Year 13 Mathematics

IAS 3.9

Bivariate Data

Robert Lakeland & Carl Nugent

Contents

- Achievement Standard 2
- Bivariate Data 3
- Scatter Plots 4
- Relationships 7
- Correlation Coefficient 12
- Regression Analysis 24
- Outliers 35
- Residuals 40
- Residual Plots 41
- Causation 46
- Non-Linear Regression 47
- Bivariate Investigation 52
- Practice Internal Assessment 54
- Answers 57

MuLake Ltd
Innovative Publisher of Mathematics Texts

NCEA 3 Internal Achievement Standard 3.9 – Bivariate Data

This achievement standard involves students investigating bivariate measurement data.

Achievement	Achievement with Merit	Achievement with Excellence
• Investigate bivariate measurement data.	• Investigate bivariate measurement data, with justification	• Investigate bivariate measurement data, with statistical insight.

- This achievement standard is derived from Level 8 of The New Zealand Curriculum and is related to the achievement objectives
  Carry out investigations of phenomena, using the statistical enquiry cycle
  - using existing data sets
  - finding, using, and assessing appropriate models (including linear regression for bivariate data), seeking explanations, and making predictions
  - using informed contextual knowledge and statistical inference
  - communicating findings and evaluating all stages of the cycle in the Statistics strand of the Mathematics and Statistics Learning Area.

- Investigate bivariate measurement data involves showing evidence of using each component of the statistical enquiry cycle.

- Investigate bivariate measurement data, with justification involves linking components of the statistical enquiry cycle to the context, and referring to evidence such as statistics, data values, trends, or features of visual displays in support of statements made.

- Investigate bivariate measurement data, with statistical insight involves integrating statistical and contextual knowledge throughout the investigation process, and may include reflecting about the process; considering other relevant variables; evaluating the adequacy of any models, or showing a deeper understanding of the models.

- Using the statistical enquiry cycle to investigate bivariate measurement data involves:
  - posing an appropriate relationship question using a given multivariate data set
  - selecting and using appropriate display(s)
  - identifying features in the data
  - finding an appropriate model
  - describing the nature and strength of the relationship and relating this to the context
  - using the model to make a prediction
  - communicating findings in a conclusion.

  Measurement data can either be discrete or continuous in nature. In regression analysis the y-variable, or response variable, must be a continuous variable. The x-variable or explanatory variable can be either a discrete or continuous variable. The relationship may be non-linear.

- Use and interpretation of R^2 is not expected at this level.

Bivariate Data

Introduction

In statistics we are often interested in identifying relationships involving more than one variable.

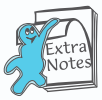
In this booklet we study Bivariate Data which deals with the study of two quantitative variables (pairs of variables) and identifying relationships between them.

We aim to identify the relationship between the variables, graphically with the use of a scatter plot and quantitatively by looking at correlation (the degree to which two or more quantities are linearly associated). In addition we use regression to enable us to make quantitative predictions (interpolation) of one variable from the other.

Much of the emphasis in this Achievement Standard is on the visual interpretation of scatter plots as well as linking statistical knowledge to the context of the question and using appropriate reasoning and reflection. To this end we are suggesting that students have access to iNZight "*a simple data analysis system which encourages exploring what data is saying without the distractions of driving complex software (iNZight website)*".

INZight can be downloaded from https://www.stat.auckland.ac.nz/~wild/iNZight/ and can be installed on either a Mac or Windows computer.

All datasets used in this booklet can be downloaded from our website at www.nulake.co.nz under the 'Downloads' link, 'Year 13', 'IAS 3.9 Bivariate Data', 'Click here' and imported directly into iNZight or Excel.



Relationships

Inspecting a Scatter Plot

When we are given a scatter plot and asked to comment on the relationship between the variables our first approach is a visual one.

The first step is to make sure we understand the precise meaning of the variables being used as well as the units applied to the variables. Sometimes it may be necessary to research the meaning of the variables so that you have a better understanding prior to investigating a possible relationship.

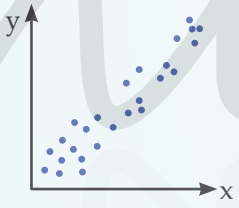
Next look at the degree of scatter of the points. Are the points clumped together or spread out? Are the points in a number of distinct clusters or is there just a single cluster of points with one or two points at the extremes of the scatter plot? Are there any unusual observations that go against the general trend of the scatter plot? Are these points still realistic, i.e. can they be explained?

