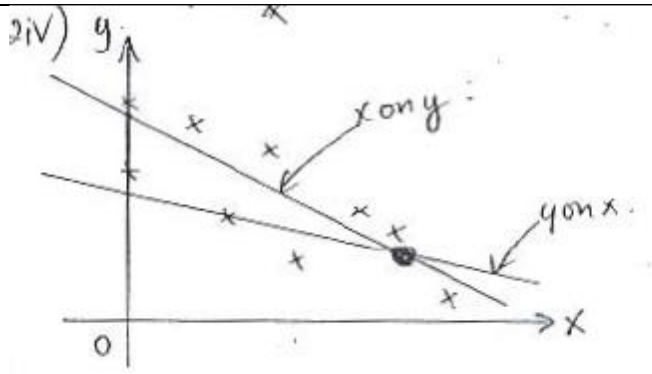
Correlation and Linear Regression

1	$\sum x^2 - \frac{(\sum x)^2}{n} = 21830 - \frac{360^2}{8} = 5630$ $\sum y^2 - \frac{(\sum y)^2}{n} = 3500 - \frac{140^2}{8} = 1050$ $\sum xy - \frac{\sum x \sum y}{n} = 3985 - \frac{(360)(140)}{8} = -2315$ $\therefore r = \frac{-2315}{\sqrt{5630 \times 1050}} = -0.952$ In general, as x increases, y decreases in an almost linear pattern. However, it would be better to sketch a scatter diagram based on the 8 pairs of data to verify.

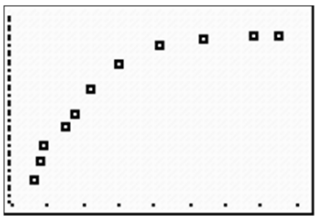
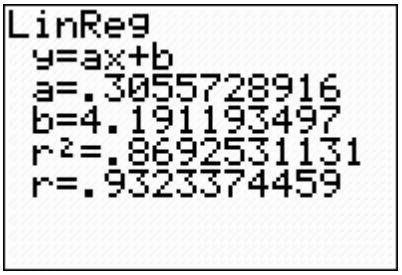
2(i)	2i) Using GC, $y = -1.121x + 8.658$ (y on x) [LinReg(ax+b) L1, L2, Y1] x values, y-values Using GC, $x = -0.537y + 5.721$ (x on y)
(ii)	2ii) Using line of regression of y on x, $y = -1.121(1.6) + 8.658$ $= 6.8644$ Thus there would be 6864 ≈ 6860 tourists 2iii) Using line of regression of x on y, $x = -0.537(3.2) + 5.721$ $= 4.00$

(iv)	 Using GC, [2-var Stat L1, L2], we get $\bar{x} = 2.689$, $\bar{y} = 5.644$ Subst $(\bar{x}, \bar{y})$ into both lines, we get $-1.121(2.689) + 8.658 = 5.644$ $-0.537(5.644) + 5.721 = 2.689$ Hence it is verified that $(\bar{x}, \bar{y})$ lies on both lines

3(i)	$y - \bar{y} = b(x - \bar{x}) \Rightarrow y = bx + (\bar{y} - b\bar{x})$ Since $y = -0.8x + 13.6$, by comparing coefficients, $b = -0.8$, and $\bar{y} - b\bar{x} = 13.6 \Rightarrow \bar{y} = -0.8(4.5) + 13.6 = 10$ $\therefore \sum y = 10 \times 8 = 80$ $b = -0.8 \Rightarrow \dfrac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sum x^2 - \dfrac{(\sum x)^2}{n}} = -0.8$ $\therefore \sum xy - \dfrac{\sum x \sum y}{n} = -0.8\left[\sum x^2 - \dfrac{(\sum x)^2}{n}\right]$ $\Rightarrow \sum xy - \dfrac{36 \times 80}{8} = -0.8\left[204 - \dfrac{36^2}{8}\right]$ $\Rightarrow \sum xy = 326.4$
(ii)	Correct: $\sum y = 80$, $\sum xy = 326.4$ From data: $\sum y = 82.7$, $\sum xy = 348$ Difference in $\sum y$ is 2.7 Difference in $\sum xy$ is 21.6 $x(2.7) = 21.6 \Rightarrow x = 8$ 9.6 is wrong. When x = 8, y = 6.9.

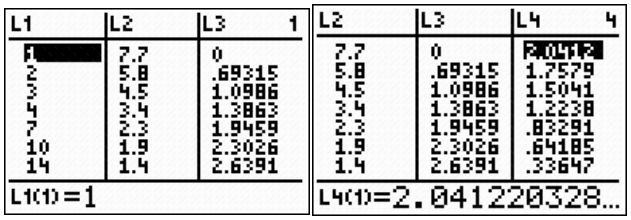
4(i)	-0.912
(ii)	Although $r_B = -0.912$ is close to -1, which indicate a strong negative linear correlation between X and Y, the scatter diagram shows that the data can be better represented by a non-linear curve
(iii)	$y = 0.0721(x-69)^2 + 46.2$ $r_{(x-69)^2, y} =$ 0.9981; proposed model is better since |r| is closer to 1 here
(iv)	54.9, As $x = 80$ is out of the range, this estimate is not reliable

5(a)(i)	$r = 0.89241 = 0.892$ (3 s.f.)
(ii)	$r = 0.95956 = 0.960$ (3 s.f.) Since $r = 0.960$ is closer to 1 than $r = 0.892$, the model in (i) is less suitable than the model in (ii).
(iii)	Regression line of F on x: $F = 0.35903 + 0.029245x$ $F = 0.359 + 0.0292x$ Regression line of x on F: $x = 204.51 + 31.484F$ $x = 205 + 31.5F$
(iv)	Using $F = 0.35903 + 0.029245x$, $100 = 0.35903 + 0.029245t^2$ $t = 58.37047 = 58.4$ s (3 s.f.)

