of a linear regression model for these data.

Ex 1B

# Hypothesis Testing for Correlation



| | B | C | D | E | G | H |
|---|---|---|---|---|---|---|
| 1 | | English Exam Mark | | | Maths Exam Mark | |
| 2 | | Mean | 60 | | Mean | 70 |
| 3 | Student | S.D. | 5 | | S.D. | 10 |
| 4 | 1 | | 63.90 | | | 70.13 |
| 5 | 2 | | 55.24 | | | 65.99 |
| 6 | 3 | | 58.80 | | | 80.18 |
| 7 | 4 | | 59.65 | | | 57.16 |
| | 5 | | 66.44 | | | 72.76 |
| | 6 | | 59.53 | | | 79.82 |
| 10 | 7 | | 57.43 | | | 71.48 |
| 11 | 8 | | 58.33 | | | 60.56 |
| 12 | 9 | | 67.43 | | | 69.56 |
| 13 | 10 | | 63.11 | | | 87.13 |
| 16 | | r= | 0.219 | | | |



| | B | C | D | E | G | H |
|---|---|---|---|---|---|---|
| 1 | | English Exam Mark | | | Maths Exam Mark | |
| 2 | | Mean | 60 | | Mean | 70 |
| 3 | Student | S.D. | 5 | | S.D. | 10 |
| 4 | 1 | | 60.22 | | | 74.64 |
| 5 | 2 | | 62.25 | | | 79.15 |
| 6 | 3 | | 61.30 | | | 75.29 |
| 7 | 4 | | 60.61 | | | 71.35 |
| | 5 | | 55.31 | | | 74.05 |
| | 6 | | 57.13 | | | 89.73 |
| 10 | 7 | | 57.16 | | | 70.41 |
| 11 | 8 | | 58.96 | | | 60.31 |
| 12 | 9 | | 56.30 | | | 71.95 |
| 13 | 10 | | 63.23 | | | 69.95 |
| 16 | | r= | -0.094 | | | |

Suppose we use a spreadsheet to **randomly** generate maths marks for students, and separately generate **random** English marks.

The **observed** PMCC between Maths and English marks in this first set of data is **0.219**

But the true PMCC between Maths and English is 0.

**This is because they were generated independently of each other and so have no correlation. The observed PMCC may vary from the true PMCC because the data is randomly sampled, just as if we threw a fair die, we wouldn't necessarily see equal counts of each outcome.**

✎ $r$ denotes the PMCC of a **sample**.

✎ $\rho$ (Greek letter *rho*) is the PMCC for the **whole population**.

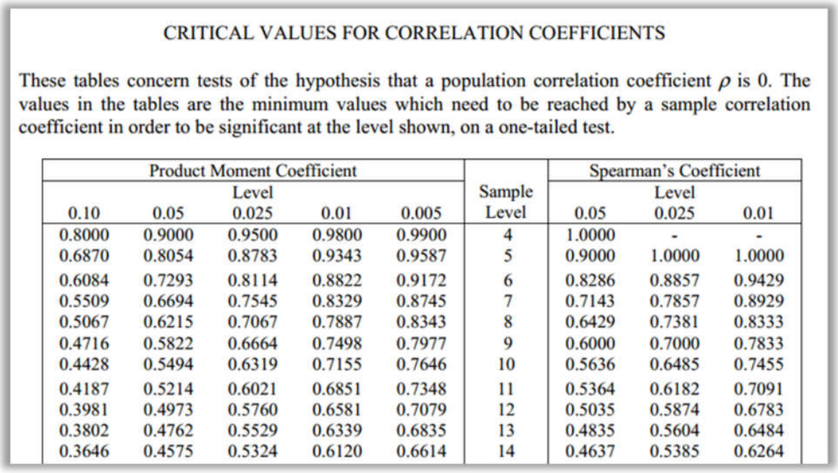
✎ $\therefore$ $r$ is the test statistic, $\rho$ is the population parameter.

