Chapter 1 - Regression, Correlation and HT

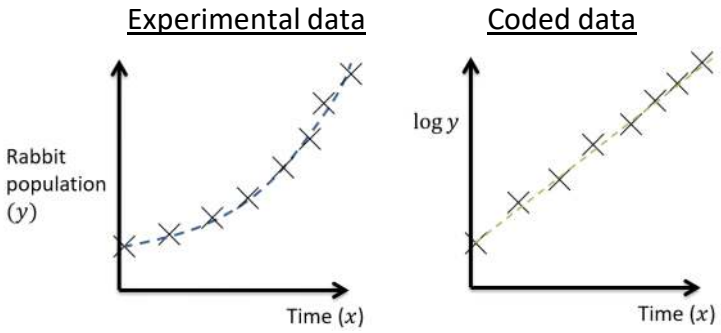
Personal notes:

1.1 - Exponential models

Notes

- A *regression line* is a line that best describes the behaviour of a set of data. In other words, line of best fit !

- Regression lines could be used to model a linear relationship between two variables.
- However, sometimes experimental data doesn't fit a linear model.
- To tackle this problem, one can use logarithms to code the data to obtain a linear relationship.



Example

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 $y = -0.2215 + 0.0792x$.

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.

1.1 - Exponential models

Practice

Mr. Fan wants to model a rabbit population R with respect to time in years t. He proposes that the population can be modelled using an exponential model: $R = kb^t$

The data is coded using $x = t$ and $y = \log R$. The regression line of y on x is found to be $y = 1 + 0.7x$.

Determine the values of k and b.

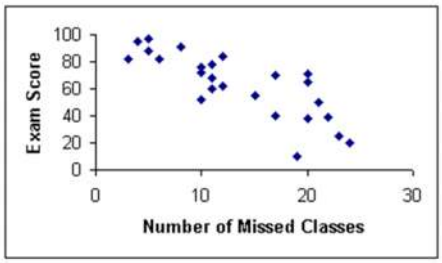
1.2 - Measuring correlation

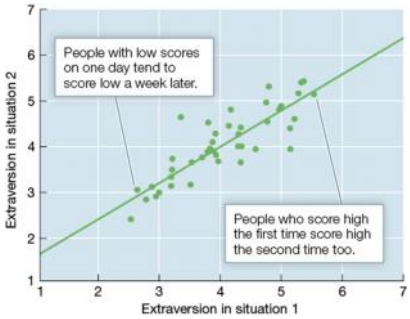
Notes

- We can quantify the strength of linear correlation between two variables. This value is known as ***product moment correlation coefficient (PMCC/Pearson's r-value)***
- The r-value is a value between -1 to +1.