6(i)	 Based on the scatter diagram, the linear model is not suitable even though r (= 0.906) is quite close to 1.
(ii)	Choose Model (b): $y = ax^b$ Reason: The graph of $y = ax^b$ fits the scatter diagram better. : From the scatter diagram, we see that as x increases, y increases at a decreasing rate.
(iii)	r = 0.932 $y = ax^b$ $\ln y = \ln(ax^b)$ $\ln y = \ln a + b \ln x$ $\ln y = 4.1912 + 0.3056 \ln x$ $\ln a = 4.1912 \Rightarrow a = 66.1$ $b = 0.306$ 
(iv)	when $y = 110$, $x = 5.29$. Extrapolation hence not reliable.

7(i)	Product moment correlation coefficient r = 0.979. There exists a strong positive linear correlation between x and y.
(ii)	Equation of least square regression line is $y = 18.5 + 0.564\,x$
(iii)	Given that $y = 30$, x = 20.4 from the equation. She left at 7am. This estimate is reliable since $r \approx 1$ so we can use the equation of y on x to estimate x, giving y and y = 30 is within the given data range. (2 reasons)
(iv)	z = time available – time taken $= 50 - x - y$ $= (50 - x) - (a + bx)$ $= (50 - x) - (18.5 + 0.564x)$ $= 31.5 - 1.564x$
(v)	For $z = 0$, $x = \dfrac{31.5}{1.564} = 20.14 \approx 20\text{min}$ The latest time when Ms Chan leaves her house is 7 a.m

8(i)	r = -0.952 For the depths sampled the moisture content **decreases approximately linearly** with an increase in the depth of the sample. Since r is close to 1 , this implies that the regression line of m on x and x on m are **almos**t identical or close to each other
(ii)	m = - 2.2048x + 83.583. When m = 50 , $x = \dfrac{83.583 - 50}{2.2048} = 15.232 = 15.2$ Reliable since the value of m is within the range of the given data. Not extrapolating.

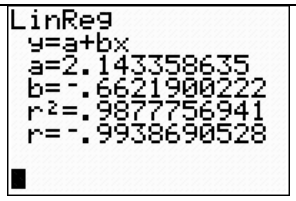
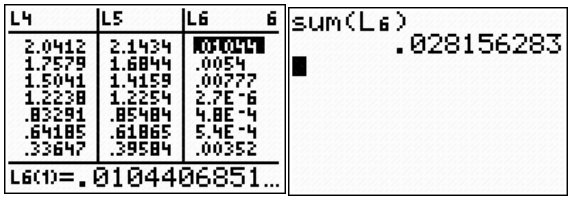
9(i)	$PV^c = k$ $\ln P + c\ln V = \ln k$ $\ln V = \dfrac{\ln k}{c} - \dfrac{1}{c}\ln P$ $\therefore y = a + bx$ is a straight line, where $a = \dfrac{\ln k}{c}$ and $b = -\dfrac{1}{c}$ are constants
(ii)	Using G.C., enter values of P and V in list L1 and L2 . Let L3 = ln$P$ and let L4 = ln$V$ 

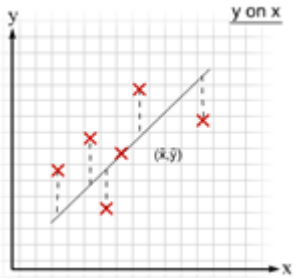
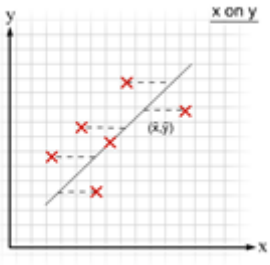
L1	L2	L3 1
1	7.7	0
2	5.8	.69315
3	4.5	1.0986
4	3.4	1.3863
7	2.3	1.9459
10	1.9	2.3026
14	1.4	2.6391

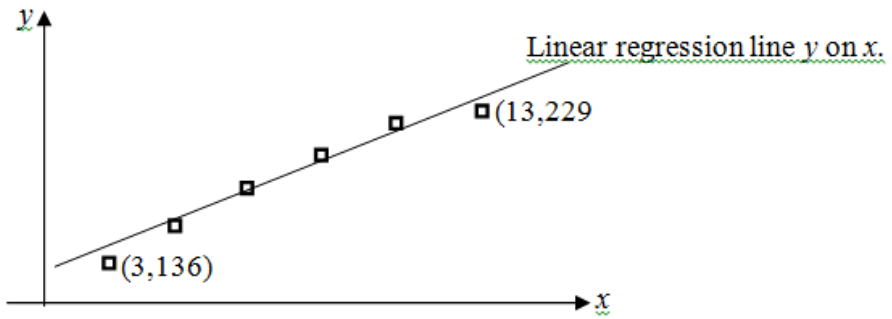
L1(1)=1

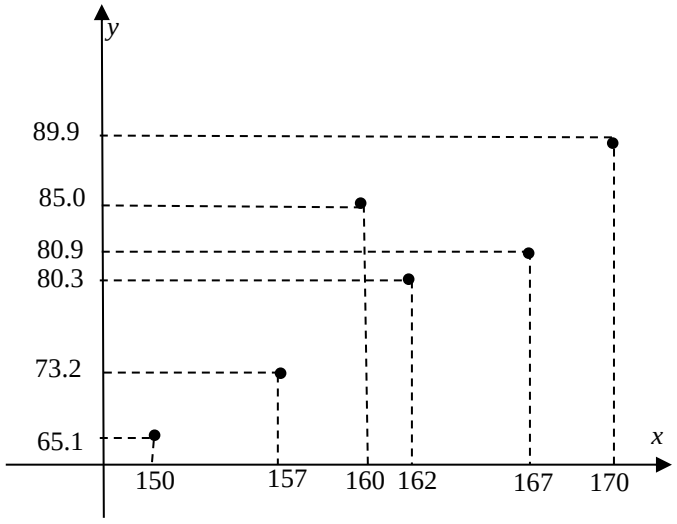
L2	L3	L4 4
7.7	0	1.7579
5.8	.69315	1.5041
4.5	1.0986	1.2238
3.4	1.3863	.83291
2.3	1.9459	.64185
1.9	2.3026	.33647
1.4	2.6391	

L4(1)=2.041220328...