What is the variation of scatter in the scatter plot? Are the points close to each other or is there significant gaps between them? Do the points follow a discernible pattern or trend, e.g. following a straight line?

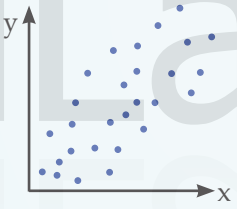
Correlation

When we draw a scatter plot the shape of the plot is usually defined as linear or non-linear (curved). A correlation exists between two variables when one of them is related or can be influenced by the other in some way. The strength of a relationship can be identified visually, on a scatter plot, by how tightly or spread out the plotted points are and the direction of the relationship can be identified by the general slope of the pattern of the data.

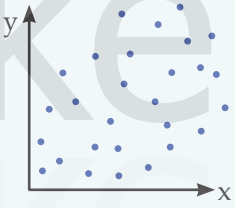
Study the scatter plots below and note the visual description of shape, strength and direction of the relationship between x and y.



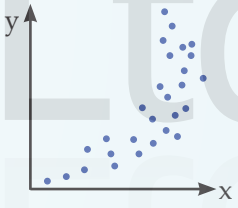
linear, strong positive relationship



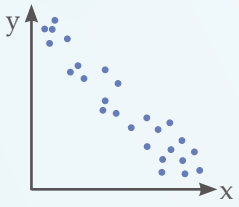
linear, moderate positive relationship



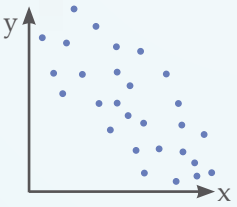
linear, weak positive relationship



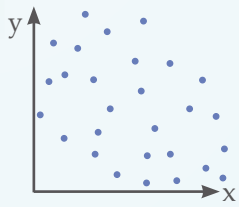
non-linear, strong positive relationship



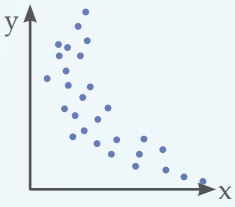
linear, strong negative relationship



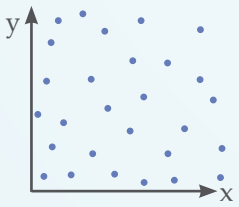
linear, moderate negative relationship



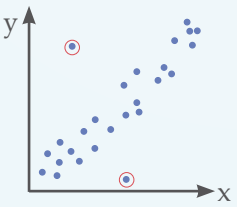
linear, weak negative relationship



non-linear, moderate negative relationship



no relationship



strong relationship with outliers

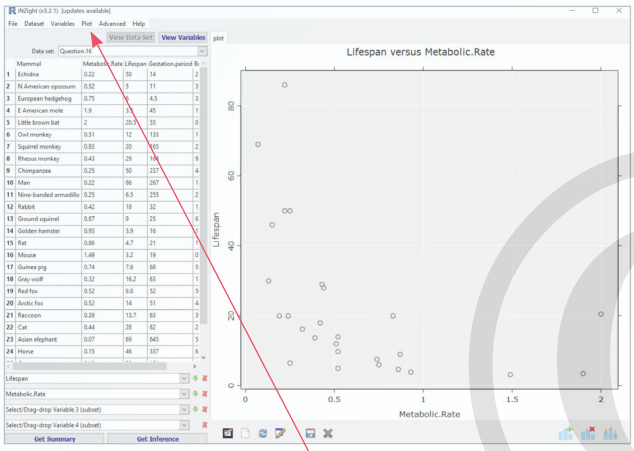
An outlier is one or more points that do not follow the trend.

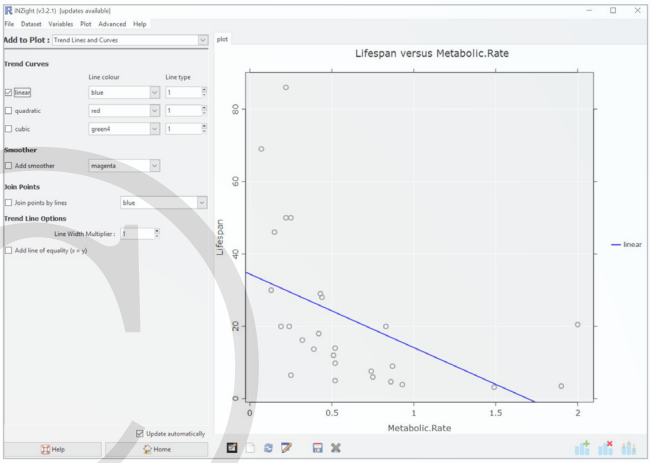
When determining whether a relationship exists between two variables we also want to be able to measure the strength and direction of a relationship quantitatively, and to do this for linear relationships we can calculate the (linear) correlation coefficient (r). This is explained in detail on Page 12.

