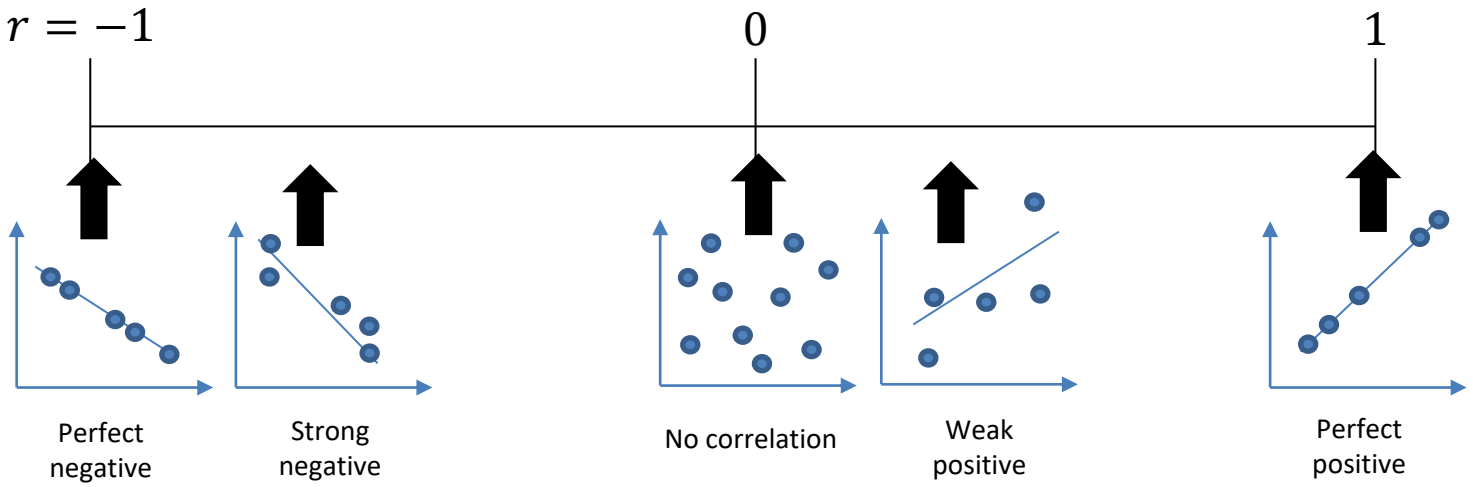
Stats2 Chapter 1: Measuring Correlation

Product Moment Coefficient

Measuring Correlation

You're used to use 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

You must have a calculator that is capable of calculating r directly: in the A Level 2017+ syllabus you are no longer required to use formulae to calculate r.

x	y
1	3
2	6
3	5
4	8


6: Statistics

$y = a + bx$

Data Entry

PMCC

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

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

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.

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 $r = 0.868$

Example

[Textbook] 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.

a ?

b ?

c ?

This is a common exam question. The important bit is **evaluating the suitability of the chosen model** (in this case a linear regression model, i.e. line of best fit). The closer r is to 1 or to -1, the more suitable this linear regression model.

Example

[Textbook] 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.

a Data on daily maximum gust is not available for these days.

b $r = 0.9533$

c r is close to 1 so there is a strong positive correlation between daily mean windspeed and daily maximum gust. This means that the data points lie close to a straight line, so a linear regression model is suitable.

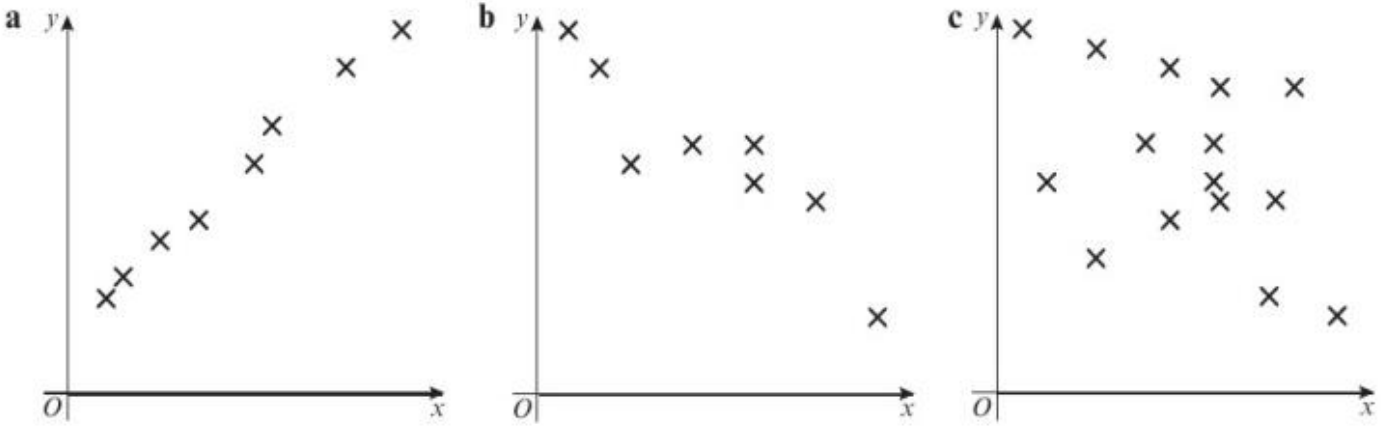
This is a common exam question. The important bit is **evaluating the suitability of the chosen model** (in this case a linear regression model, i.e. line of best fit). The closer r is to 1 or to -1, the more suitable this linear regression model.