	 $r = -0.994$
(iii)	By using G.C., the required estimated regression line of y on x is $y = 2.1434\text{-}0.66219x$ $y = 2.14 - 0.662x$ (to 3 s.f.) $\ln V = 2.1434 - 0.66219 \ln P$ $\ln V = 2.1434 - 0.66219 \ln(8) = 0.76641$ $V = e^{0.76641} = 2.15203 = 2.2$ (to 1 d.p.) (ans) Since $r = -0.994 \approx -1$ which indicates that the sample values of x and y are almost perfectly linearly correlated and $\ln(1) \le \ln(8) \le \ln(14)$, therefore the prediction is reliable.
(iv)	$\ln V = \frac{\ln k}{c} - \frac{1}{c} \ln P$ $y = 2.1434 - 0.66219x$ $\frac{1}{c} = 0.66219$ $c = \frac{1}{0.66219} = 1.51014 = 1.51$ (to 3 s.f.) $\frac{\ln k}{c} = 2.1434$ $\ln k = 2.1434 \times 1.51014 = 3.2368$ $k = e^{3.2368} = 25.452 = 25.5$ (to 3 s.f.)
(v)	For $\sum (y - Y')^2$ to be minimum, Then $Y' = a + bx$ must be $y = 2.1434\text{-}0.66219x$ By using G.C., let L5 = 2.1434–0.66219 L3 and $L6 = (L4 - L5)^2$  Minimum value of $\sum (y - Y')^2 = 0.0282$ (to 3 s.f.)

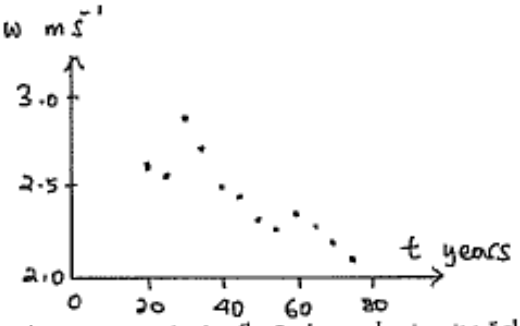
10(a)	  The least squares regression line of y on x is the line with the minimum sum of the squared vertical distances from each point to the line while the least squares regression line x on y is the line with the minimum sum of the squared horizontal distances from each point to the line. $(\bar{x},\bar{y})$ is the point through which the estimated regression line of y on x and that of x on y both pass.
(b)(i)	From the regression line y on x, we have $\bar{y}=\frac{68}{7}\bar{x}+\frac{786}{7}$. Since $\bar{x}=8$, $\bar{y}=\frac{68}{7}(8)+\frac{786}{7}=190$. $\Rightarrow \frac{160+\alpha+229+\beta+225+136}{6}=190 \Rightarrow \alpha+\beta=390$ …. (1) $b=\frac{\Sigma(x-\bar{x})\Sigma(y-\bar{y})}{\Sigma(x-\bar{x})^2}=\frac{68}{7} \Rightarrow \left(\Sigma xy-\frac{\Sigma x\Sigma y}{6}\right)=\frac{68}{7}\left(\Sigma x^2-\frac{(\Sigma x)^2}{6}\right)$ Hence, $\left((6660+9\alpha+7\beta)-\frac{(48)(750+\alpha+\beta)}{6}\right)=\frac{68}{7}\left(454-\frac{2304}{6}\right)$ $\Rightarrow \left((6660+9\alpha+7\beta)-8(750+\alpha+\beta)\right)=680 \Rightarrow \alpha-\beta=20$ … (2) (1) + (2) : $2\alpha=410 \Rightarrow \alpha=205$ (1) – (2) : $2\beta=375 \Rightarrow \beta=185$
(ii)	When y = 210, $7(210)=68x+786$ $\Rightarrow x=10.05882353$ = 10.1 (to 3 s.f.) The predicted value of x is likely to be reliable as estimation is done by interpolation and r = 0.985 is close to 1.

(iii)	

11(i)	$\bar{x} = 161$ (from calculator or computation) when $\bar{x} = 161$, $\bar{x} = 103.6 + 0.726\bar{y}$ $\bar{y} = (161 - 103.6)/0.726$ $= 79.06336088$ using $\bar{y} = \sum y / n$ $79.06336088 = \frac{1}{6}(65.1 + 73.2 + 85 + k + 80.9 + 89.9)$ $k = 80.3$ Use G.C. to find regression line of y on x: $y = -97.593 + 1.097x$
(ii)	Use y on x line to predict weight. When $x = 165$, $y = -97.593 + 1.097(165)$ $y = 83.4$ (1 d.p.) – using 3 d.p. of a and b to compute. or $y = 83.5$ - using full accuracy of a and b to compute
(iii)	Using G.C., $r = 0.893$  C is unusually overweight.

12

7(i)



(ii) w on t, because t is the independent variable.