Let’s carry out a hypothesis test on whether there is **positive** correlation between English and Maths marks, at 10% significance level:

$H_0$: $\rho = 0$

$H_1$:

Sample size =



CRITICAL VALUES FOR CORRELATION COEFFICIENTS

These tables concern tests of the hypothesis that a population correlation coefficient $\rho$ is 0. The values in the tables are the minimum values which need to be reached by a sample correlation coefficient in order to be significant at the level shown, on a one-tailed test.

| Product Moment Coefficient | | | | | Sample | Spearman’s Coefficient | | |
|---|---|---|---|---|---|---|---|---|
| Level | | | | | | Level | | |
| 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | Level | 0.05 | 0.025 | 0.01 |
| 0.8000 | 0.9000 | 0.9500 | 0.9800 | 0.9900 | 4 | 1.0000 | - | - |
| 0.6870 | 0.8054 | 0.8783 | 0.9343 | 0.9587 | 5 | 0.9000 | 1.0000 | 1.0000 |
| 0.6084 | 0.7293 | 0.8114 | 0.8822 | 0.9172 | 6 | 0.8286 | 0.8857 | 0.9429 |
| 0.5509 | 0.6694 | 0.7545 | 0.8329 | 0.8745 | 7 | 0.7143 | 0.7857 | 0.8929 |
| 0.5067 | 0.6215 | 0.7067 | 0.7887 | 0.8343 | 8 | 0.6429 | 0.7381 | 0.8333 |
| 0.4716 | 0.5822 | 0.6664 | 0.7498 | 0.7977 | 9 | 0.6000 | 0.7000 | 0.7833 |
| 0.4428 | 0.5494 | 0.6319 | 0.7155 | 0.7646 | 10 | 0.5636 | 0.6485 | 0.7455 |
| 0.4187 | 0.5214 | 0.6021 | 0.6851 | 0.7348 | 11 | 0.5364 | 0.6182 | 0.7091 |
| 0.3981 | 0.4973 | 0.5760 | 0.6581 | 0.7079 | 12 | 0.5035 | 0.5874 | 0.6783 |
| 0.3802 | 0.4762 | 0.5529 | 0.6339 | 0.6835 | 13 | 0.4835 | 0.5604 | 0.6484 |
| 0.3646 | 0.4575 | 0.5324 | 0.6120 | 0.6614 | 14 | 0.4637 | 0.5385 | 0.6264 |

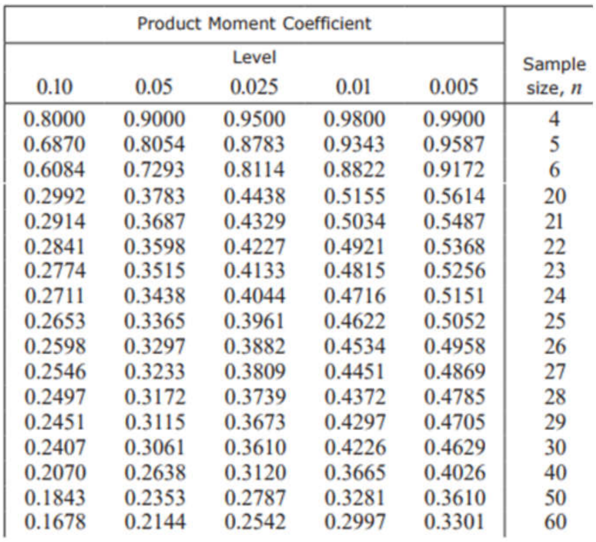
**Note**: you take the negative value from the table if looking at significance for negative correlation

# Two-tailed test

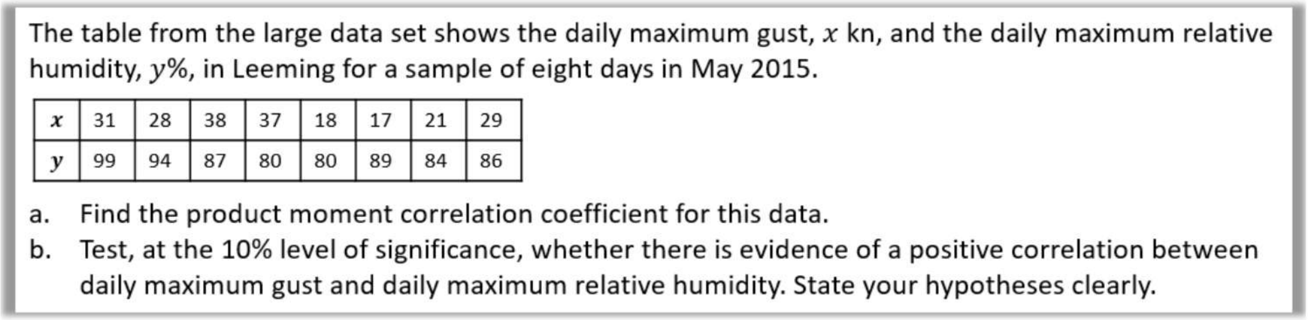
In the previous example we hypothesised that English/Maths marks were positively correlated. But we could also test whether there was **any** correlation, i.e. positive **or** negative.

