Answers

Page 5

1. Time spent brushing teeth per day (x). Number of cavities (y).

2. Number of hours spent studying for an exam (x). Percentage result (y).

3. Calorie intake per day (x). Weight (y).

4. Either in this case. Dependent on question being asked.

Page 6

5. Speed (x). Petrol consumption (y).

6. Money spent on tobacco products (x). Alcoholic beverage spending (y). Use of the word compared.

7. Maximum speed limit in km/h (x). Petrol consumption litres/100 km (y).

8. Wattage of heater (x). Effective heating area (y).

9. Countries (x). Recycling rates for paper and glass (y).

10. Area in m^2 (x). Selling price (y).

11. GDP (x). Unemployment rate (y).

12. Age of a person (x). Resting heart rate (bpm) (y).

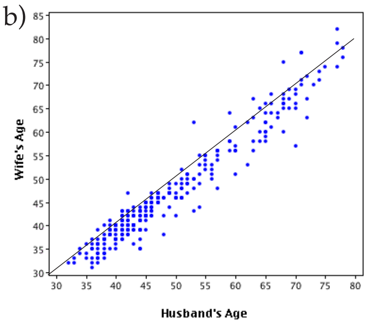
13. Cigarettes smoked per day (x). Age at death (y).

14. Thickness of insulation (mm) (x). Heat loss (%) (y).

Page 8

15. a) Wife's age.

b)



Page 8 Q15 cont...

c) Husband's and wife's ages are the same.

d) Wives older than their husbands.

e) More points below the line.

Page 9 Q15 cont...

f) Yes as more points fall below the line y = x.

g) There might appear to be some truth to this as the points are more densely clustered and there appears to be less variation at younger ages.

h) Husband's versus wife's age shows a strong linear positive association. The majority of points fall below the line y = x, indicating that husbands are generally older than their wives. The points are tightly clustered and the variation at each age is within approximately 10 years indicating that wives are up to ten years younger than their husbands. There are one or two 'unusual' observations, e.g. the point (54, 63) that goes against the 'trend' i.e. where a wife is 9 years older than her husband but well within acceptable norms. The density of points is less as age increases likely due to separation and death.

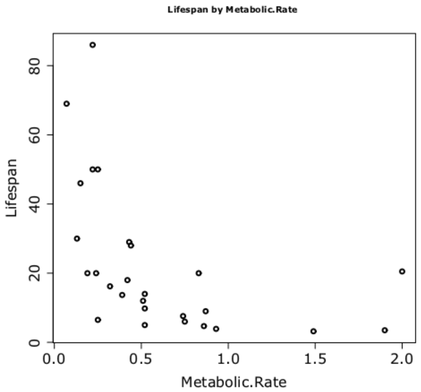
Page 9

16. a) Metabolic rate (explanatory) and lifespan (response).

More sense to investigate how metabolic rate predicts lifespan not the other way round.

Page 9 Q16 cont...

b)

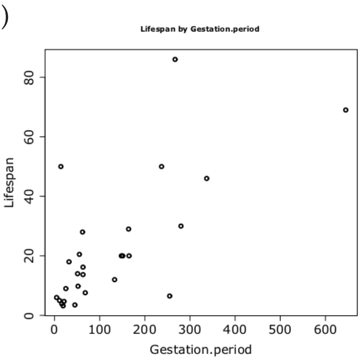


c) Little brown bat. The trend is that the higher your metabolic rate the lower your lifespan. You would expect the little brown bat to have a lifespan of approximately 3 to 4 years.

Page 10 Q16 cont...

d) There appears to be a negative non-linear association between the two variables. Mammals with increased metabolic rates have a shorter lifespan. Most points fall within the metabolic range 0 to 1.0 and within this range lifespan varies from 4 to 86. One point appears to go against the trend (the little brown bat) with a metabolic rate of 2.0 but a lifepan of 20.5 years.

e)



f) Man has a long lifespan given the gestation period compared to other mammals and the Asian elephant even though it has a long gestation period its lifespan is not as great as could be expected.