(iii) The regression line is suitable since the points on the scatter diagram, for $30 \leq t \leq 55$, lie close to a straight line and the absolute value of the correlation coefficient is close to 1.

(iv)(a) With w and $\frac{1}{t}$,

$r = 0.99328 \simeq 0.993$

Since the absolute value of this correlation coefficient is closer to 1 than the previous value of -0.965, the linear model for w and $\frac{1}{t}$ is a better model compared to the linear model for w and t.

(b) Eq of reg line of w on $\frac{1}{t}$ is

$w = 1.5763 + 37.419\frac{1}{t}$

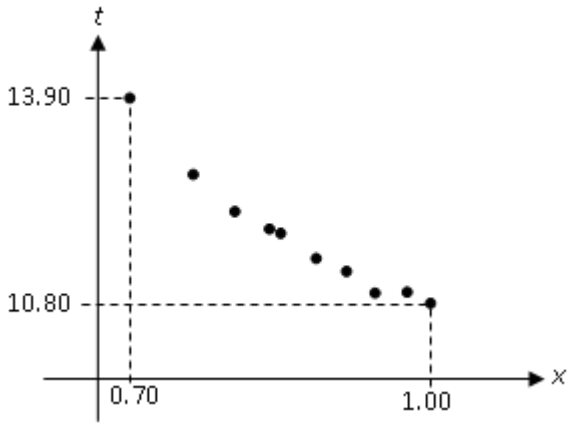
$\therefore w = 1.58 + 37.4\frac{1}{t}$

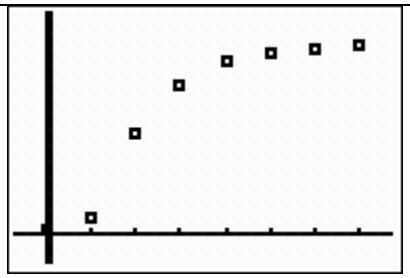
(c) When $t = 43$, $w = 2.45$.

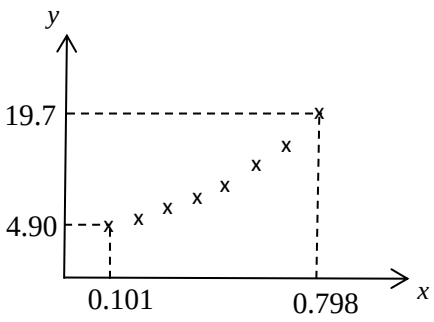
Hence the max walking speed at the age of 43 years is $2.45\ ms^{-1}$.

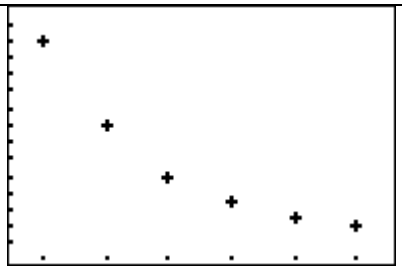
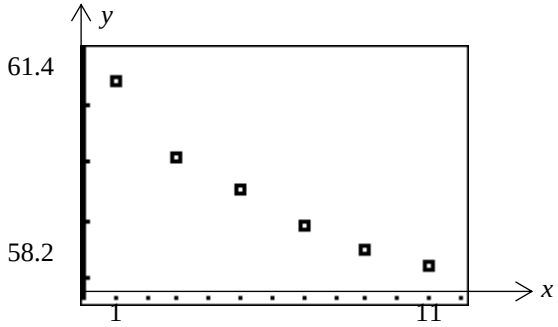
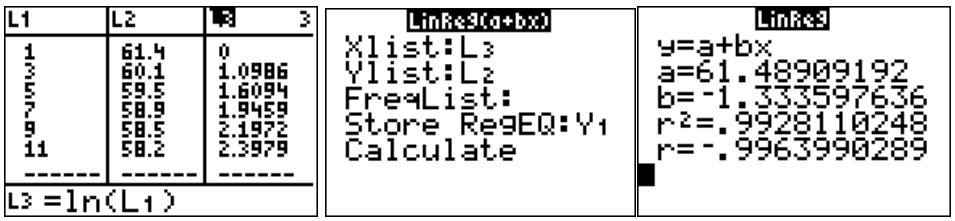
13(a)(i)	For x on y, $x = -0.3085y + 31.13 \Rightarrow x = -0.309y + 31.1$ (3 s.f.) For y on x, $y = -2.8526x + +95.999 \Rightarrow y = -2.85x + 96.0$ (3s.f.)
(ii)	Since chemical Y is the controlled variable, use regression line of x on y. $0 = -0.3085y + 31.13 \Rightarrow y = 100.91$ The estimation is not valid as this is an extrapolation, linear relation may not hold outside the range of data
(b)(i)	By comparing the linear product moment correlation for the 3 models, Model C is the most appropriate with the highest value of $\lvert r \rvert = 0.993$ as it best describes the data given.

