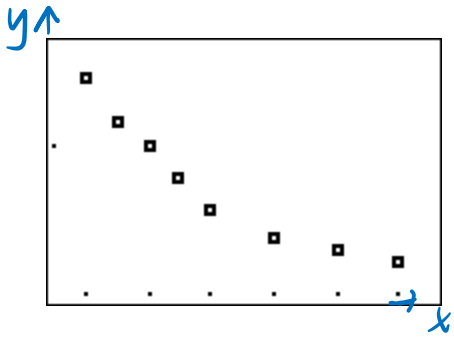
Solutions to Statistics 8 Tutorial: Correlation & Linear Regression

Additional Practice Questions

1(i)



The product moment correlation coefficient between x and y,
r is -0.943(3 s.f.).

(ii) Since $|r|$ is close to 1, it suggests a linear model is appropriate. However the scatter diagram shows the relationship is non-linear.

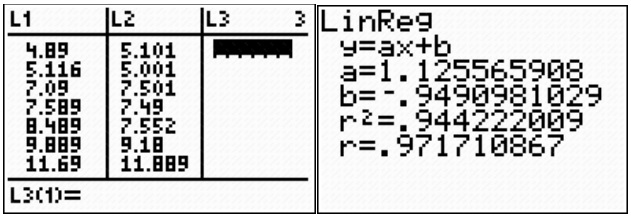
(iii) The scatter diagram shows that as x increases, the rate of decrease in y becomes smaller and y appears to approach a value. For a linear model, the rate of decrease is constant.
OR
The product moment correlation coefficient between $1/x$ and y, r_2 is 0.993 where $|r_2|$ is almost 1 as compared to $|r|$=0.943.

Therefore the model $y = a + \frac{b}{x}$ where $a > 0, b > 0$ is better.

(iv) $a = 2.16$ (3s.f.), $b = 2.33$ (3s.f.)

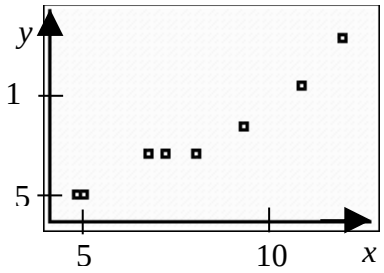
(v) When $x = 4.5, y = 2.68$ (3 s.f..)
Since $x = 4.5$ is within the data range of the x values and value of $|r_2|$ is close to 1, the estimate is reliable.

2(i)



r = 0.972
Even though r is close to 1, it does not mean that there is a cause and effect relationship between x and y.

(ii)



The points on the scatter diagram lie close to a straight line with positive gradient. This agrees with the value of r obtained in part (i).

(iii)

L1	L2	L3
4.89	5.101	1.6294
5.116	5.001	1.6096
7.09	7.501	2.015
7.589	7.49	2.0136
8.489	7.552	2.0218
9.889	9.18	2.217
11.69	11.889	2.4756

L3 ="ln(L2)"

LinReg
y=ax+b
a=.1283667212
b=.9993992111
r²=.9753341311
r=.9875900622

For the model $y = ab^x$, i.e. $\ln y = \ln a + x \ln b$, $r = 0.98759$

L2	L3	L4
5.101	1.6294	1.5872
5.001	1.6096	1.6324
7.501	2.015	1.9587
7.49	2.0136	2.0267
7.552	2.0218	2.1388
9.18	2.217	2.2914
11.889	2.4756	2.4587

L4 ="ln(L1)"

LinReg
y=ax+b
a=1.038473708
b=-.07652448
r²=.9562758273
r=.9778935664

For the model $y = ax^b$, i.e. $\ln y = \ln a + b \ln x$, $r = 0.97789$

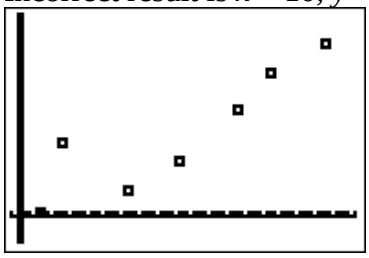
Since $|r| = 0.98759$ is closest to 1 for the model $y = ab^x$, this is the most suitable model.