Exercise 1.2

Pearson Stats/Mechanics Year 2
Pages 3-5

Homework Exercise

1 Suggest a value of r for each of these scatter diagrams:



2 The following table shows 10 observations from a bivariate data set.

v	50	70	60	82	45	35	110	70	35	30
m	140	200	180	210	120	100	200	180	120	60

a State what is measured by the product moment correlation coefficient.

b Use your calculator to find the value of the product moment correlation coefficient between v and m.

3 In a training scheme for young people, the average time taken for each age group to reach a certain level of proficiency was measured. The table below shows the data.

Age, x (years)	16	17	18	19	20	21	22	23	24	25
Average time, y (hours)	12	11	10	9	11	8	9	7	6	8

a Use your calculator to find the value of the product moment correlation coefficient for these data.

b Use your answer to part **a** to describe the correlation between the age and average time taken based on this sample.

Homework Exercise

4 The number of atoms of a radioactive substance, n, is measured at various times, t minutes after the start of an experiment. The table below shows the data.

Time, t	1	2	4	5	7
Atoms, n	231	41	17	7	2
log n					

Hint For part **b** enter corresponding values of t and log n into your calculator.

The data is coded using $x = t$ and $y = \log n$.

a Copy and complete the table showing the values of $\log n$. **(2 marks)**

b Calculate the product moment correlation coefficient for the coded data. **(1 mark)**

c With reference to your answer to **b**, state whether an exponential model is a good fit for these data. **(2 marks)**

The equation of the regression line of y on x is found to be $y = 2.487 - 0.320x$.

d Find an expression for n in terms of t, giving your answer in the form $n = ab^t$, where a and b are constants to be found. **(3 marks)**

5 The width, w cm, and the mass, m grams, of snowballs are measured. The table below shows the data.

Width, w	3	4	6	8	11
Mass, m	23	40	80	147	265
log w					
log m					

The data are coded using $x = \log w$ and $y = \log m$.

a Copy and complete the table showing the values of $\log w$ and $\log m$. **(3 marks)**

b Calculate the product moment correlation coefficient for the coded data. **(1 mark)**

c With reference to your answer to **b**, state whether a model in the form $y = kx^n$ where k and n are constants is a good fit for these data. **(2 marks)**

d Determine the values of k and n. **(3 marks)**

Homework Exercise

6 From the large data set, the daily mean air temperature, t °C, and the rainfall, f mm, were recorded for Perth on seven consecutive days in August 2015.

Temp, t	18.0	16.4	15.3	15.0	13.7	10.2	12.0
Rainfall, f	3.0	13.0	4.6	32.0	28.0	63.0	22.0

a Calculate the product moment correlation coefficient for these data. **(1 mark)**

b With reference to your answer to part **a**, comment on the suitability of a linear regression model for these data. **(2 marks)**

7 From the large data set, the daily total rainfall, x mm, and the daily total sunshine, y hours, were recorded for Camborne on seven consecutive days in May 2015.

Rainfall, x	2.2	tr	1.4	4.4	tr	0.2	0.6
Sunshine, y	5.2	7.7	5.6	0.3	5.1	0.1	8.9

a State the meaning of 'tr' in the table above. **(1 mark)**

b Calculate the product moment correlation coefficient for these 7 days, stating clearly how you deal with the entries marked 'tr'. **(2 marks)**

c With reference to your answer to part **b**, comment on the suitability of a linear regression model for these data. **(2 marks)**

Challenge

Data are recorded for two variables, x and y.

x	3.1	5.6	7.1	8.6	9.4	10.7
y	3.2	4.8	5.7	6.5	6.9	7.6

By calculating the product moment correlation coefficients for suitably coded values of x and y state, with reasons, whether these data are more closely modelled by a relationship of the form $y = ab^x$ or a relationship of the form $y = kx^n$, where a, b, k and n are constants.

Homework Answers

1 Answers close to

a 0.9 **b** −0.7 **c** −0.3

2 **a** The type and strength of linear correlation between v and m.

b 0.870

3 **a** −0.854

b There is negative correlation. The relatively older young people took less time to reach the required level.

4 **a**

Time, t	1	3	4	5	7
Atoms, n	231	41	17	7	2
log n	2.36	1.61	1.23	0.845	0.301

b −0.980 (3 s.f.)

c There is an almost perfect negative correlation with data in the form log n against t, which suggests an exponential decay curve.

d $a = 307$ (3 s.f.), $b = 0.479$ (3 s.f.)

5 **a**

Width, w	3	4	6	8	11
Mass, m	23	40	80	147	265
log w	0.477	0.602	0.778	0.903	1.04
log m	1.36	1.60	1.90	2.17	2.42

b 0.9995

c A graph of log w against log m is close to a straight line as the value of r is close to 1, therefore $m = kw^n$ is a good model for this data.

d $n = 1.88$ or 1.89 (3 s.f.), $k = 2.91$ (3 s.f.)

6 **a** −0.833

b −0.833 is close to −1 so the data values show a strong to moderate negative correlation. A linear regression model is suitable for these data.

7 **a** A 'trace or tr' of rain is an amount less than 0.05mm.

b −0.473 (3 s.f.), treating 'tr' values as 0.

c The data shows a weak negative correlation so a linear model may not be best, there may be other variables affecting the relationship or a different model might be a better fit.

Challenge

r for x and y: 0.999 (3 s.f.)

r for log x and log y: 1.00 (3 s.f.)

r for x and log y: 0.985 (3 s.f.)

Therefore the most suitable model would be in the form $y = ax^n$