	Using linear transformation $w = \ln x$, Regression line of w on y is $w = -0.026136y + 3.8294 \Rightarrow w = -0.0261y + 3.83$ (3 s.f.)
(ii)	Change in w = $-0.026136(5) = -0.13068 \approx -0.131$ (3 s.f.) w decreases by 0.131

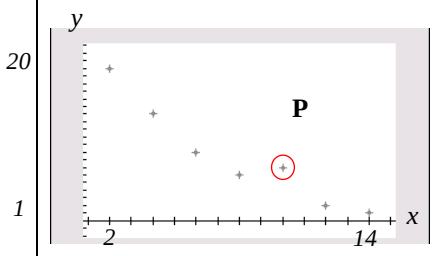
14(i)	There will be no difference as the product moment correlation coefficient is independent of the units in which the data is measured.
(ii)	The **regression line of *t* on *x*** should be used because the running time t is **dependent** on the leg length, x
(iii)	
(iv)	**Yes**. Aaron has reason to disagree because the **scatter diagram suggests** that t and x has a **curvilinear relationship** rather than a linear one.
(v)(a)	Product moment correlation coefficient between t and $\frac{1}{x^2}$ is **0.992** (3 s.f.) The new model is a **better model** because $\lvert 0.992 \rvert$ is **closer to 1** than $\lvert -0.963 \rvert = 0.963$.
(b)	Regression line is $t = 7.8603 + 2.8616\frac{1}{x^2}$ i.e. $t = 7.86 + 2.86\frac{1}{x^2}$ (3 s.f.) When t = 10, $10 = 7.8603 + 2.8616\frac{1}{x^2}$ $x^2 = \frac{2.8616}{2.1397}$ $x = 1.16$ (to 2 dec places) since $x > 0$ Thus minimum length of leg required is **1.16m**. This estimate **may not be reliable** as t = 10 is **outside the sample data range for *t***. OR Extrapolated values are unreliable

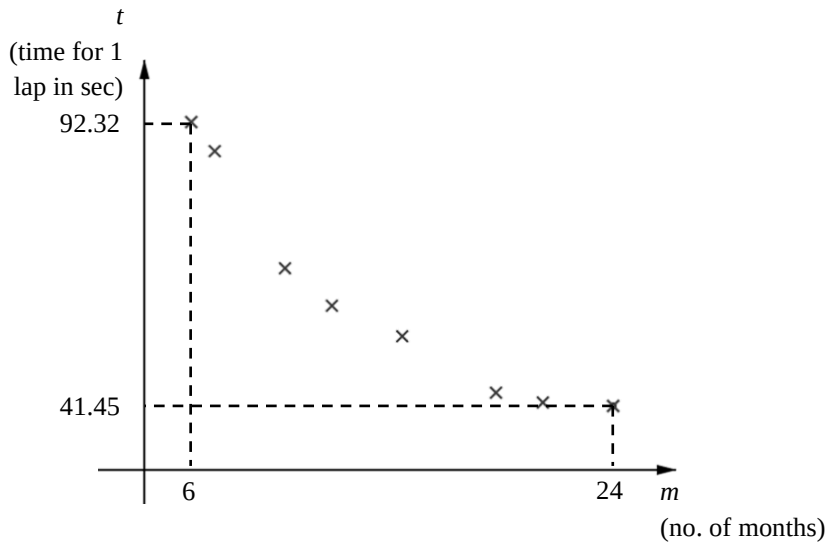
15(i)	
(ii)	A linear model is not likely to be appropriate as the area covered would then increase continuously, eventually to an infinite area.
(iii)	$D = 53$: $r = -0.99349$ Since $D = 53$ gives a value of r closest to -1, it is the most appropriate.
(iv)	$a = 4.26272$; $b = -0.60899$ Equation of regression line is $\ln(D - A) = 4.26272 - 0.60899t$ When $t = 20$, $A = 52.99964$ cm^2
(v)	D is the maximum area of the petri dish

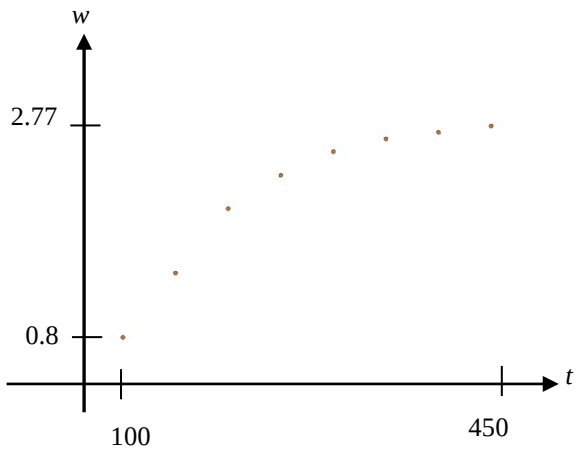
16(i)	A regression line of y on x is more appropriate as the bacteria population depends on the concentration of nutrients in the water body
(ii)	$r = 0.98119 \approx 0.981$ (3 s.f.)
(iii)	 Although $r \approx 0.981$ suggests a strong linear correlation, the scatter diagram shows that **as x increases, y increases at an increasing rate**. Therefore, a linear model **is not necessarily the best model** for the relationship between x and y.
(iv)	By G.C., $\ln y = 1.3869 + 1.9984x$ $\ln y = 1.39 + 2.00x$ (3 s.f.) When $x = 1$, $y = 29.527 \approx 29.5$ (3 s.f.) The bacteria population is 29500 (3 s.f.) Since $x = 1$ lies **outside of the data range**, the estimate is **not reliable**

17(i)	
(ii)	Scatter diagram: 
(iii)	$y = a + b\ln x$ is the better model. **Reason:** y is decreasing at a **decreasing rate** as x increases.
(iv)	**Interpretation:** Value of a represents <u>Amy's predicted weight one month</u> after she began her healthy eating lifestyle and exercise regime. $r = -0.996$ (3sf) **Useful Screenshots for reference:**  [**Remark:** Those who indicated that the linear model is the better model in (iii) will be given marks accordingly.] **<u>For those who chose $y = c + dx$ in (iii):</u>** Value of c represents <u>Amy's predicted initial weight</u> before she began her healthy eating lifestyle and exercise regime $r = -0.967$
(v)	Equation of regression line: $y = 61.489 - 1.3336\ln x$ When $y = 55$:

	$55 = 61.489 - 1.3336\ln(x)$ $x = 129.77$ (5sf) She will reach a weight of 55kg in the <u>130th month</u> after she started. **<u>For those who chose $y = c + dx$ in (iii):</u>** Equation of line as $y = 61.268 - 0.30571x$ Substituting y = 55 into equation to obtain x = 20.503 She will reach a weight of 55kg in the <u>21st month</u> after she started.
(vi)	**<u>For any model chosen in (iii):</u>** As $x \to \infty, y \to -\infty$. Thus, this implies that Amy's weight will decrease to a **negative value over time**, which is unrealistic.