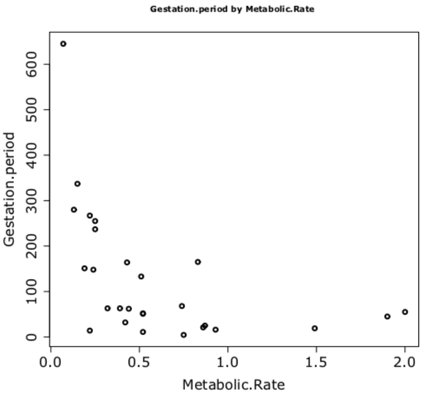
Page 10 Q16 cont...

g) There appears to be a moderate positive linear association between the two variables. Mammals with a longer gestation period have a longer lifespan. Two unusual observations are the Echnida which has a gestation period of only 14 days but a long lifespan of 50 years and man with a gestation period of 267 days and a very long lifespan of 86 years. Most points fall within the gestation range of 0 to <200 days and lifespan of 0 to 20 years.

h) Metabolic rate (explanatory) and gestation period (response).

Investigating how metabolic rate predicts gestation period not the other way round.

i)



Page 11 Q16 cont...

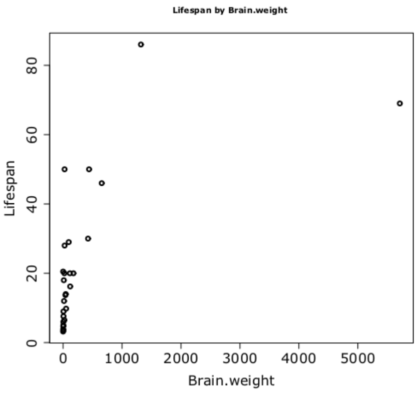
j) They have a low gestation period.

k) Not an obvious linear one although greater metabolic rate does result in a lower gestation period. More likely a negative non-linear relationship.

Page 11 Q16 cont...

l) There appears to be weak negative non-linear association between the two variables. Mammals with a greater metabolic rate have a shorter gestation period. Most points are clustered in the metabolic range 0 to < 1.0 and gestation period of 0 to < 300 days. Extreme points are that of the Asian elephant and little brown bat although they would still fit a non-linear model.

m)



n) Gestation period is a reasonable predictor of lifespan although man and the Echnida go against the trend. A linear model could be suitable for these two variables.

Metabolic rate versus lifespan would be better suited to a non-linear model. The little brown bat goes against the trend. A suitable non-linear model would probably be a good predictor of lifespan.

With brain weight versus lifespan a linear model is the best model, but the Asian elephant goes against the trend.

Page 11 Q16 n) cont...

Overall brain weight looks like the better predictor from a linear perspective if you exclude the Asian elephant from the dataset, but metabolic rate from a non-linear perspective. Gestation period is a reasonable linear predictor when including all data elements.

Page 16

17.

x	y	xy	x^2	y^2
5	40.1	200.5	25	1608.01
15	32.2	483	225	1036.84
18	35.1	631.8	324	1232.01
20	34.3	686	400	1176.49
25	23.6	590	625	556.96
30	26.9	807	900	723.61
38	24.1	915.8	1444	580.81
50	20.0	1000	2500	400
201	236.3	5314.1	6443	7314.73

$r = {}^{-}0.912$

18.

x	y	xy	x^2	y^2
1	3.1	3.1	1	9.61
2	4.4	8.8	4	19.36
3	7.2	21.6	9	51.84
4	6.6	26.4	16	43.56
5	15	75	25	225
6	14.1	84.6	36	198.81
9	20.3	182.7	81	412.09
10	25.3	253	100	640.09
40	96.0	655.2	272	1600.36

$r = 0.975$

19. $r = 0.920$

Page 17 (Answers may vary)

20. 1. Constant difference.

21. Close to ${}^{-}1$. The older a car the less its value.

22. Somewhat negative. Less well educated people smoke.

23. Close to 1. People with big feet are generally taller.

24. Somewhat negative. More cigarettes smoked the lower the level of fitness.

25. ${}^{-}1$. The more you spend the less you save.

26. Close to 1. Tall men often choose tall wives.

27. Somewhat positive. More land area greater price.

Page 17 cont...

28. Close to $^{-}1$. More insulation less heat loss.

29. 0. No relationship.

