6.6 Correlation & linear regression

Solutions

1. (a)

The data points on the scatter diagram clearly do not show a linear form. Thus, using either regression line would not be suitable. (ans)

(b)

Compute $\overline{x}=\frac{9.62}{6}$ and $\overline{y}=\frac{19.78+y}{6}$.

Substitute them into $y=1.069x+1.083$ since $(\overline{x},\overline{y})$ must lie on the regression line.

Thus, $y=-3.00$ (ans)

Key the data into GC and use LinReg(ax+b) L1, L2:
$r=0.392$ (ans)

Adding the point $(\overline{x},\overline{y})=(1.603,2.797)$ will not cause a change in r. (ans)

(c)

Using GC, with L1 = x, L2 = y, L3 = (x-1.108)^2 and LinReg(ax+b) L2, L3:

$(x-1.108)^2=0.662y+1.718$ (ans)

Substitute $y=3$ into the equation,

$x=3.03$ or $x=0.817$ (ans)

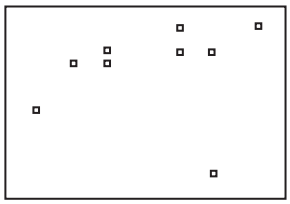
Since $r=0.985$ is close to 1 and $x=3$ is within data range, we can use this line to estimate y. (ans)

2. (a)(i)

Using GC, $r=0.147$ (ans)

As r is small, expect x and y to be not linearly correlated. (ans)

(a)(ii)



(ans)

(a)(iii)

As (9,1) Is an outlier, the interpretation in (i) should be amended. (ans)

If (9,1) is removed, the new $r=0.823$ indicated that x and y are linearly correlated. (ans)

(b)(i)

$\overline{x}=8$, $\overline{y}=68$

$$b=\frac{\Sigma xy-\frac{\Sigma x\Sigma y}{6}}{\Sigma x^2-\frac{(\Sigma x)^2}{6}}=\frac{400}{88}=\frac{50}{11}=4.54545$$

Estimate least squares regression line of y on x is

$y-\overline{y}=b(x-\overline{x})$

$y=4.54545x+31.63636$

$y=4.55x+31.6$ (ans)

(b)(ii)

When $x=8$, $y=68$

Estimated evaporation loss for a drum kept in storage for 8 weeks is 68 ml. (ans)

(b)(iii)

When the storage time is more than a year, x will be outside the range $1\le x\le 15$. Hence, we would not expect to get good estimates from the line of regression for evaporation loss. (ans)

3. (i)

Since $y=1.0031x-4.3018$, $b=1.0031$ and $\overline{y}-b\overline{x}=-4.3018$

$\overline{y}-1.0031(.752)=-4.3018$

$\overline{y}=48.611725$

From the table, $\Sigma y=338.7+k=48.611725(8)$

$k=50.2$ (ans)

(ii)

From GC, $r = 0.863$ (ans)

(iii)

Model C is suitable since it is observed from the scatter plot diagram that as x increases, there is a decreasing rate at which y increases. Since $b > 0$, Model is not suitable. (ans)

(iv)

$L1 \to x$, $L2 \to y$, $L3 \to \ln L1$

Using GC, LinReg L3, L2 yields

$y = 53.72\ln x - 163.14$

$a = -163.14, b = 53.72, r = 0.905$ (ans)

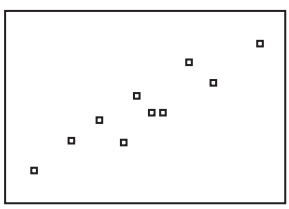
(v)

$80 = 53.72\ln x - 163.14$

$x = 92.39$ (ans)

Since $y = 80$ is out of range, value of x obtained may not be accurate. (ans)

4. (i)



(ans)

$r = 0.932$ (ans)

The scatter diagram and the value of r suggests a strong positive linear correlation between x and y. (ans)

(ii)

Regression line of y on x: $y = 0.348 + 0.273x$ (ans)

When $x = 58, y = 0.348 + 0.273(58) = 16.2$,

Moisture content is 16.2%. (ans)

(iii)

When $x = 10, y = 0.348 + 0.273(10) = 3.08$

Moisture content is 3.08%.

The estimation may be unreliable as $x = 10$ is outside the range of the given data. (ans)

(iv)

No. r will be the same because the value of r is not affected by a linear transformation on y. (ans)

5. (i)

$\overline{x} = 33.857, \overline{y} = \dfrac{142 + a}{7}$

Substituting in the given regression line,

$\dfrac{142 + a}{7} = 43.5 - 0.602 \times 33.857 = 23.11$

$a = 19.82 \approx 20$ (ans)

(ii)

Using GC,

Regression line x on y:

$x = 64.349 - 1.318y$

$x = 64.4 - 1.32y$

$r = -0.891$ (ans)

6.