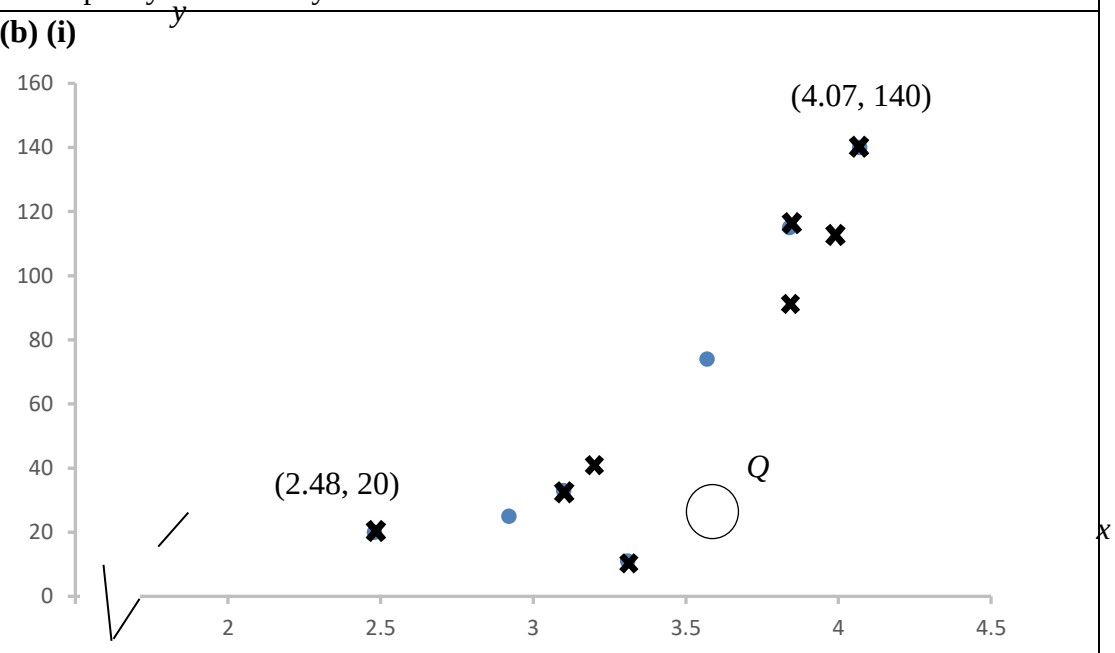
18(i)	$r = -0.9550961661 = -0.955$ (to 3 sig.fig) Since r value is close to -1, it suggests a strong negative **<u>linear</u>** correlation **<u>between x and y</u>**, hence a linear model is appropriate.
(ii)	
(iii)	With point P removed, the remaining points lie close to an exponential curve, as x increases, y decreases at a decreasing rate, hence consistent with a model of the form $y = Ae^{bx}$.
(iv)	$y = Ae^{bx} \Rightarrow \ln y = \ln A + bx$ Since <u>x is the controlled variable</u>, we use the regression line of $\ln y$ on x. From GC, $\ln y = -0.24888x + 3.6276$ When $y = 7$, $\ln 7 = -0.24888x + 3.6276$ $x = 6.75 = 7$ (nearest whole no)

19	(i)  A linear model would imply that in the long run, the time taken to swim a lap would be negative, which is unrealistic. *(Note: Extrapolation is not accepted as a reason, as the question isn't looking for a reason based on the data obtained.)*
(ii)	Using GC, for $C = 37$, $r = \underline{-0.992555}$
(iii)	The most appropriate value for C is $\underline{38}$, as the magnitude of its corresponding value of r is closest to 1.
(iv)	From GC, least squares regression line of $\ln(t-38)$ on m is $\ln(t-38) = 5.01236 - 0.16349m$ $\Rightarrow \underline{\ln(t-38) = 5.01 - 0.163m}$ (3 s.f.) $C = 38$ is the fastest time that a student can expect to complete a lap of breaststroke after spending a long time at the swim school. (Making t the subject in the equation of the regression line gives us $t = 38 + e^{5.01-0.163m}$, so as $m \to \infty$, $t \to 38$.)
(v)	When $m = 9$, $t = 38 + e^{5.01236-0.16349(9)}$ $= 72.50$ (2 d.p.) A timing of 60.33 seconds is well below the expected timing of 72.50 seconds. Therefore, we can say that the student is exceptionally strong in his/her swimming ability.
(vi)	The 8 randomly selected students might have been of different genders and ages. To make the results fairer, data could be collected separately based on genders and age ranges.