Page 18

30. 0.87

31. $^{-}0.45$

32. $^{-}0.93$

33. 0.05

Page 19

34. 0.50

35. $^{-}0.72$

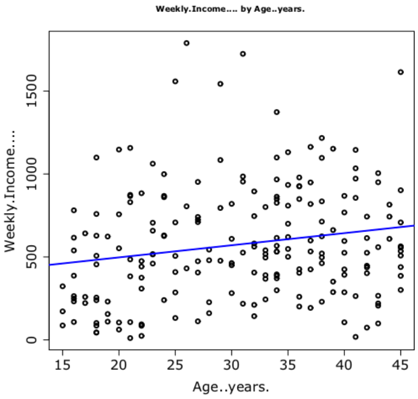
36. $^{-}0.99$

37. 0.25

Page 20

38. a) $529

b) and c)

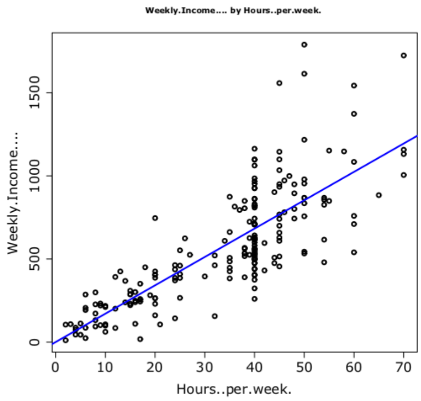


r = 0.19

d) Little association between age and weekly income. Range of income is evenly spread throughout each age. There is a gradual positive tendency (r positive) for income to increase slightly over time perhaps indicating that as some people become more experienced in their job or career they earn more. Dataset is only limited to ages 15 to 45 so we are not getting the full picture to retirement (65). Scatter plot does not support the statement.

Page 20 Q38 cont...

e) and f)

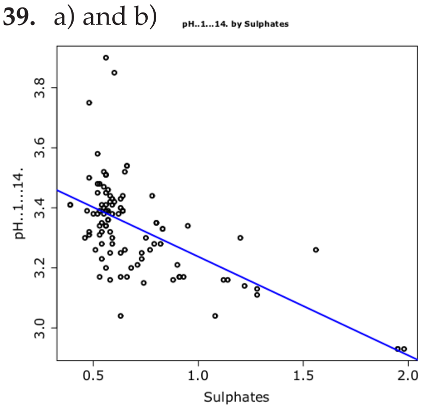


r = 0.80

g) Strong positive association between hours worked and weekly income. As most people get paid based on an hourly rate this is expected. Those working long hours and earning less are likely people on a salary e.g. teachers etc. In the age group 50 to 60 a cluster of four to five individuals are earning significantly more perhaps reflecting a significantly higher hourly rate or salary. Scatter plot and linear correlation coefficient supports the statement.

Page 21

39. a) and b)



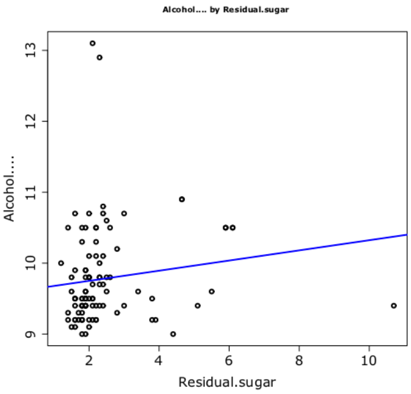
r = $^{-}0.62$

c) For the majority of wines the statement appears to be true. Wines with an increased sulphate concentration generally have a lower pH level hence the negative correlation coefficient.

Page 21 Q39 c) cont...

There are three wines that stand out and go against the trend but the majority follow the negative trend. Most wines in the dataset are clustered in the 3.1 to 3.5 pH range and have a level of sulphates of between 0 and 1.0, probably because they all come from the same region in Portugal where conditions (soil and weather are likely to be similar). The relationship between the variables can be described as moderate.

d)



r = 0.14

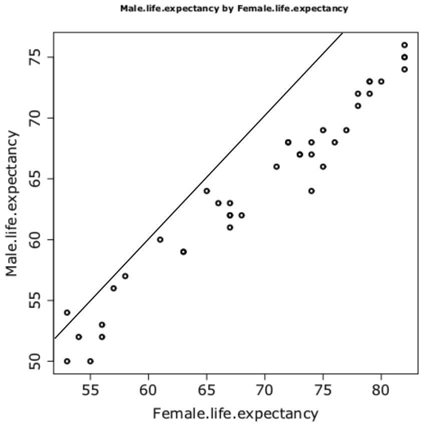
Scatter plot does not support this. No correlation between residual sugar and alcohol content of a wine. Correlation coefficient is 0.14. The majority of wines (cluster) have a residual sugar level of between 0 and 3 and their alcohol content ranges from 9 to 11%. There are two unusual wines with a low residual sugar level and a high alcohol content. The trend line if anything indicates that the more residual sugar the greater the alcohol content but this is not reflected in the dataset.