$S_{xx} = \Sigma x^2 - \dfrac{(\Sigma x)^2}{n} = 448,$

$S_{yy} = \Sigma y^2 - \dfrac{(\Sigma y)^2}{n} = 148$

$\overline{x} = 13, \overline{y} = 39$

Regression equation y on x:

$y - 39 = b(x - 13)$

Substitute $x = 39$ and $y = 37$,

$b = 0.5$

$\dfrac{S_{xx}}{S_{yy}} = 0.5$

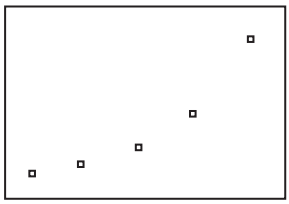
Therefore, $S_{xy} = 0.5(448) = 224$

Regression equation x on y:

$x - 13 = \frac{224}{148}(y - 39)$

$x = 1.514y - 46.03$ (ans)

7. (i)



(ans)

(ii)

Model C.

Reasons:

Shape of the points follows the shape of an exponential graph.

p increases as x increases. (ans)

(iii)

Form L3 as ln p.

LinReg (a+bx) L1, L3

$\ln p = -12.47 + 0.1978x$

$a = -12.47 = -12.5$

$b = 0.1978 = 0.198$

$r = 0.9937 = 0.994$ (ans)

(iv)

$\ln p = -12.47 + 0.1978(19) = -8.7118$

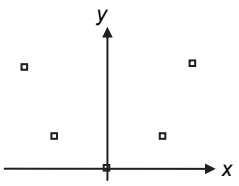
$p = 0.0001646 = 0.000165$ (ans)

Extrapolation done in calculating p when $x = 19$. There may not be a linear relationship between p and x for $x < 25$. (ans)

8.

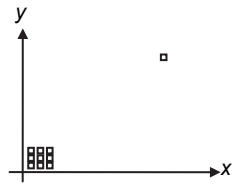
The scatter diagram gives a good overview of how the two variables are related. (ans)

(i)



(ans)

(ii)



(ans)

9. (i)

$r = 0.98$

High positive linear correlation between X and Y. (ans)

(ii)

r remains unchanged. r is a measure of the degree of scatter and it is unchanged when there is a change of scales. (ans)

10.

Regression line of x on y:

$y = -\frac{9}{8}x + 5.5$

$x = -\frac{8}{9}y + \frac{44}{9}$

$r^2 = \frac{16}{25}$

Since $(\overline{x}, \overline{y}) = (4,1)$ lies on the regression line of y on x, required regression line:

$y - 1 = -\frac{18}{25}(x - 4)$

$y = -\frac{18}{25}x + \frac{97}{25}$ (ans)

11.

The point is $c = 51, l = 34$ (ans)

(i)

A suitable regression line is $l = 19.722 + 0.86981c$ (ans)

(i)(a)

When $c = 34.5, l = 49.7$ (ans)

(i)(b)

When $c = 45, l = 58.9$ (ans)

(ii)

The extrapolation has gone beyond the interval of the provided data. Relationship between the variables may not follow the same pattern as the given set of data. (ans)

12.

$\Sigma t = 80, \bar{t} = 10, n = 8$

$\Sigma m = 503 + p$

$\bar{m} = 0.94\bar{t} + 62.6$

$\frac{503+p}{8} = 0.94(10) + 62.6$

$p = 73$ (ans)

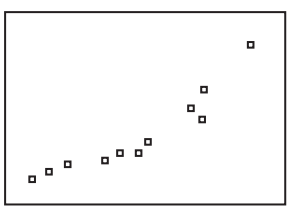
From GC, regression line of t on m,

$t = m - 62$ (ans)

Using the regression line of m on t,

$\frac{\Delta m}{\Delta t} = 0.94$

$\Delta m = 0.94\Delta t = 0.94(2) = 1.88$ (ans)



(ans)

Since the points for model A lie close to an increasing curve with increasing positive gradient, model A is most appropriate. (ans)

From GC, $m = 64.1 + 0.0799t^2$

$a = 64.1$

$b = 0.0799$

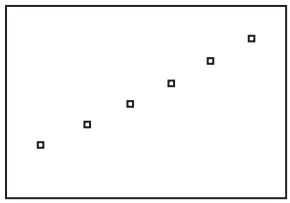
$r = 0.919$ (ans)

13. (i)

$y = AB^x$

$\log y = \log A + x \log B$

The scatter diagram is plotted with $Y = \log y$ again x.