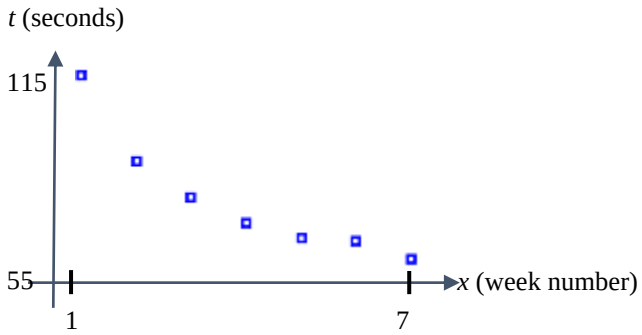
20(i)	$r = 0.925$ (3 s.f.) Acceptable answers: • As the pmcc value is close to 1, indicating a **strong positive <u>linear</u>** correlation, it suggests that a linear model is appropriate. • A linear model with positive linear correlation would suggest that the weight loss may exceed 100%, which is impossible. Thus a linear model is not appropriate.
(ii)	 (A) $w = a + b\ln t$, from the scatter diagram, as length of time increases, percentage of weight loss also **increases at a decreasing rate**.
(iii)	$w = -5.31 + 1.35\ln t$ The product moment correlation coefficient between ln$t$ and w is $r = 0.9828402622 = 0.983 \quad (3s.f.)$ This pmcc value of 0.983 is closer to 1 than the earlier pmcc value of 0.925, indicating stronger positive linear correlation between w and ln$t$ compared to the linear model.
(iv)	$2.4 = -5.3074 + 1.3522\ln t$ $\ln t = \dfrac{2.4 + 5.3074}{1.3522}$ $t = e^{5.6999}$ $t = 298.84 \approx 299$ (3 s.f.) This estimate is reliable since (1) The estimate is an interpolation, because w = 2.4 is within the data range of w. (2) the product moment correlation coefficient between ln t and w is $r = 0.983$ which is very close to 1, showing a very strong positive linear correlation between ln t and w.

10(v)	1 day = 24 hours $w = -5.31 + 1.35\ln(24t)$

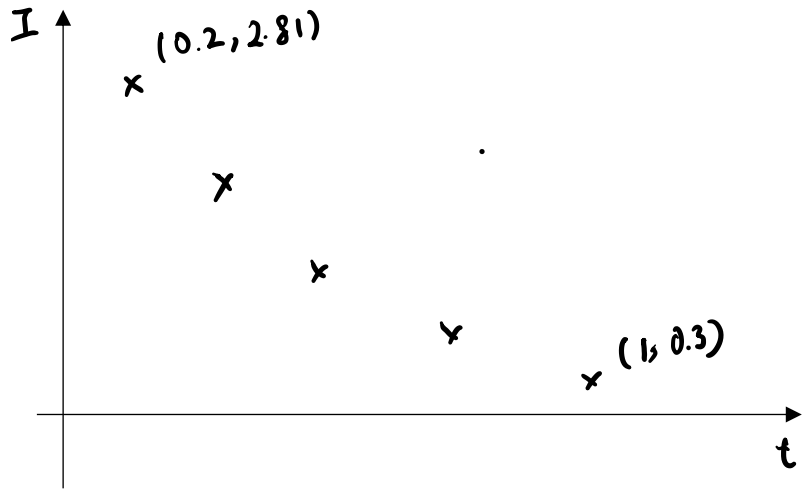
21	**(a)(i)** Given $y = 79.695x - 205.86$,
	$\bar{x} = \dfrac{4.07+3.84+4+3.31+2.48+2.92+3.1+3.57}{8}$ $= 3.41125$
	$\bar{y} = \dfrac{140+115+m+11+20+25+33+74}{8}$ $= \dfrac{418+m}{8}$
	Substitute $\bar{x}$, $\bar{y}$ into the regression line,
	$\dfrac{418+m}{8} = 79.695(3.41125) - 205.86$
	$m = 109.997$ $m = 110$
	(ii) It means that the total fuel use increases by 79.695×10^3 tonnes when the seat
	capacity increases by 100.
	(b) (i)
	(ii) The scatter diagram in (b)(i), excluding point *Q*, suggests that as *x*
	increases, *y* increases at an increasing rate so model A is not the most
	appropriate.
	(iii) (A): $r = 0.98265$
	(B): $r = 0.97787$
	As $\lvert r \rvert$ for model (A) is the closest to 1, therefore, model A is the most appropriate model.



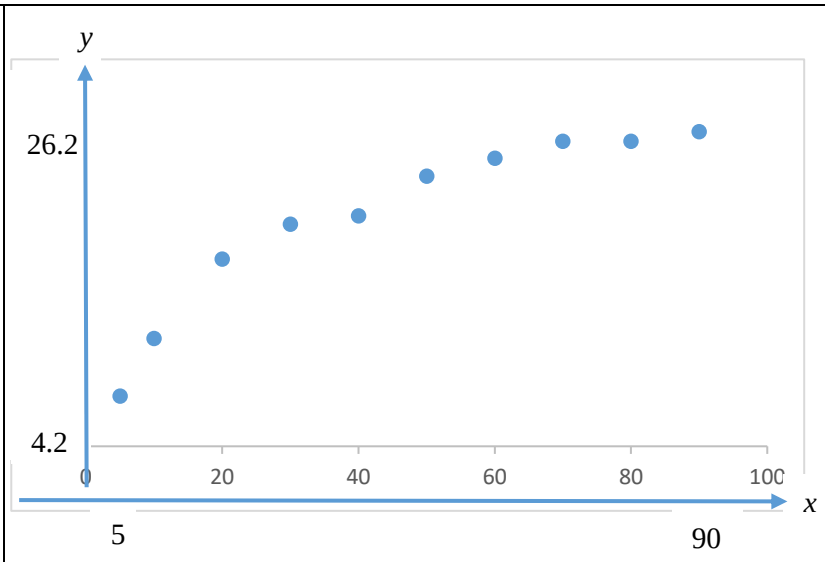
	(iv) Least squares regression line using model A:
	$y = 2.6061e^{x} - 18.429$
	When $x = 3.31$, $y = 2.6061e^{3.31} - 18.429$ $= 52.939$ So total amount of fuel used is 52 939 tonnes.
	(v) Since $r = 0.98265$ is close to 1, which suggests that there is a strong positive linear relationship between e^{x} and y, and $x = 3.31$ is between 2.48 and 4.07, so interpolation is reliable.