Page 21 Q39 cont...

e) No good linear relationship between one variable and wine experts' evaluation. Unlikely to be as it is a combination of factors that make a good wine not just one.

Page 22

40. a) and b)



c) Women have a greater life expectancy than males because all points fall below the line y = x except for one country.

d) Bangladesh.

e) No. In the countries with higher life expectancy the points deviate further from the line y = x indicating that the gap between male and female life expectancy in these countries is greater. Women live longer by more than five years in these countries.

f) 0.98

g) Definite correlation between male and female life expectancy across nearly all countries. Females live longer than males. The gap between male and female life expectancy is greater in those countries with a higher overall life expectancy. Linear correlation coefficient is 0.98 indicating a strong positive association between male and female

Page 22 Q40 g) cont...

life expectancy. One observation goes against the trend in Bangladesh where males outlive females.

h) The wealthier a country the better its health system and the greater the life expectancy. An indicator of a 'wealthy' country is likely to be the number of TVs per person.

i) More likley to be the low doctor patient ratio. Scatter plot gives a linear correlation coefficient of $^{-}0.67$ (the more people per doctor the lower the life expectancy). TVs have become an 'essential' item and more readily available even in poorer countries so not as good an indicator ($^{-}0.52$). In poorer countries the number of people per doctor tends to be high and life expectancy is directly related to a good health system and access to affordable health care.

Page 23 Q40 cont...

j) No difference between the sexes. TV's per person are not as good an indicator as people per doctor. Correlation coefficient for doctors was $^{-}0.67$ for all, $^{-}0.69$ for males and $^{-}0.64$ for females. For TV's per person the correlation coefficients were $^{-}0.52$ for all, $^{-}0.51$ for males and $^{-}0.52$ for females. Results for the different sexes are consistent with those for both sexes combined.

Page 23

41. a) Moderate to strong negative relationship. As it gets warmer demand for heating is less and so electricity usage decreases.

Page 23 Q41 b) cont...

b) In summer when it gets hot, air conditioning is used so the demand for electricity increases.

Page 30

42.

	M (x)	T (y)	xy	x^2
	3	26.8	80.4	9
	4	25.4	101.6	16
	3	30.2	90.6	9
	7	20.5	143.5	49
	6	21.9	131.4	36
	8	19.8	158.4	64
	11	15.5	170.5	121
	15	14.9	223.5	225
	2	30.5	61	4
	5	18.5	92.5	25
Σ	64	224	1253.4	558

T = 30.17 – 1.214M
T = 15.6 minutes

43. y = $^{-}1.560$ + 4.127x

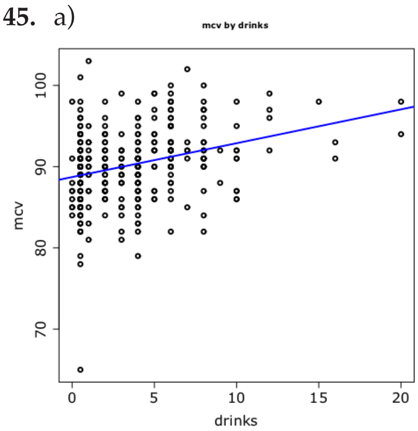
y = 6.03 m

44. Bod.wght = 0.91757 x Brn.wght – 105.82

Bod.wght = 630 kg (2 sf)

Page 31

45. a)



Wide variation of MCV for each drink (240 ml) consumed. Most points clustered from 0 to 10 drinks. Very few points for greater than 10 drinks. No unusual observations. Scatter plot indicates different MCV for different people drinking the same number of drinks.

Page 31 Q45 cont...

b) Linear correlation coefficient r = 0.31 reflects a weak positive linear relationship between number of drinks and MCV.

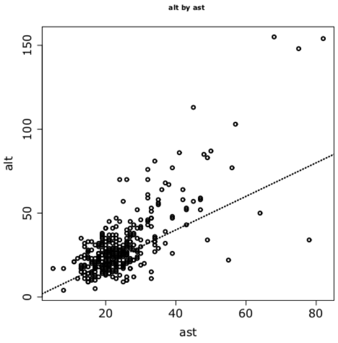
c) mcv = 0.4167 * drinks + 88.72

mcv = 90.8 for (drinks = 5)

d) Wide variation for each number of drinks and linear correlation weak so prediction estimate weak.

e) Drawing two equidistant parallel lines one each side of the trend line that encompass most of the points gives us a prediction range of 82 to 100.

f)

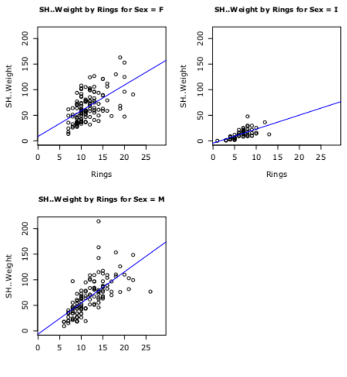


Significant number of males from the dataset with AST higher than ALT at the lower levels but from (AST) 40 onwards very few. All should be tested further to check for liver disease especially those with AST > 40.