A scientist takes 30 observations of the masses of two reactants in an experiment. She calculates a product moment correlation coefficient of $r = -0.45$.

The scientist believes there is no correlation between the masses of the two reactants. Test at the 10% level of significance, the scientist’s claim, stating your hypotheses clearly.



| Product Moment Coefficient | | | | | Sample size, $n$ |
|---|---|---|---|---|---|
| Level | | | | | |
| 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | |
| 0.8000 | 0.9000 | 0.9500 | 0.9800 | 0.9900 | 4 |
| 0.6870 | 0.8054 | 0.8783 | 0.9343 | 0.9587 | 5 |
| 0.6084 | 0.7293 | 0.8114 | 0.8822 | 0.9172 | 6 |
| 0.2992 | 0.3783 | 0.4438 | 0.5155 | 0.5614 | 20 |
| 0.2914 | 0.3687 | 0.4329 | 0.5034 | 0.5487 | 21 |
| 0.2841 | 0.3598 | 0.4227 | 0.4921 | 0.5368 | 22 |
| 0.2774 | 0.3515 | 0.4133 | 0.4815 | 0.5256 | 23 |
| 0.2711 | 0.3438 | 0.4044 | 0.4716 | 0.5151 | 24 |
| 0.2653 | 0.3365 | 0.3961 | 0.4622 | 0.5052 | 25 |
| 0.2598 | 0.3297 | 0.3882 | 0.4534 | 0.4958 | 26 |
| 0.2546 | 0.3233 | 0.3809 | 0.4451 | 0.4869 | 27 |
| 0.2497 | 0.3172 | 0.3739 | 0.4372 | 0.4785 | 28 |
| 0.2451 | 0.3115 | 0.3673 | 0.4297 | 0.4705 | 29 |
| 0.2407 | 0.3061 | 0.3610 | 0.4226 | 0.4629 | 30 |
| 0.2070 | 0.2638 | 0.3120 | 0.3665 | 0.4026 | 40 |
| 0.1843 | 0.2353 | 0.2787 | 0.3281 | 0.3610 | 50 |
| 0.1678 | 0.2144 | 0.2542 | 0.2997 | 0.3301 | 60 |



The table from the large data set shows the daily maximum gust, $x$ kn, and the daily maximum relative humidity, $y$%, in Leeming for a sample of eight days in May 2015.

| $x$ | 31 | 28 | 38 | 37 | 18 | 17 | 21 | 29 |
|---|---|---|---|---|---|---|---|---|
| $y$ | 99 | 94 | 87 | 80 | 80 | 89 | 84 | 86 |

a. Find the product moment correlation coefficient for this data.

b. Test, at the 10% level of significance, whether there is evidence of a positive correlation between daily maximum gust and daily maximum relative humidity. State your hypotheses clearly.

Ex 1C

**2.** A meteorologist believes that there is a relationship between the daily mean windspeed, $w$ kn, and the daily mean temperature, $t$ °C. A random sample of 9 consecutive days is taken from past records from a town in the UK in July and the relevant data is given in the table below.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $t$ | 13.3 | 16.2 | 15.7 | 16.6 | 16.3 | 16.4 | 19.3 | 17.1 | 13.2 |
| $w$ | 7 | 11 | 8 | 11 | 13 | 8 | 15 | 10 | 11 |

The meteorologist calculated the product moment correlation coefficient for the 9 days and obtained $r = 0.609$

(a) Explain why a linear regression model based on these data is unreliable on a day when the mean temperature is 24 °C

(1)

(b) State what is measured by the product moment correlation coefficient.

(1)

(c) Stating your hypotheses clearly test, at the 5% significance level, whether or not the product moment correlation coefficient for the population is greater than zero.

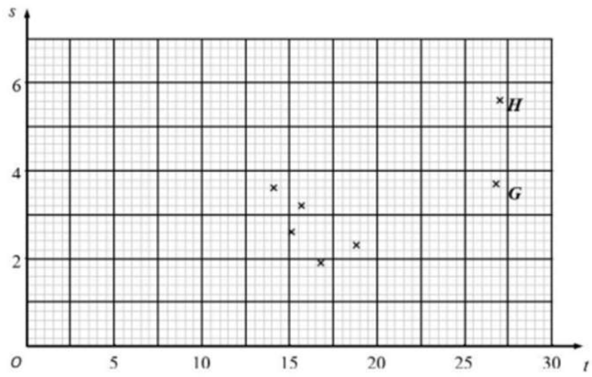
(3)

Using the same 9 days a location from the large data set gave $\bar{t} = 27.2$ and $\bar{w} = 3.5$

(d) Using your knowledge of the large data set, suggest, giving your reason, the location that gave rise to these statistics.