Example

Calculate the correlation coefficient, r, for each of the three scatter plots involving lifespan created using iNZight for Question 16 parts b), e) and m).

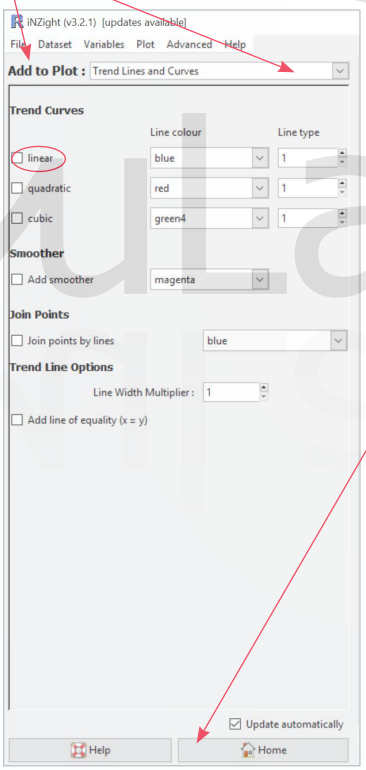
Begin by drawing your scatter plot with the required variables, e.g. metabolic rate and lifespan.



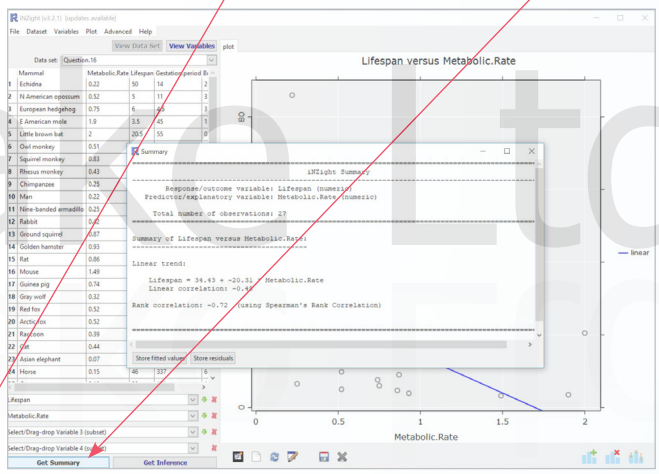


Select the menu option 'Plot' and then 'Add to Plot'. Click on the 'Add to Plot' pop down menu and choose 'Trend Line and Curves'.

Click on the checkbox 'linear'. If the 'Update automatically' checkbox is highlighted the linear trend line (regression line) will appear on your scatter plot (see the screenshot at the top of the right hand column).

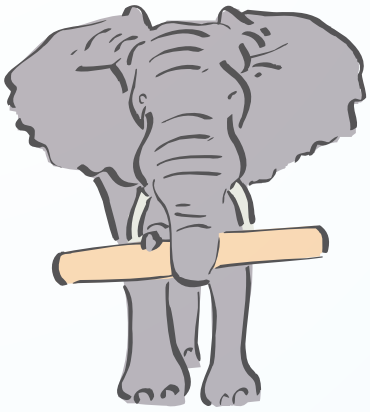


To find the linear correlation coefficient once you have created your scatter plot and added the trend curve, click on the 'Home' button and then the 'Get Summary' button.



The correlation (r) values is r = ⁻0.48 for metabolic

Repeat the same procedure for the other two pairs of variables: i.e. gestation period, lifespan and brain weight, lifespan.



Regression Analysis

Simple Linear Regression

Once we have calculated the linear correlation coefficient for our paired data and we are convinced that a linear relationship exists, the next stage is to obtain a formula that allows us to predict the response variable from the independent variable. This is done through a process called least squares regression.

Regression allows us to model, examine and explore relationships as well as acts as a predictor.

Simple linear regression is used when two variables are thought to be connected by a linear relationship.

Linear regression fits a straight line to the data in a scatter plot.