Page 32

46. a) Male abalone have a moderate positive linear relationship (r = 0.68) between rings and shucked weight as larger number of rings generally gives a larger shucked weight.

b)



Page 32 Q46 b) cont...

Moderate positive association between number of rings and shucked weight of abalone. r = 0.55 for F, r = 0.65 for I and r = 0.68 for Males.

Infant scatter plot has majority of points clustered below 10 rings because of age. These infant abalone could not be identified as M or F until older.

Female scatter plot shows a large tight cluster between 10 and 15 rings (age 11.5 to 16.5 years) with greater variation in shucked weight.

Male scatter plot shows a greater variation in age. One unusual observation is where shucked weight is 214 g.

Page 33 Q46 cont...

c) Male SH.W = 6.08 * Rings – 6.13

Fem. SH.W = 5.016 * Rings + 8.742

SH.Weight (M) = 85 g (50 to 120 g)

SH.Weight (F) = 84 g (40 to 120 g)

Drawing two equidistant parallel lines one each side of the trend line to encompass most points.

d) Yes it does. Correlation (M) = 0.72. Correlation (F) same.

e) From scatter plots approximately 6 to 7 rings although there is some overlap.

Male scatter plot shows males from 6+ rings and females from 7+ rings while infants are still designated as such up to 8 or 9 rings.

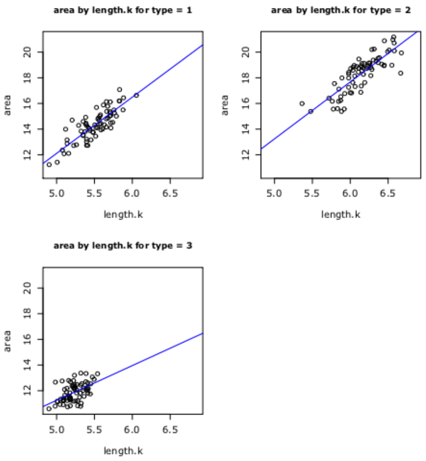
f) Median shucked weight for M is 65 g compared to 61.6 g for F, but mean shucked weight is similar for both and the mean gives a better measure of total,

Page 33 Q46 f) cont...

so shucked weight of male and female abalones looks much the same.

Page 34

47. a)



Kama and Rosa kernels show a strong positive correlation (0.83) between length and area and Canadian kernel a moderate positive correlation of 0.52. All three seeds are tightly clustered within their length range 5 – 5.8 mm for Kama, 5 – 5.5 mm for Canadian and 5.8 – 6.6 mm for Rosa. No unusual observations in any of the three scatter plots.

b) area = 4.384 * length.k – 9.811

area = 15.4 mm

c) 5 to 5.8 mm as this is the length range in which most of the Kama kernels lie.

d) area = 4.436 * length.k – 8.937

area = 19.9 mm

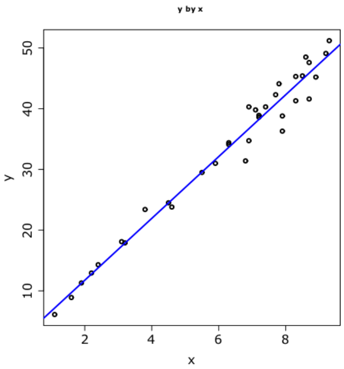
e) area = 6.239 * length.k – 20.27

area = 15.6 and 20.3 mm

Correlation coefficient for all kernels is 0.95 which is a strong positive correlation. Results compare favourably with those for individual kernels as the gradients for Kama and Rosa kernel regression lines were similar.

Page 37

48. a)



Correlation coefficient is 0.99 which indicates a strong positive association between x and y.

Majority of points are clustered from 7 to 10. No unusual observations.

Regression line is:
y = 5.091 * x + 1.566
y = 27.0 when x = 5.0

b) (4.3, 41.2)

c) Reduces correlation coefficient to 0.96.