(1)

**2.** A researcher believes that there is a linear relationship between daily mean temperature and daily total rainfall. The 7 places in the northern hemisphere from the large data set are used. The mean of the daily mean temperatures, $t$ °C, and the mean of the daily total rainfall, $s$ mm, for the month of July in 2015 are shown on the scatter diagram below.



(a) With reference to the scatter diagram, explain why a linear regression model may not be suitable for the relationship between $t$ and $s$.
(1)

The researcher calculated the product moment correlation coefficient for the 7 places and obtained $r = 0.658$

(b) Stating your hypotheses clearly, test at the 10% level of significance, whether or not the product moment correlation coefficient for the population is greater than zero.
(3)

(c) Using your knowledge of the large data set, suggest the names of the 2 places labelled $G$ and $H$.
(1)

(d) Using your knowledge from the large data set, and with reference to the locations of the 2 places labelled G and H, give a reason why these places have the highest temperatures in July.
(1)

(e) Suggest how you could make better use of the large data set to investigate the relationship between daily mean temperature and daily total rainfall.
(1)

**2.** An ornithologist believes that there is a relationship between the tail length, $t$ mm, and the wing length, $w$ mm, of female hook-billed kites. A random sample of size 10 is taken from a database of these kites and the relevant data is given in the table below.

| $t$ **(mm)** | 191 | 197 | 208 | 180 | 188 | 210 | 196 | 191 | 179 | 208 |
|---|---|---|---|---|---|---|---|---|---|---|
| $w$ **(mm)** | 284 | 285 | 288 | 273 | 280 | 283 | 288 | 271 | 257 | 289 |

The ornithologist plans to use a linear regression model based on these data and interpolate or extrapolate as necessary to estimate the wing length of other female hook-billed kites from their tail length.

(a) (i) Explain what is meant by extrapolation.

**(1)**

(ii) Explain the dangers of extrapolation.

**(1)**

The ornithologist attempts to calculate the product moment correlation coefficient, $r$, and obtains a value of 1.3

(b) Explain how she should be able to identify that this is incorrect without carrying out any further calculations.

**(1)**

(c) Use your calculator to find the correct value of the product moment correlation coefficient, $r$.

**(1)**

(d) Stating your hypotheses clearly test, at the 1% significance level, whether or not there is evidence that the product moment correlation coefficient for the population is positive.

**(3)**

(e) Explain what your test in part (d) suggests about female hook-billed kites.

**(1)**

**2.** Tessa owns a small clothes shop in a seaside town. She records the weekly sales figures, £$w$, and the average weekly temperature, $t$°C, for 8 weeks during the summer.
The product moment correlation coefficient for these data is $-0.915$

(a) Stating your hypotheses clearly and using a 5% level of significance, test whether or not the correlation between sales figures and average weekly temperature is negative.
**(3)**

(b) Suggest a possible reason for this correlation.
**(1)**

Tessa suggests that a linear regression model could be used to model these data.

(c) State, giving a reason, whether or not the correlation coefficient is consistent with Tessa's suggestion.
**(1)**

(d) State, giving a reason, which variable would be the explanatory variable.
**(1)**

Tessa calculated the linear regression equation as $w = 10\ 755 - 171t$

(e) Give an interpretation of the gradient of this regression equation.
**(1)**

## Critical Values for Correlation Coefficients

These tables concern tests of the hypothesis that a population correlation coefficient $\rho$ is 0. The values in the tables are the minimum values which need to be reached by a sample correlation coefficient in order to be significant at the level shown, on a one-tailed test.