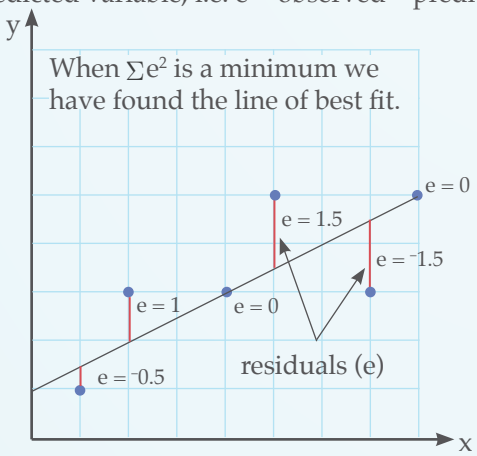
Linear regression is different from correlation analysis where the purpose is to examine the strength and direction of the linear relationship between the two variables.

The main goal of linear regression is to fit a straight line through the data that predicts y based on x.

Simple linear regression uses the least squares method. A least squares regression line is the line which produces the smallest value of the sum of the squares of the residuals.

A residual is the vertical distance from a point on a scatter plot to the line of best fit. Therefore the least squares regression line can be seen as the best line of best fit. The term simple refers to the fact that the least squares regression method is one of the simplest in statistics.

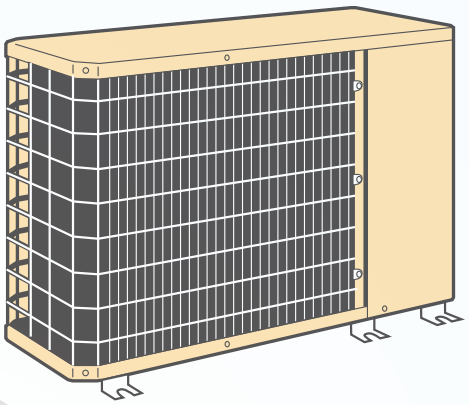
Study the scatter plot below. Our objective is to find the line of best fit so that the sum of the squares of the residuals, usually denoted by the variable e is a minimum. The residual e is the difference between the observed value of the dependent variable and the predicted variable, i.e. e = observed – predicted.



$\Sigma e^2 = {}^{-}0.5^2 + 1^2 + 0^2 + 1.5^2 + {}^{-}1.5^2 + 0^2 = 5.75$

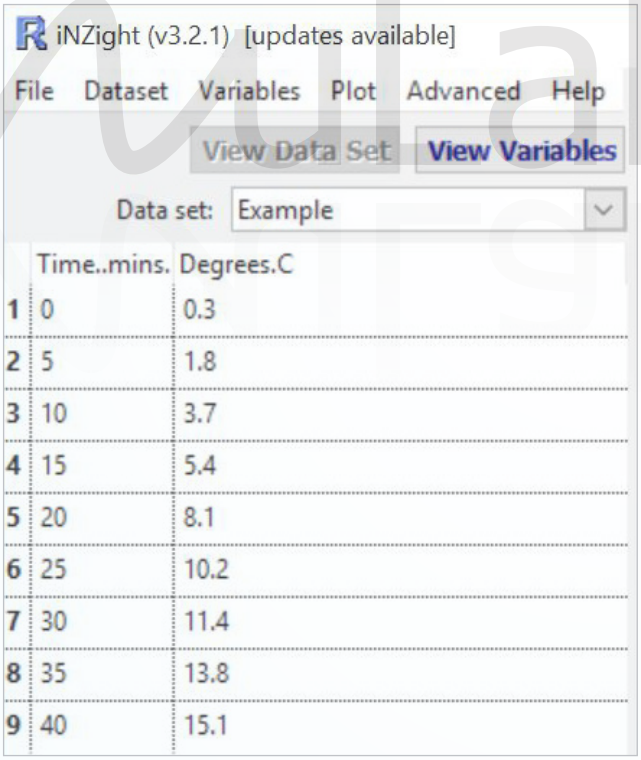
Example

A heater is turned on in a cold room and the temperature (°C) of the room is recorded every five minutes. The results are given below. Find the least squares regression line using iNZight.