Regression line is:
y = 4.942 * x + 3.017
Slope similar.

d) Slope is similar, little effect on correlation coefficient.
Point is an outlier as it goes against the general trend but not that influential.

e) May consider (15, 94.2).

f) Correlation 0.98 which is similar.
Regression line:
y = 5.555 * x – 0.8797

g) Influential but may not be an outlier as it follows the general trend.

Page 38 Q48 cont...

h) (15.6, 9.4)

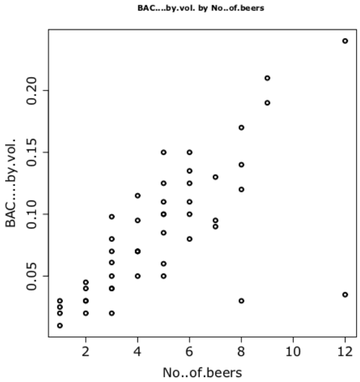
i) Correlation 0.65
Regression line:
y = 2.939 * x + 13.42

Page 38 Q48 cont...

j) Both influential and an outlier as it affects the correlation coefficient and the gradient of the regression line significantly.

Page 38

49. a)



(12, 0.035) and (8, 0.03)

Not following the same trend as the other points and also distant from the other points.

b) Included correlation coefficient is 0.7 and regression line:
BAC = 0.01397 * Beers
+ 0.0177

Removing points results in correlation coefficient of 0.89 and regression line of :
BAC = 0.01939 * Beers
– 0.001869

Included points are influential outliers.

Page 39 Q49 cont...

c) From scatter plots BAC higher for females drinking the same number of drinks. Gradient of female scatter plot is roughly double that of the males. (0.018 compared to 0.008).

Using regression lines:
Females (r = 0.82)
BAC = 0.01783 * Beers
+ 0.009568
Males (r = 0.54)
BAC = 0.008413 * Beers
+ 0.03129

Page 39 Q49 c) cont...

Removing outliers.
Using regression lines:

Females (r = 0.93)
BAC = 0.02002 * Beers
+ 0.003833
Males (r = 0.85)
BAC = 0.01647 * Beers
+ 0.001528
Regression lines now similar for both males and females to predict BAC.

d) Based on the dataset with outliers removed.

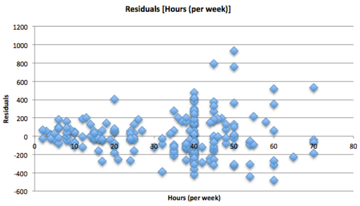
Predicting BAC based on the number of drinks consumed is similar for all weight groups up to 125 kg, i.e. 47 – 60, 60 – 79, 79 – 97 and 97 – 125 essentially have the same regression line (BAC = 0.02 * Beers).

Correlation coefficients of these four weight groups (0.96, 0.92, 0.85 and 0.72) indicate a moderate / strong positive relationship between drinks consumed and BAC.

e) Based on the previous results there is little difference in predicting BAC based on the number of beers consumed, by gender or weight.

Page 44

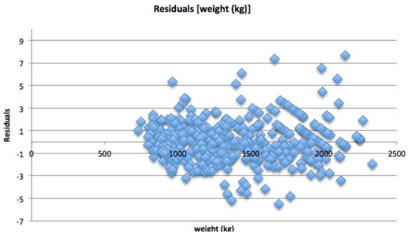
50.



Residuals are randomly scattered. They tend to increase in magnitude for more hours worked. This indicates there is a greater variation in weekly income for more hours worked. Overall a linear model appears to be appropriate for hours worked per week versus weekly income.

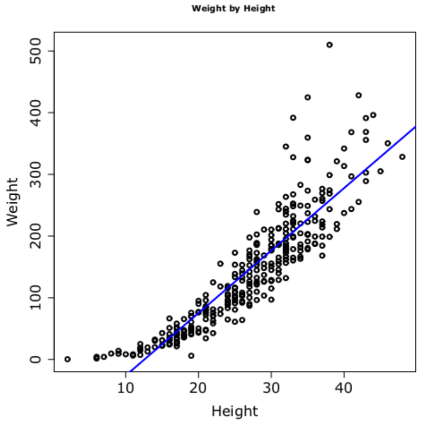
Page 44 cont...

51.



The residuals are randomly scattered with an even split above and below the horizontal axis. A linear model would appear to be appropriate for weight versus economy rate of vehicles.