| Product Moment Coefficient | | | | | | Spearman's Coefficient | | |
|---|---|---|---|---|---|---|---|---|
| | | Level | | | Sample size, $n$ | | Level | |
| 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | | 0.05 | 0.025 | 0.01 |
| 0.8000 | 0.9000 | 0.9500 | 0.9800 | 0.9900 | 4 | 1.0000 | – | – |
| 0.6870 | 0.8054 | 0.8783 | 0.9343 | 0.9587 | 5 | 0.9000 | 1.0000 | 1.0000 |
| 0.6084 | 0.7293 | 0.8114 | 0.8822 | 0.9172 | 6 | 0.8286 | 0.8857 | 0.9429 |
| 0.5509 | 0.6694 | 0.7545 | 0.8329 | 0.8745 | 7 | 0.7143 | 0.7857 | 0.8929 |
| 0.5067 | 0.6215 | 0.7067 | 0.7887 | 0.8343 | 8 | 0.6429 | 0.7381 | 0.8333 |
| 0.4716 | 0.5822 | 0.6664 | 0.7498 | 0.7977 | 9 | 0.6000 | 0.7000 | 0.7833 |
| 0.4428 | 0.5494 | 0.6319 | 0.7155 | 0.7646 | 10 | 0.5636 | 0.6485 | 0.7455 |
| 0.4187 | 0.5214 | 0.6021 | 0.6851 | 0.7348 | 11 | 0.5364 | 0.6182 | 0.7091 |
| 0.3981 | 0.4973 | 0.5760 | 0.6581 | 0.7079 | 12 | 0.5035 | 0.5874 | 0.6783 |
| 0.3802 | 0.4762 | 0.5529 | 0.6339 | 0.6835 | 13 | 0.4835 | 0.5604 | 0.6484 |
| 0.3646 | 0.4575 | 0.5324 | 0.6120 | 0.6614 | 14 | 0.4637 | 0.5385 | 0.6264 |
| 0.3507 | 0.4409 | 0.5140 | 0.5923 | 0.6411 | 15 | 0.4464 | 0.5214 | 0.6036 |
| 0.3383 | 0.4259 | 0.4973 | 0.5742 | 0.6226 | 16 | 0.4294 | 0.5029 | 0.5824 |
| 0.3271 | 0.4124 | 0.4821 | 0.5577 | 0.6055 | 17 | 0.4142 | 0.4877 | 0.5662 |
| 0.3170 | 0.4000 | 0.4683 | 0.5425 | 0.5897 | 18 | 0.4014 | 0.4716 | 0.5501 |
| 0.3077 | 0.3887 | 0.4555 | 0.5285 | 0.5751 | 19 | 0.3912 | 0.4596 | 0.5351 |
| 0.2992 | 0.3783 | 0.4438 | 0.5155 | 0.5614 | 20 | 0.3805 | 0.4466 | 0.5218 |
| 0.2914 | 0.3687 | 0.4329 | 0.5034 | 0.5487 | 21 | 0.3701 | 0.4364 | 0.5091 |
| 0.2841 | 0.3598 | 0.4227 | 0.4921 | 0.5368 | 22 | 0.3608 | 0.4252 | 0.4975 |
| 0.2774 | 0.3515 | 0.4133 | 0.4815 | 0.5256 | 23 | 0.3528 | 0.4160 | 0.4862 |
| 0.2711 | 0.3438 | 0.4044 | 0.4716 | 0.5151 | 24 | 0.3443 | 0.4070 | 0.4757 |
| 0.2653 | 0.3365 | 0.3961 | 0.4622 | 0.5052 | 25 | 0.3369 | 0.3977 | 0.4662 |
| 0.2598 | 0.3297 | 0.3882 | 0.4534 | 0.4958 | 26 | 0.3306 | 0.3901 | 0.4571 |
| 0.2546 | 0.3233 | 0.3809 | 0.4451 | 0.4869 | 27 | 0.3242 | 0.3828 | 0.4487 |
| 0.2497 | 0.3172 | 0.3739 | 0.4372 | 0.4785 | 28 | 0.3180 | 0.3755 | 0.4401 |
| 0.2451 | 0.3115 | 0.3673 | 0.4297 | 0.4705 | 29 | 0.3118 | 0.3685 | 0.4325 |
| 0.2407 | 0.3061 | 0.3610 | 0.4226 | 0.4629 | 30 | 0.3063 | 0.3624 | 0.4251 |
| 0.2070 | 0.2638 | 0.3120 | 0.3665 | 0.4026 | 40 | 0.2640 | 0.3128 | 0.3681 |
| 0.1843 | 0.2353 | 0.2787 | 0.3281 | 0.3610 | 50 | 0.2353 | 0.2791 | 0.3293 |
| 0.1678 | 0.2144 | 0.2542 | 0.2997 | 0.3301 | 60 | 0.2144 | 0.2545 | 0.3005 |
| 0.1550 | 0.1982 | 0.2352 | 0.2776 | 0.3060 | 70 | 0.1982 | 0.2354 | 0.2782 |
| 0.1448 | 0.1852 | 0.2199 | 0.2597 | 0.2864 | 80 | 0.1852 | 0.2201 | 0.2602 |
| 0.1364 | 0.1745 | 0.2072 | 0.2449 | 0.2702 | 90 | 0.1745 | 0.2074 | 0.2453 |
| 0.1292 | 0.1654 | 0.1966 | 0.2324 | 0.2565 | 100 | 0.1654 | 0.1967 | 0.2327 |