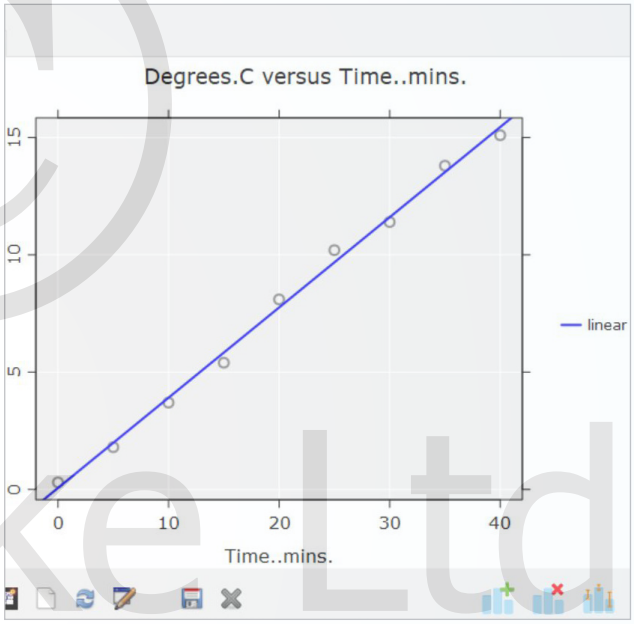
Time (x)	C (y)
0	0.3
5	1.8
10	3.7
15	5.4
20	8.1
25	10.2
30	11.4
35	13.8
40	15.1

Import the dataset 'Example.csv' into iNZight and then draw a scatter plot of time versus temperature.

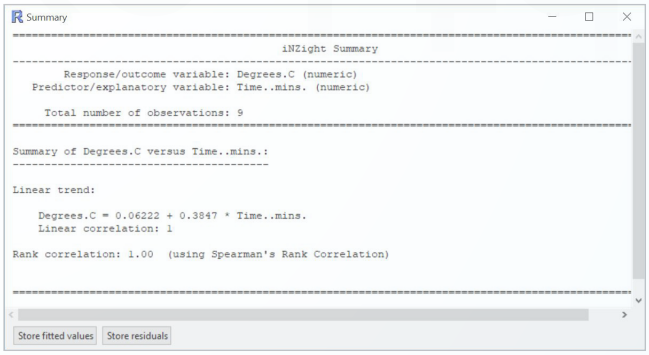


	Time..mins.	Degrees.C
1	0	0.3
2	5	1.8
3	10	3.7
4	15	5.4
5	20	8.1
6	25	10.2
7	30	11.4
8	35	13.8
9	40	15.1

Add a linear trend line (regression line) to the scatter plot by selecting the menu option 'Plot' and then 'Add to Plot'. Click on the 'Add to Plot' pop down menu and choose 'Trend Line and Curves'. Click on the checkbox 'linear. If the 'Update automatically' checkbox is highlighted the linear trend line will appear on the scatter plot.



To find the regression line click on the 'Get Summary' button. Both the correlation coefficient and the regression line are displayed in the window.



The regression line is written in terms of the variable names rather than x and y,

i.e. Degrees.C (y) = 0.06222 + 0.3847 * Time.mins (x)

Residual Plots

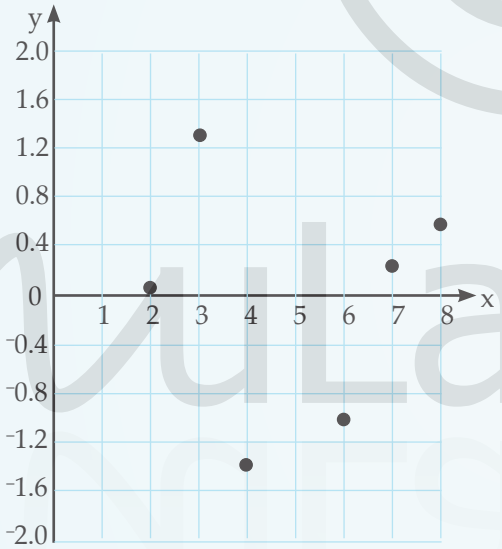
Residual Plots

By examining residual plots we can assess the appropriateness of a linear regression model.

A residual plot is a graph which shows the residuals on the vertical (y) axis and the explanatory variable (independent variable) on the horizontal axis (x).

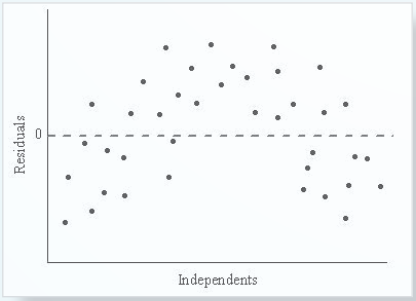
If the points in a residual plot are randomly placed around the horizontal axis then a linear regression model is applicable for the data, otherwise a non-linear model is more appropriate.