52. a)

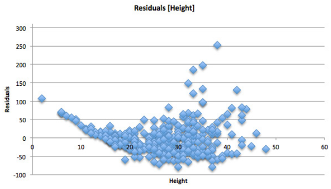


Linear correlation coefficient of 0.89 indicating a strong positive association. Variation of weight increases as height increases. Points are tightly packed between 12 and 35 mm. At extremes points deviate from the linear trend.

Page 45 Q52 cont...

b) Exponential.

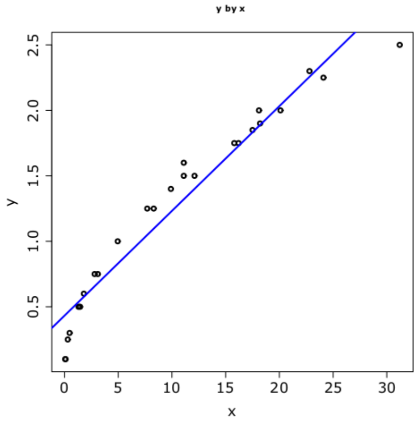
c)



Approximate U pattern of residuals indicates that a non-linear model would be more suitable. Suggest exponential (growth curve model) as it 'visually' fits the data.

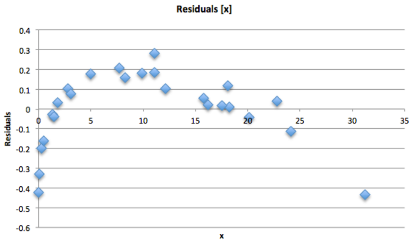
Page 45

53. a)



Linear correlation of 0.97 indicates a very strong positive relationship. Linear model does not 'fit' the points at the start and appears to be moving away from the y values as x becomes larger. For x values in the range 2 to 25 linear trend is an excellent 'fit'.

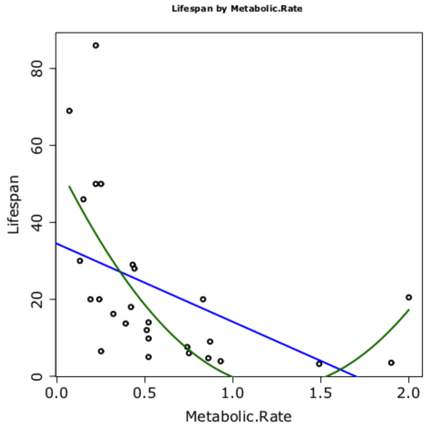
b)



Inverted U pattern of residuals indicates that a non-linear model would be more suitable. Suggest power function as it 'visually' fits the data.

Page 50

54. a) and b) (iNZight)



$L = {}^{-}20.31M + 34.43$

$L = 36.69M^2 - 92.54M + 55.61$

Page 50 Q54 cont...

c) Trendline predictions: 26.3, 10.1 and $^{-}6.2$

Trend curve predictions: 24.5, $^{-}2.6$, 17.3

d) Regression line is not a good fit and lifespan predictions are too high for metabolic rates between 0.1 and 1.4. Quadratic does not predict well for 0 to 0.25 metabolic range and between 1.0 and 1.5 predicts a negative lifespan.

e) $y = 7.6049x^{-0.822}$

f) Power trend predictions: 16.2, 6.5, 4.3

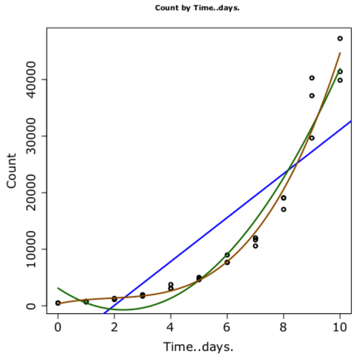
g) Power trend, because it deals with asymptotic behaviour whereas linear and quadratic don't. As metabolic rate increases lifespan decreases at a declining rate.

Page 51.

55. a) 500

b) As days increase bacteria count increases at an increasing rate so not a linear model.

c) (iNZight)



Best model would be cubic.

$C = 76.84T^3 - 431.5T^2 + 1067T + 355.7$

Predictions:
1750 and 44 700

d) Cubic fits plotted points very well but as days increase (> 10) model will not be as reliable.

Page 51 Q55 cont...

e) Exponential curve:
$C = 487.25e^{0.4587T}$

f) Predictions:
1930 and 47 800.
Exponential curve fits very well and will deal with the bacteria count increasing at an increasing rate after 10 days which the cubic model would not.