(iv) $\ln y = 0.128367x + 0.9994$

$\ln a = 0.9994 \Rightarrow a = 2.72$

$\ln b = 0.128367 \Rightarrow b = 1.14$

3. Incorrect result is $x = 10$, $y = 42$.



From the scatter diagram, $x = 10$, $y = 42$ is an outlier.

$$y = -14.16830 + 1.61302x \quad \text{and} \quad x = 9.97265 + 0.59168y$$

Using GC to solve the above simultaneous equations,

$$\bar{x} = 34.8522 \quad \bar{y} = 42.04899$$

Let k be the correct value when $x = 10$,

$$k = 42.04899(7) - (1.3 + 14.8 + 30.1 + 60.8 + 81.3 + 98.6)$$

= 7.4 (to 1 decimal place)

As x is the independent variable, y on x should be used.

The estimate is reliable since $r = 0.977$ is close to 1, indicates a strong positive linear correlation between x and y and $x = 40$ is within data range.

c is the sum of least square deviation between the observed value y and the predicted value on the regression line y on x.
Using GC,
$c = 401.541488 = 402$ (to 3 sig figs) (ans)

4.

(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)

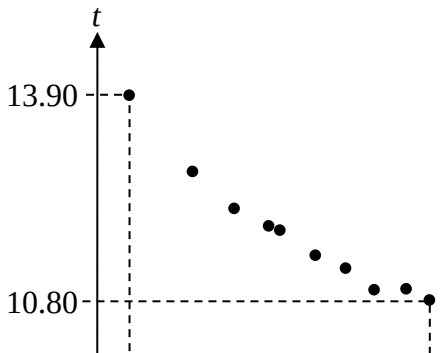
(iv) z = time available – time taken
$= 50 - x - y$
$= (50 - x) - (a + bx)$
$= (50 - x) - (18.5 + 0.564x)$
$= 31.5 - 1.564x$.

(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.

5(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 $|0.992|$ is **closer to 1** than $|-0.963| = 0.963$.

(v)(b) Regression line is

$$t = 7.8603 + 2.8616\frac{1}{x^2}$$

$$t = 7.86 + 2.86\frac{1}{x^2} \quad (3 \text{ s.f.})$$

When $t = 10$,

$$10 = 7.8603 + 2.8616\frac{1}{x^2}$$

$$x^2 = \frac{2.8616}{2.1397}$$

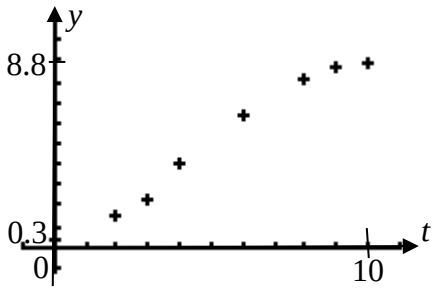
$x = 1.16$ (to 2 dec places) since $x > 0$

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.

6(i)



(ii) Using GC, $r = 0.989$.
Using regression line of y on t, $y = 0.041899 + 0.94916t$

When $t = 7$, $y = 6.6860 = 6.69\,\text{cm}^2$ (3s.f.)
Since $t = 7$ is within the data range and $r = 0.989$ is close to 1, the answer is reliable.

(iii) When $t = 80$, $y = 75.974 = 76.0\ \text{cm}^2$ (3s.f)

Note that $76.0\text{cm}^2 > 64\text{cm}^2$ (the total area of the slice of bread.) The regression line may not be suitable as it is impossible for the bread to keep growing mould.

Also, from the scatter diagram, it shows that as t increases (after 8 days, the mould starts to grow at a decreasing rate.

Hence, a linear model may not be appropriate.