(ans)

From GC, the points on the scatter diagram lie close to a straight line, so the relation is a reasonable model. (ans)

(ii)

From GC,

LinReg

$y = ax + b$

$a = 0.3271848077$

$b = 0.7809873679$

$r^2 = 0.9895655134$

$r = 0.9947690754$

The line of regression of Y on x is

$Y = 0.781 + 0.327x$ (ans)

Since $\log A = 0.781, A = 6.04$ (ans)

(iii)

From GC, $r = 0.995$ which is very close to +1, it supports the reasonability of the model. (ans)

(iv)

Lines must pass through $(\bar{x}, \bar{Y})$.

$\bar{Y} = 0.781 + 0.327\bar{x}$

$\bar{x} = 2.00\bar{Y} - 0.297$

Solving both equations,

$\bar{Y} = 1.98$ (shown) (ans)

$\sum Y = 1.97653 \times 7 = 13.83574$

$\log y_0$

$= 13.83574 - (\log 11 + \log 29 + \log 68 + \log 138 + \log 215 + \log 560)$

$= 2.27894$

$y_0 = 10^{2.27894} = 190.08 \approx 190$ (ans)

14. (i)

$\overline{x} = 15.75, \overline{y} = \dfrac{23.3 + k}{8}$ (ans)

(ii)

$y = 0.197x + 0.184$

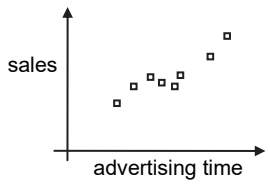
$\overline{y} = 0.197\overline{x} + 0.184$

$\dfrac{23.3 + k}{8} = 0.197(15.74) + 0.18$

$k = 2.994$

$r = 0.9287$ (ans)

(iii)



(ans)

(iv)

r remains unchanged because *r* is a measure of the degree of scatter and this is unchanged by a change of scaling. (ans)

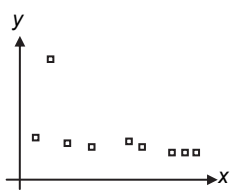
(v)

It is reasonable because $y = 3.4$ lies within the range of the given data and $r = 0.9287 \approx 1$, the regression line *y* on *x* is almost identical to *x* on *y*. (ans)

15. (i)

$r = -0.542$ (ans)

(ii)



(ans)

The scatter diagram indicates a strong negative linear correlation between *x* and *y*. This is not consistent with the value of *r* in (i) due to the presence of the outlier. (ans)

(iii)

The data pairs which should be removed are (5,2) and (11,0)

For the revised data, $r = -0.916$

$y = 0.83992 - 0.021267x$

$y = 0.840 - 0.0213x$ (ans)

(iv)

Substitute a large value of *x* $(x > 39)$ and obtain a negative value of *y*, which is impossible.

OR $0.83992 - 0.021267x < 0 \Rightarrow x > 39.494$

For $x \geq 40, y < 0$ which is impossible. (ans)

(v)

$(a,b) = (\overline{x},\overline{y}) = (13.75, 0.5475)$

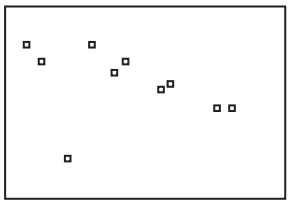
OR

Any pair (a,b) that lies on the regression line of *y* on *x*.

a	*b*
4	0.755
5	0.734
6	0.712
7	0.691
8	0.670
9	0.649
10	0.627
11	0.606
12	0.585
13	0.563
14	0.542
15	0.521
16	0.500
17	0.478
18	0.457
19	0.436
20	0.415
21	0.393
22	0.372

(ans)

16. (i)



There is a negative correlation between the 2 data sets with one clear outlier, when $A=54, B=58$. (ans)

(ii)

The doctor should ignore the pair of data points when $A=54$ and $B=58$.

The new linear product moment correlation coefficient, $r=-0.928$

(iii)(a)

Diastolic blood pressure. (ans)

(iii)(b)

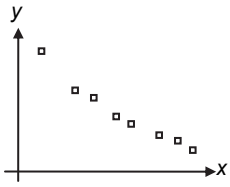
Regression of x on y: $x=158.36-1.163y$

When $y=80, x=65$. (ans)

(iv)

$x=90$ is outside the range of the data given for x. (ans)

17. (i)



(ans)

(ii)

$y=73.3-18.4\ln x$ (ans)

(iii)

Draw the regression line in (ii) in the scatter diagram in (i). If the data points are close to the straight line, data firs the calculated equation. (ans)

(iv)

$a=e^{73.3}=6.89\times10^{31}$ (ans)

$b=-18.4$ (ans)

(v)

1.84% (ans)

Since $x=49$ is out of the given data range of $1\le x\le 40$, prediction is unreliable. (ans)

(vi)

For large values of x, the model gives $y<0$. So the model is not valid for large values of x. (ans)