HIGHER

H11-30min Regression and the PMCC: 30-minute Pack

Total Marks

EDEXCEL 1ST0 · Calculator allowed · about 30 minutes

Name: _____ Date: ____ / ____ / _____

Answer ALL questions. Show all your working.

1 For each statement about the PMCC and regression lines, write down whether it is True or False.

(i) The value of the product moment correlation coefficient, r, must always lie between -1 and 1 inclusive. **(1)**

.....

(ii) A product moment correlation coefficient of r = 0 always means there is no relationship at all between the two variables. **(1)**

.....

(iii) The regression line of y on x always passes through the point (x-bar, y-bar), the means of the x and y data. **(1)**

.....

(iv) If two variables have a strong linear correlation, this always means that changes in one variable directly cause changes in the other. **(1)**

.....

(Total for Question 1 is 4 marks)

2 A personal trainer records, for 8 gym members, the number of personal training sessions they have attended and the increase in the weight (kg) they can bench press since starting. The scatter graph shows the results, with the points plotted but no line of best fit drawn.

(a) State one feature, other than being drawn with a ruler, that a well-drawn line of best fit should have. **(1)**

(b) On the grid, draw a straight line of best fit for this data. **(2)**

(c) Use your line of best fit to estimate the increase in bench press weight after 4.5 sessions. **(2)**

(d) State whether your estimate in part (c) was found by interpolation or extrapolation. **(1)**

(Total for Question 2 is 6 marks)

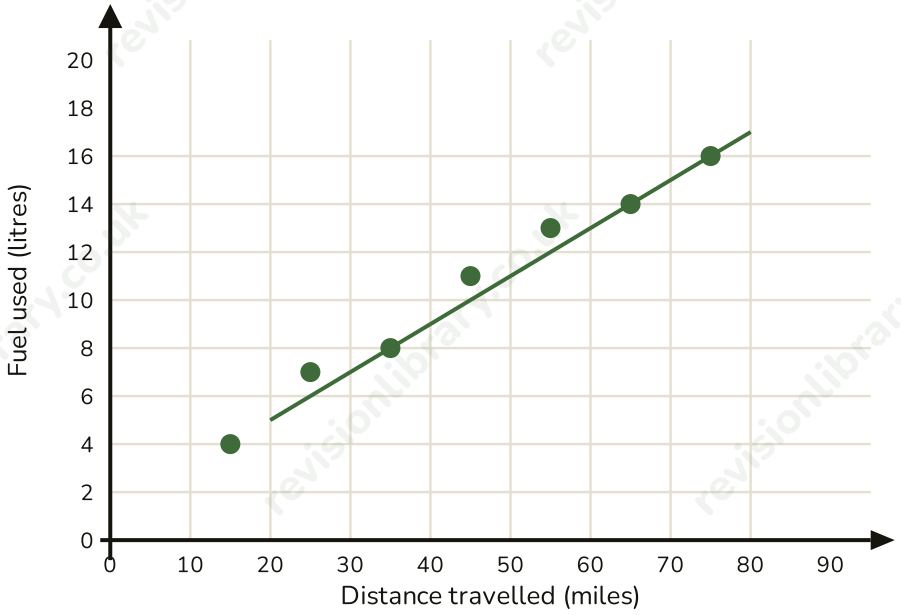
3 For a set of paired data, the regression line of y on x is $y = 1.5x + 8$. The mean of the x-values is $\bar{x} = 6$ and the mean of the y-values is $\bar{y} = 17$.

(a) Show that the point $(\bar{x}, \bar{y})$ lies on this regression line. **(2)**

(b) Explain why this result is not a coincidence. **(1)**

(Total for Question 3 is 3 marks)

4 A delivery company records, for 7 journeys, the distance travelled (miles) and the fuel used (litres). The scatter graph shows the results, together with a line of best fit that passes through the points (20, 5) and (80, 17).



