Regression, correlation and hypothesis testing

Previously in Chapter 4 of Stats/Mech Year 1, you learnt how to interpret correlation and regression line equations for bivariate data. The methods you used were reliant on the two variables having a linear correlation. In this chapter we will look at how we can analyse data where a correlation exists between two variables but is not linear. We will also define the product moment correlation coefficient and explore its role in hypothesis testing for correlation.

Interpreting models of the forms $y = ax^n$ and $y = kb^x$

Linear models are very useful because they allow us to analyse data with relative ease. However, in reality not all models that display a pattern between two variables are linear. We will now look at one such model, of the form $y = ax^n$. You can use the coding $Y = logy$ and $X = logx$ to obtain a linear relationship:

- If you have a model of the form $y = ax^n$, then a linear relationship is given by $logy = loga + nlogx$.

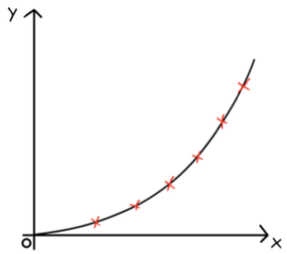
If we plot $logy$ against $logx$, we will obtain a linear model (straight line) since the above equation is in the form $Y = MX + C$. We will now look at how we can obtain the linear form using the original form:

We start with the original equation $y = ax^n$
Taking logs of both sides: $logy = \log(ax^n)$
Using the multiplicative property of logs: $logy = loga + \log(x^n)$
Using the power rule for logs: $logy = loga + n \cdot logx$

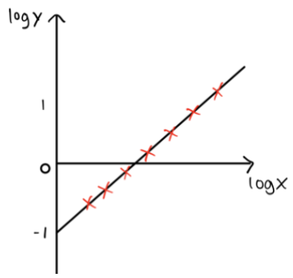
Recall that if a base of a logarithm is not explicitly written then you can assume it is 10.

This is now in a linear form.

$$Y = C + M \cdot X$$



Here is a graph showing points on the curve $y = 0.1x^{1.8}$, which is of the above form $y = ax^n$. We can see a pattern here, but it isn't linear.



Plotting $logy$ against $logx$, we can see that we now have a straight line. The gradient of this line, n, is equal to 1.8 and the y-intercept is equal to $log0.1$.

To obtain a linear relationship corresponding to a model of the form $y = kb^x$, we use the coding $Y = logy$ and $X = x$:

- If you have a model of the form $y = kb^x$, then the linear relationship between y and x is given by $logy = logk + xlogb$.

In this case, we need to plot $logy$ against x to obtain a linear model.

Example 1: The heights, h cm, and masses, m kg, of a sample of Galapagos penguins are recorded. The data are coded using $y = logm$ and $x = logh$ and it is found that a linear relationship exists between x and y. The equation of the regression line of y on x is $y = 0.0023 + 1.8x$.
Find an equation to describe the relationship between m and h, giving your answer in the form $m = ah^n$, where a and n are constants to be found.

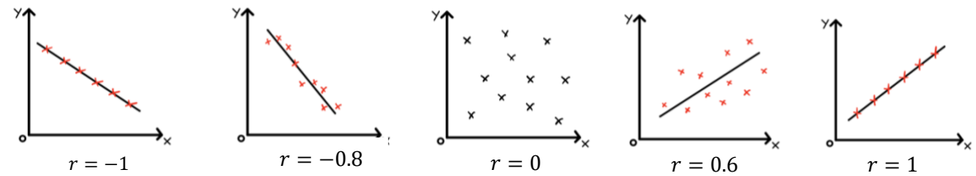
Using the regression line and substituting the coding:	$logm = 0.0023 + 1.8logr$
Using the power rule and taking the $logr$ term to the other side:	$logm - \log(r^{1.8}) = 0.0023$
Using the division law for logs and then using the relationship between exponentials and logs:	$log\left(\frac{m}{r^{1.8}}\right) = 0.0023 \Rightarrow \frac{m}{r^{1.8}} = 10^{0.0023}$
Simplifying: We can see that $a = 10^{0.0023}, n = 1.8$	$m = 10^{0.0023}(r^{1.8})$

Measuring correlation

The product moment correlation coefficient (PMCC) is a measure that describes the strength of the linear correlation between two variables. The PMCC **for a sample of data** is denoted by r, while **for a population** we denote the PMCC by ρ. The PMCC can only take values between -1 and 1.

- If $r = 1$ then there is a perfect positive linear correlation. All points will lie on a straight line.
- If $r =$ -1 then there is a perfect negative linear correlation. All points will lie on a straight line.
- If $r = 0$ then there is no linear correlation.

Here are a selection of scatter graphs that help to better understand how to interpret the PMCC:



You need to be able to use your calculator to find the PMCC for bivariate data (data involving two variables). The method for doing so depends on which calculator you are using, so refer to your calculator's handbook or a relevant online tutorial if you are unsure how to calculate the PMCC.

Hypothesis testing for zero correlation

You need to be able to carry out hypothesis tests on a sample of bivariate data to find out if we can establish a linear relationship for the entire population. The idea is that we calculate the PMCC for the sample and compare it to a critical value which will tell us whether or not a linear relationship is likely to exist. The procedure for these questions can be consolidated into four steps:

1. First write down your null and alternative hypotheses. Your null hypothesis is always $\rho = 0$, while your alternative hypothesis will depend on what you are told in the question.
2. Using your calculator, work out the PMCC of the sample data, r.
3. Use the significance level and the sample size given to you in the question to find the critical value. You will need to refer to the "product moment coefficient" table in the formula booklet (or at the back of the Edexcel textbook) to find this value.
4. Take the absolute value of your PMCC, r, and compare to the critical value. If your absolute value is greater than the critical, then you should reject the null hypothesis. Otherwise you should accept the null hypothesis. Don't forget to write a full conclusion in the context of the question.

Example 2: Twelve students sat two biology tests, one theoretical the other practical. Their marks are shown below:

Marks in theoretical test, t	5	9	7	11	20	4	6	17	12	10	15	16
Marks in practical test, p	6	8	9	13	20	9	8	17	14	8	17	18

a) Find the product moment correlation coefficient for these data, correct to 3 significant figures.
b) A teacher claims that students who do well in their theoretical test tend to do well in their practical test. Test this claim at the 0.05 significance level, stating your hypotheses clearly.

a) Using a calculator, we find the PMCC:	Using a calculator, $r = 0.935$.
b) The teacher claims that better scores in the theoretical test are likely to give better practical scores, so we are testing for a positive correlation. Therefore, our hypotheses are:	$H_0: \rho = 0, H_1: \rho > 0$.
To find the critical value, note that the sample size, n, is 12 and the significance level is 5% (one-tail). We use this to find that our critical value is 0.4973.	Significance level: 5%, $n = 12 \Rightarrow$ critical value $= 0.4973$
We compare the absolute value of our PMCC to the critical value. Our absolute value is 0.935, which is greater than 0.4973.	$0.935 > 0.4973$
As a result, we choose to reject the null hypothesis. We finish by writing a conclusion in context of the question.	∴ Reject the null hypothesis. We can conclude that there is sufficient evidence to suggest that students who do well in theoretical Biology tests also do well in practical Biology tests.