Pages 54 – 56

Practice Internal Assessment Task

Achievement

The student has investigated bivariate measurement data.

They have shown evidence of using each component of the statistical enquiry cycle.

The student has:

Posed an appropriate relationship question, selected and used appropriate display(s). For example, produced appropriate scatter graph(s) with axes labelled. The explanatory and response variables are clear.

Identified features in the data and described the nature and strength of the relationship. For example, stated from a visual inspection that the trend is linear rather than non-linear. They have described the strength with reference to visual aspects of scatter about the regression line and the nature, in context, by stating, for example, that as one variable increases the other also tends to increase. Other features and unusual points have been identified.

Found an appropriate model. For example, fitted an appropriate regression model to the data.

Made a prediction. For example, used the model to make a prediction for the response variable. The prediction is sensible with respect to the context and uses units and sensible rounding.

Communicated findings in a conclusion. For example, clearly communicated each component of the cycle. They have made a conclusion which is consistent with their question.

Merit

The student investigated bivariate measurement data, with justification.

They have shown evidence, linked components of the statistical enquiry cycle to the context and made supporting statements by referring to evidence such as statistics, data values, trends, or features of visual displays.

The student has:

Posed an appropriate relationship question.

Selected and used appropriate display(s). For example, produced appropriate scatter plot(s) with axes labelled. The explanatory and response variables are clear.

Identified features in the data and described the nature and strength of the relationship. For example, stated from a visual inspection, with evidence, that the trend is linear rather than non-linear. They have described the strength with references to visual aspects of the scatter about the regression line. They have described, in context, the nature, by stating, for example, as one variable increases the other also tends to increase and justified this with reference to visual aspects of the display(s). They have included consideration of possible contextual reasons for the features. They may have acknowledged that they have found only a statistical relationship between the variables but that this does not imply causation.

Found an appropriate model. For example, fitted an appropriate regression model to the data. The appropriateness of the model is justified by discussion of fit throughout the range of x-values in the data or the number of data points.

Made a prediction. For example, used the model to make a prediction for the response variable. The prediction is interpreted in context with the use of units and sensible rounding. It is justified with discussion on how precise it might be and has supported comments with references to statistical evidence from the analysis.

Communicated findings in a conclusion. For example, clearly communicated each component of the cycle. There is contextual support for the conclusion which is consistent with their question. They may have considered, giving contextual reasons, the possibility of investigating their chosen relationship separately for males and females.

Excellence

The student has investigated bivariate measurement data, with statistical insight.

They have shown evidence of integrating statistical and contextual knowledge throughout the process. They may have reflected on the process, considered other relevant variables, or evaluated the adequacy of any models.

The student has:

Posed an appropriate relationship question.

Selected and used appropriate display(s). For example, produced appropriate scatter graph(s) with axes labelled. The explanatory and response variables are clear. Identified features in the data and described the nature and strength of the relationship. For example, stated, with evidence, that the trend is linear/ non-linear. They have described the strength with references to visual aspects of the scatter about the regression line. They have described, in context, the nature by stating, for example, that as one variable increases the other also tends to increase and justified this with reference to visual aspects of the display(s). Other features of the data have been discussed with supporting comments. They have reflected on the strength and nature of the relationship with contextual comments. They may have acknowledged that they have found only a statistical relationship between the variables but that this does not imply causation. They may have acknowledged that other factors (named) would impact on the variable. They have reflected on features by discussing their relevance to a wider population.

Found an appropriate model. For example, fitted an appropriate regression model to the data. The appropriateness of the model is justified by discussion of fit throughout the range of x-values in the data. Consideration has been given to aspects such as the number of data points. They may have considered improvements to the model by considering other models that might be more appropriate. They may have extended the investigation by developing models with data that has been separated into relevant subsets.

Made a prediction. For example, used the model to make a prediction for the response variable. The prediction is interpreted in context with the use of units and sensible rounding. It is justified with discussion on how precise it might be and is supported by comments with references to statistical evidence from the analysis. They may have justified the choice of variable to use for predictions by giving reasons for using the selected one rather than others. They have reflected on predictions by discussing their relevance to a wider population. The prediction may have been made using alternative models and the accuracy of these has been discussed in the context of the investigation.

Communicated findings in a conclusion. For example, clearly communicated each component of the cycle. They have extended their initial investigation (for example, investigating males and females separately), giving contextual reasons to justify the extension and discussing the results of the extended investigation with respect to their question.