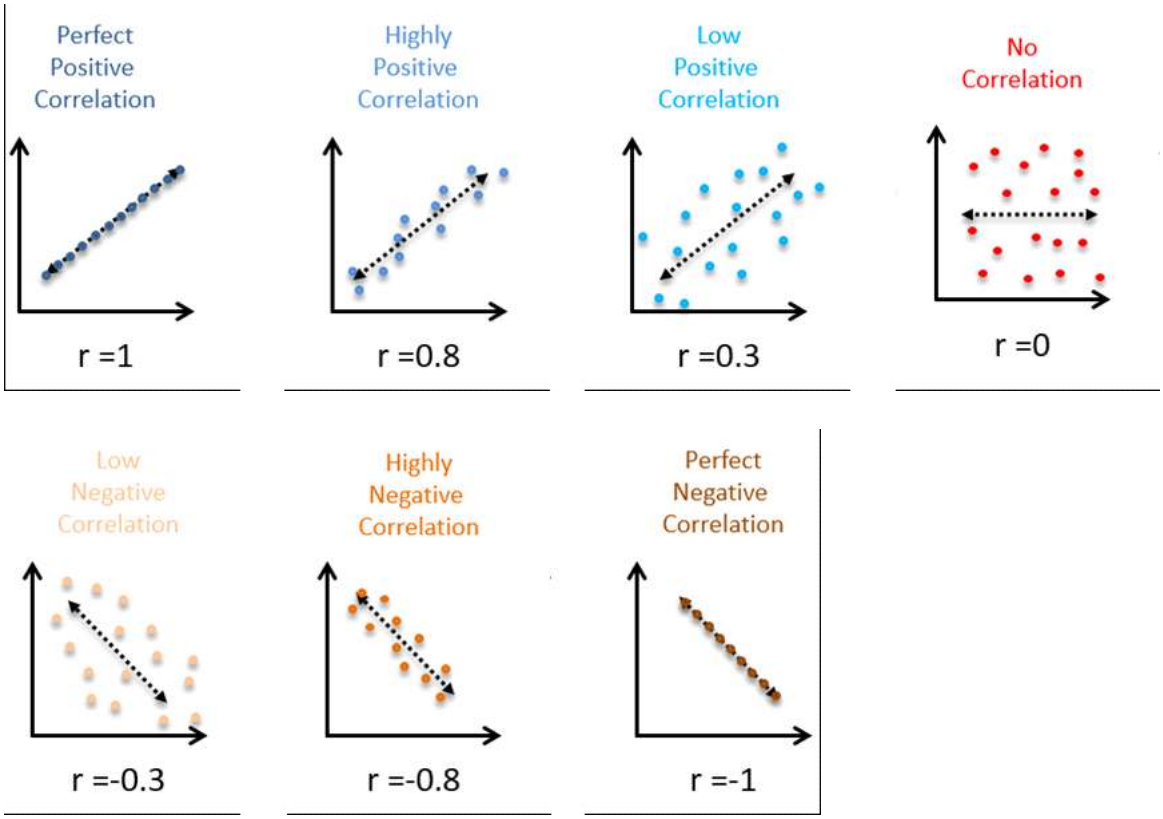
Types of correlation

- *Positive correlation* (r-value > 0): high scores on one variable associated with high scores on the other variable
- *Negative correlation* (r-value < 0): high scores on variable 1 associated with low scores on variable 2, and vice versa
- *Zero correlation* (r-value $= 0$): the two variables do not covary, there is no relationship.





Correlation coefficients spectrum



What do you notice about the position of data (points) in relation to the regression line?

1.2 - Measuring correlation

Example (r-value in calculator)

Work out the product moment correlation coefficient of the following sets of data

x	2	5	10	18	23
y	4	5	7	8	10

x	6	9	13	19	20
y	5	10	15	12	15

x	1	3	5	6	8
y	9	3	5	4	4

Example

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.

Practice

The following data is taken from the large data set showing the daily mean air temperature and the daily mean wind speed in Beijing in July 1987:

Daily mean air temperature (°C)	23.8	22.1	26.0	26.4	25.8	25.7	26.1	27.2
Daily mean wind speed (kn)	4.0	3.3	7.0	4.0	6.0	3.5	3.3	3.0

a) Calculate the product moment correlation coefficient.
b) Given your answer to part a), how suitable is a linear regression model for this data?

1.3 - Hypothesis testing for zero correlation

Theory

Correlation could be used as an *inferential test*, which allows us to infer something about the population from the sample data collected. We can test whether the correlation coefficient we found is different from zero. (Why zero? See below explanation)

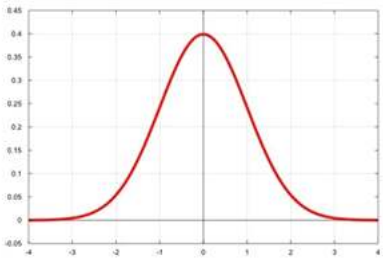
We may hypothesis that two variables are positively (or negatively) correlated, or maybe we predict a relationship but we are not sure whether it will be positive or negative (statistical relationship we are talking here *obviously*).

Null hypothesis is that there is no relationship.

Imagine if we keep drawing samples of, say, 50 pairs of **uncorrelated** scores (e.g. extroversion and IQ) from a population (e.g. the whole UK).
Then we calculate the Pearson's r value of each pair to find the correlation coefficient, we would expect

- Lots of pairs that give a r-value close to 0
- Some moderate positive and negative r-values
- Few very high positive and negative r-values

This results a reasonably *normally distributed* around a **mean of zero.**



The "tails" of the distribution would be where the high values of r would be. These high values of r are unlikely (given the pairs of scores are coming from **uncorrelated** variables).

So by chance, there could indeed be some high correlations (i.e. high positive/negative r, in right and left tail of normal distribution).

With Pearson's hypothesis testing, we are *asking how likely it is that we would obtain the r value we have in our sample*, given that the null hypothesis is true.

1.3 - Hypothesis testing for zero correlation

Notes

Use *one-tailed test* to test whether the population PMCC, p, is greater than zero or less than zero.

- $H_0: p = 0$, $H_1: p > 0$ or
- $H_0: p = 0$, $H_1: p < 0$

Use *two-tailed test* to test whether the population PMCC, p, is not equal to zero.

- $H_0: p = 0$, $H_1: p \neq 0$

We can determine the critical region for *r* for our hypothesis test by using the table of critical values in the formulae booklet.

Product moment coefficient					
		Level			Sample
0.10	0.05	0.025	0.01	0.005	size
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
0.5509	0.6694	0.7545	0.8329	0.8745	7
0.5067	0.6215	0.7967	0.7887	0.8343	8
0.4716	0.5822	0.6664	0.7498	0.7977	9

Example

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.

1.3 - Hypothesis testing for zero correlation

Practice

The table from the LDS shows the daily maximum gust, x kn, and the daily maximum relative humidity, $y\%$, in Camborne for a sample of ten days in June 2015.

x	y
42	98
37	99
22	97
26	99
22	99
21	90
22	88
29	85
30	86
24	74

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

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

Product Moment Coefficient					
		Level			Sample size, n
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
0.5509	0.6694	0.7545	0.8329	0.8745	7
0.5067	0.6215	0.7067	0.7887	0.8343	8
0.4716	0.5822	0.6664	0.7498	0.7977	9
0.4428	0.5494	0.6319	0.7155	0.7646	10

Exam Practice (Oct 2020 Q2)

Stav believes that there is a correlation between Daily Total Sunshine and Daily Maximum Relative Humidity at Heathrow.

He calculates the product moment correlation coefficient between these two variables for a random sample of 30 days and obtains $r = -0.377$

(c) Carry out a suitable test to investigate Stav's belief at a 5% level of significance. State clearly

- your hypotheses
- your critical value

(3)