22	**Suggested Solution**	
(i)		
(ii)	A linear model would predict her timing to decrease at a constant rate and eventually negative, which is not possible as there is a limit to how fast a person can swim. A quadratic model would predict that her timings would have a minimum and then increase at an increasing rate, which is also not appropriate.	
(iii)	Based on the scatter diagram and the model, as x increases t decreases at a decreasing rate, therefore b is positive. a has to be positive as it represents the best possible timing that Sharron can swim in the long run.	
(iv)	From GC, $r = 0.991$ $b = 67.69$ $a = 49.50$	
(v)	Let m be the best timing Sharron has at the 8th month.	

	$\left(\overline{\frac{1}{x}}\right) = 0.33973$ We know that $\left(\overline{\frac{1}{x}}, \bar{t}\right)$ is on the regression line $t = 48.28 + 69.45\left(\frac{1}{x}\right)$. $\bar{t} = 48.28 + 69.45(0.33973) = 71.874$ $\frac{522 + m}{8} = 71.874$ $m = 52.992$ Sharron best timing is 53 seconds at the 8th month.	

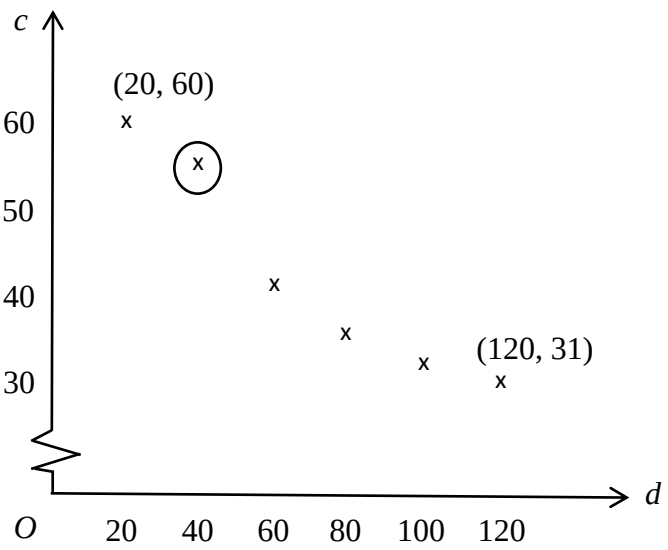
23	**(a)** **There may be a strong negative linear correlation between the amount of red wine intake and the risk of heart disease, but we cannot conclude that amount of red wine intake causes risk of heart disease to decrease, as causality cannot be inferred from correlation.**
	(b)(i) The variable *t* is the independent variable, as we are able to control, or determine, the intervals at which we measure the corresponding radiation.
	(b)(ii)  From the scatter diagram, we can see that the points lie along a curve, rather than a straight line. Hence $I = at + b$ is not a likely model.
	(b)(iii) r between I and t = -0.9565 r between ln I and t = -0.9998

	(b)(iv) $I = ae^{bt} \Rightarrow \ln I = bt + \ln a$ Equation of regression line: $\ln I = -2.7834239t + 1.6007544 \Rightarrow \ln I = -2.78t + 1.60$ $\ln a = 1.600754 \Rightarrow a = 4.96$ (3 s.f.) $b = -2.78$ (3 s.f.)
	(b)(v) $t = 0.7$, $I = 0.706$ (to 3 sig fig) The answer is reliable as r is close to -1, and $t = 0.7$ is within the data range (0.2 to 1.0) and thus the estimate is obtained via interpolation.

24 (i)	
(ii)	(a) $r = 0.925588 \approx 0.9256$ (b) $r = 0.996583 \approx 0.9966$
(iii)	From the scatter diagram, it is observed that as x increases, y increases by decreasing amounts, and the product moment correlation coefficient between $\ln x$ and y is closer to 1 than that of x and y. Hence $y = c + d\ln x$ is the better model.
(iv)	Since the fertilizer is applied at various concentrations, x (and hence $\ln x$) is the predetermined/controlled/independent variable. Thus, neither the regression line of x on y nor the regression line of $\ln x$ on y should be used. To estimate x given $y = 20.0$, the regression line of y on $\ln x$ should be used.
(v)	From GC, the equation of the regression line of y on $\ln x$ is $y = -8.4492 + 7.8126\ln x$ $\therefore y = -8.45 + 7.81\ln x$ When $y = 20.0$, $x = 38.147 \approx 38$
(vi)	Replace y by $0.1y$, $0.1y = -8.4492 + 7.8126\ln x$ $\therefore y = -84.5 + 78.1\ln x$

Solution

25 **(i)**



(ii) From the scatter diagram (after removing the outlier), as *d* increases, *c* decreases at a decreasing rate.

Also, the concentration of the herbicide will not decrease indefinitely and become a negative percentage.

Hence a linear model should not be used to model this set of data.

(iii) Using GC, $r_A = -0.92958$ while $r_B = -0.97521$.

Since the *r* value for model B is closer to -1 than model A, model B is more appropriate for modelling this set of data.

(iv) $c = ae^{bd}$

$\ln c = \ln a + bd$

From GC, $\ln c = 4.1696 - 0.0066478d$

$\ln c = 4.16 - 0.00665d$

When $d = 140$, $\ln c = 4.1696 - 0.0066478(140)$

$c = 25.5059 \approx 25.5$

(v) The estimate is unreliable because the data substituted is outside the data range [20,120] and so the linear relationship between *d* and ln *c* may not hold.

(vi) Initially, the concentration of herbicides in the soil is 64.7%.