(a) Find the equation of the line of best fit in the form $y = mx + c$, where x is the distance travelled and y is the fuel used. **(3)**

.....

(b) Use the equation to estimate the fuel used for a distance of 50 miles. **(1)**

(c) The distances in the sample ranged from 20 to 80 miles. Comment on the reliability of using this equation to estimate the fuel used for a 150-mile journey. **(2)**

(Total for Question 4 is 6 marks)

5 For a sample of 10 runners whose weekly training distance ranged from 10 to 40 miles, the regression line relating weekly training distance, d (miles), to 10 km race time, T (minutes), is:

$T = 65 - 0.8d$

(a) Use the equation to estimate the race time for a runner who trains 25 miles per week. **(2)**

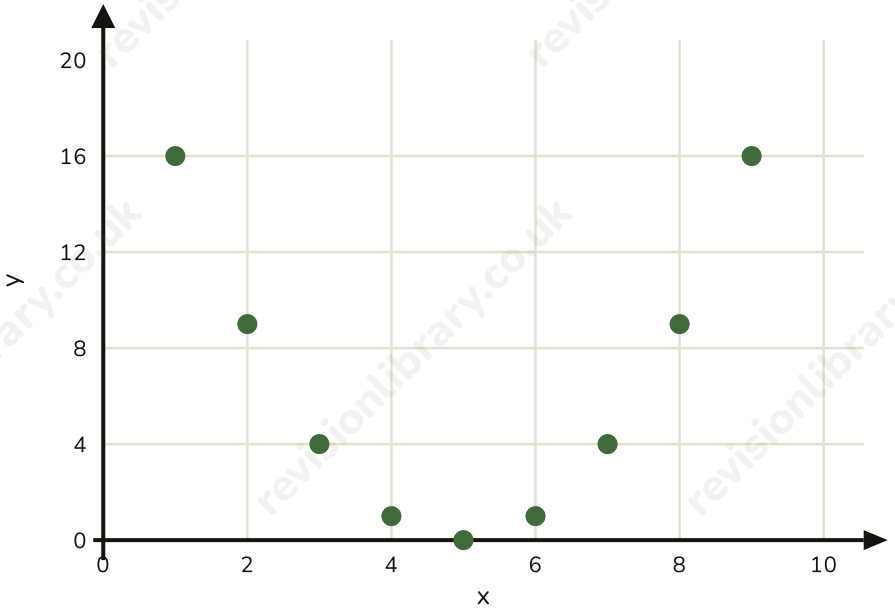
(b) Explain why it would not be appropriate to use this equation to estimate the race time for a runner who trains 80 miles per week. **(2)**

(c) Interpret what the value -0.8 represents in this context. **(1)**

(d) State whether the value 65 (the intercept) has a sensible real-world interpretation in this context, and explain your answer. **(2)**

(Total for Question 5 is 7 marks)

6 A researcher plots data for two variables, x and y. The scatter graph shows a clear curved (U-shaped) pattern. She calculates the product moment correlation coefficient and finds r = 0, and concludes that 'there is no relationship at all between x and y'.



(a) State what a value of r close to 0 normally suggests about the linear relationship between two variables. **(1)**

(b) Explain why the researcher's conclusion that 'there is no relationship between x and y' is not fully justified, given the shape of the scatter graph. **(2)**

(Total for Question 6 is 3 marks)

Mark scheme · H11-30min Regression and the PMCC: 30-minute Pack

Question 1

- (i) **B1** True cao
- **(i) Answer: True**
- (ii) **B1** False cao
- **(ii) Answer: False**
- (iii) **B1** True cao
- **(iii) Answer: True**
- (iv) **B1** False cao
- **(iv) Answer: False**