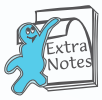
Consider the residual plot below for the example from the previous page. For the six data points in our example the residuals coordinates are (explanatory, residual) (2, 0.0833) (3, 1.3333) (4, ⁻1.4167), (6, ⁻0.9167), (7, 0.3333) and (8, 0.5833).



The residual plot shows a fairly random pattern. The first two residuals are positive, the next two negative and the last two positive. Such a random pattern indicates that a linear model is a good model for the data.

If a residual plot displays a non-random structure then it indicates that a linear model is not a good fit for the data. An example of a non-random residual plot is given below.





A curved pattern in a residual plot indicates that a linear model should not be used and another non-linear one should be investigated, e.g. quadratic, cubic etc.

55. Download the dataset called Question 55.csv from under the 'Downloads' link on our website (www.nulake.co.nz) which gives the bacterium count in culture of three different samples of the same bacteria after so many days.

Import the file into iNZight or Excel.

a) What was the approximate initial bacterium culture count for the samples at the start?

b) Justify why a least squares regression line is not the best model for this data, even though the linear correlation coefficient is 0.87.

c) Carry out an appropriate non-linear regression choosing the best option available within iNZight. Justify your choice and predict the bacteria count on days 3 and 10.

d) Give the strengths and weaknesses of your non-linear model from part c)

e) Import the same file (Question 55.csv) into Excel then draw a scatter plot and fit an appropriate non-linear regression curve to it.

f) Comment on the non-linear regression curve you have used, justifying your choice.

Predict the bacteria count at days 3 and 10 using your curve.

Comment on the strengths and weakness of the model and your predictions.

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

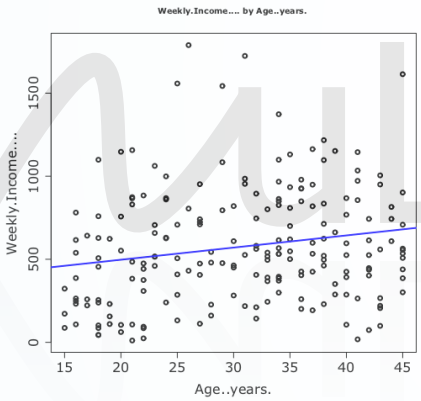
36. ⁻0.99

37. 0.25

Page 20

38. a) $529

b) and c)

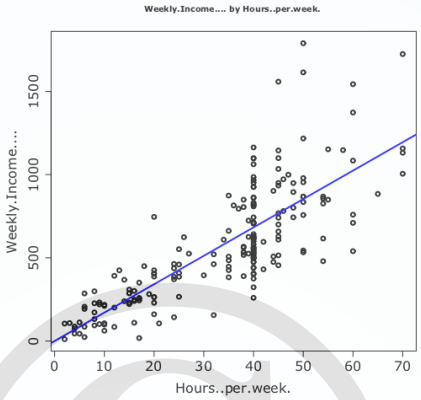


r = 0.19

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)

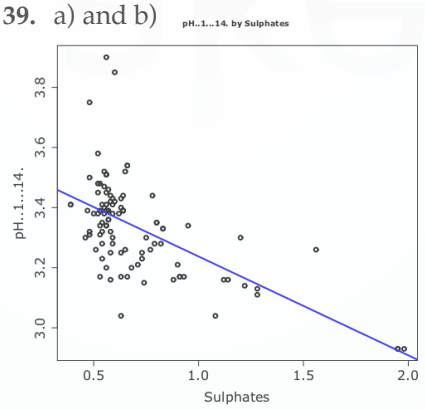


r = 0.80

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)



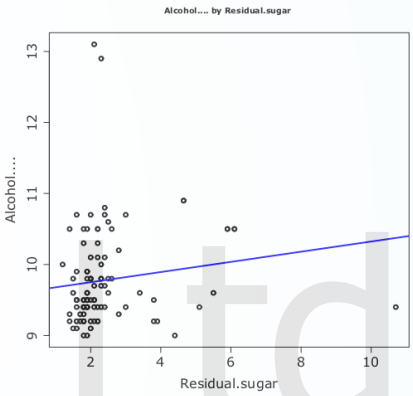
r = ⁻0.62

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.