Correlation and Regression

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
x	*w*	4	4	8	7	12	12	3	4	7	10
y	*g*	13	12	19	23	33	37	10	n/a	n/a	23

Correlation

1. What is bivariate data? Data which has pairs of values for two variables. It can be shown on a scatter graph.

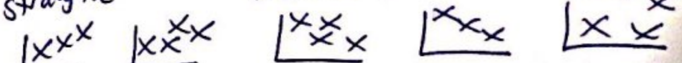
g ⌊ w

2. What is the independent or explanatory variable?
variable that is controlled (x axis)

3. What is the dependent or response variable?
variable that's measured. (y axis)

4. State the meaning of n/a in the table
not applicable - no data on those days.

5. Draw a scatter graph to represent the data labels!
no linear correlation.
straight



6. What is correlation?
correlation describes the nature of the linear relationship between two variables.

7. Describe the correlation between windspeed and gust
Strong positive correlation.

8. Interpret your answer (context)
As windspeed increases so does gust (increases)

9. Do windspeed and gust have a causal relationship? need to consider context of Qu.
correlation does not imply causation. (Need proof).

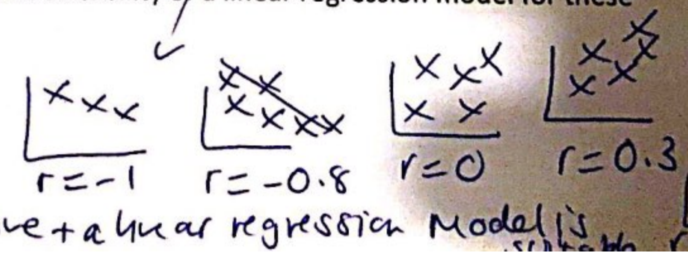
find on calc

10. What is the product moment correlation coefficient (pmcc or r)?
r = 0.9532973672   r = 0.9533

11. Calculate the product moment correlation coefficient (pmcc or r)
PMCC describes the linear correlation between two variables. It takes values between 1 and -1

12. With reference to r comment on the suitability of a linear regression model for these data

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



Regression equation of line of best fit (gan w)

1. What is the least squares regression line of y on x?

A straight line that minimises the sum of the squares $d_1^2 + d_2^2 + d_3^2$ etc of the distances of each data point from the line. Σd_i^2 minimises

plot two points - when $x = 10$, $y = 27.71$

2. Calculate the least squares regression line of y on x

$y = a + bx$ $y = 1.8478 + 2.5869x$

3. Draw the least squares regression line of ~~y on x~~ on your scatter graph (gan w)

4. Interpret a When there is zero windspeed you can expect a gust of 1.8478 knots

5. Interpret b

for every extra 1 knot in windspeed the gust goes up by 2.5869 knots.

6. Use your regression line to estimate the gust for a windspeed of 6 knots.
   Comment on the reliability of this estimate.

$y = 1.8478 + 2.5869(6)$ $y = 17.3692$ knots.
$W = 6$ is within the range of the data (interpolation) so is likely to be accurate.

7. Use your regression line to estimate the gust for a windspeed of ~~13~~ 23 knots.
   Comment on the reliability of this estimate.

$y = 61.3258$ $W = 23$ is ~~at~~ side the range of data (extrapolation) so unlikely to be accurate

8. Use your regression line to estimate the windspeed for a gust of 20 knots.
   Comment on the reliability of this estimate.

The independent (explanatory variable) is (x) W (windspeed) so you cannot use this model to predict x given a y value.

9. Calculate the regression line of x on y

y on x $y = a + bx$

x on y $x = a + by$ $x = 0.035055 + 0.35129y$.

$x = 0.035055 + 0.3512 \times 20$

$x = 7.059$.