Question 2

- (a) **B1** follows the direction/trend of the points, passing through (or close to) the mean point, with a roughly equal number of points above and below the line, oe
- **(a) Answer: It should follow the trend of the data with roughly equal numbers of points on each side of the line.**
- (b) **B1** a single ruled straight line drawn across the full range of the plotted points
- (b) **B1** line follows the trend of the data with a roughly equal number of points on each side, e.g. close to a line from about (1,2) to (8,13)
- **(b) Answer: A ruled line running from approximately (1,2) to (8,13).**
- (c) **M1** reads a value from a single ruled line of best fit at x = 4.5
- (c) **A1** answer consistent with their line; expected value approximately 7.5 kg (accept 6.5 to 8.5), ft
- **(c) Answer: Approximately 7.5 kg (accept 6.5-8.5)**
- (d) **B1** Interpolation cao
- **(d) Answer: Interpolation**

Question 3

- (a) **M1** substitutes x = 6 into y = 1.5x + 8
- (a) **A1** y = 1.5(6) + 8 = 17, which equals y-bar, confirming the point lies on the line
- **(a) Answer: y = 1.5(6) + 8 = 17 = y-bar, so (6,17) lies on the line.**
- (b) **B1** the regression line of y on x always passes through the mean point (x-bar, y-bar), so this must be true for any regression line, oe
- **(b) Answer: The regression line of y on x always passes through the mean point (x-bar, y-bar), so it must pass through (6,17).**

Question 4

- (a) **M1** gradient = (17-5)/(80-20), oe
- (a) **A1** m = 0.2
- (a) **A1** y = 0.2x + 1 cao (c found correctly, e.g. from 5 = 0.2(20) + c)
- **(a) Answer: y = 0.2x + 1**
- (b) **B1** 11 litres cao, ft from (a)
- **(b) Answer: 11 litres**
- (c) **B1** identifies that 150 miles is well outside the range of the original data (20 to 80 miles), so this is extrapolation
- (c) **B1** correct conclusion, e.g. the linear pattern may not continue that far (fuel use per mile could change on longer/motorway journeys), so the estimate would be unreliable, oe
- **(c) Answer: Unreliable - 150 miles is far outside the 20-80 mile range of the data, so this is extrapolation and the same linear relationship may not continue.**

Question 5

- (a) **M1** substitutes d = 25 into T = 65 - 0.8d
- (a) **A1** 45 minutes cao
- **(a) Answer: 45 minutes**
- (b) **B1** 80 miles is well outside the range of the original data (10 to 40 miles), so this would be extrapolation
- (b) **B1** correct conclusion, e.g. the linear relationship may not continue at such high training distances (e.g. risk of injury/diminishing returns), so the estimate would be unreliable, oe
- **(b) Answer: 80 miles is far outside the 10-40 mile range of the data, so using the equation here is extrapolation and may give an unreliable estimate.**
- (c) **B1** for each extra mile trained per week, race time decreases by about 0.8 minutes, on average, oe
- **(c) Answer: For each extra mile trained per week, race time decreases by about 0.8 minutes, on average.**
- (d) **B1** No cao
- (d) **B1** d = 0 lies outside the range of the data (10 to 40 miles), so it is not appropriate to interpret it directly - a runner who does no weekly training would not be represented by this model, oe
- **(d) Answer: No; d = 0 lies outside the range of the data (10-40 miles), so it should not be directly interpreted - it would represent a runner with no training at all, which is not represented in the data.**

Question 6

- (a) **B1** there is little or no LINEAR relationship between the variables, oe
- **(a) Answer: There is little or no linear relationship between the variables.**
- (b) **B1** the PMCC only measures the strength of a LINEAR relationship, so a small r does not rule out a strong non-linear (e.g. curved) relationship
- (b) **B1** the scatter graph shows a clear curved (U-shaped) pattern, so x and y are actually strongly related, just not in a straight-line way, oe
- **(b) Answer: r only measures linear relationships; the scatter graph clearly shows a strong curved (U-shaped) pattern, so x and y are strongly related - just not in a straight line, which is why r is close to 0 despite the clear pattern.**