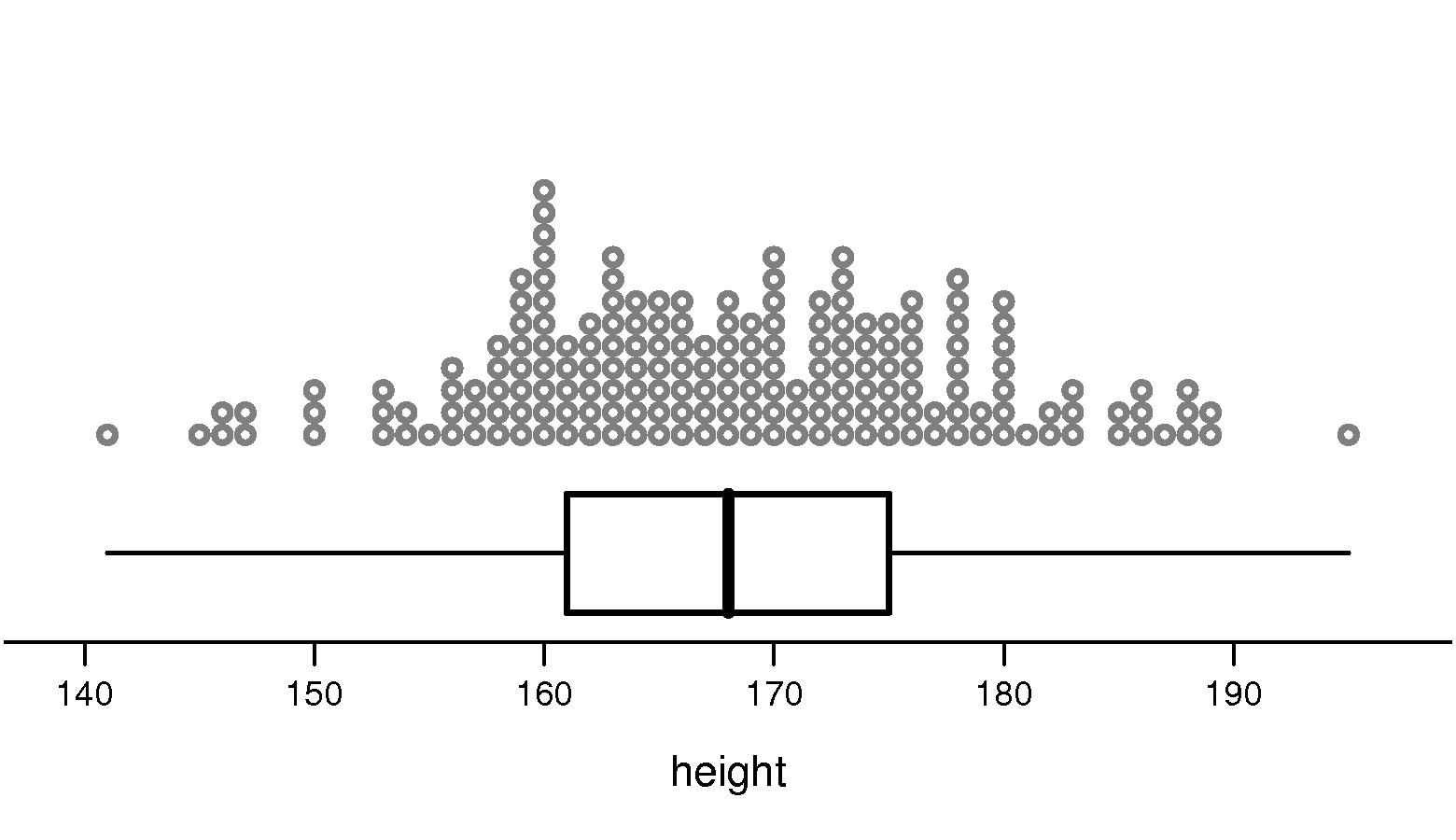
**February 2013**

**AS 91581 Investigate Bivariate measurement data**

In the past we have focused mainly on data from a single variable. Using the data we were able to make a very rough estimate of the value of a member of the population using an average.



of 200 Year 10 students from Census at schools data base

In this topic we turn our attention to data where we have **two pieces of information** about each **unit** we measure. A **unit** may be an individual or an object.

The information we gather about our units are called **attributes**. The attributes that we gather about each unit are more commonly called **variables.** We will be investigating the relationship between the two variables.

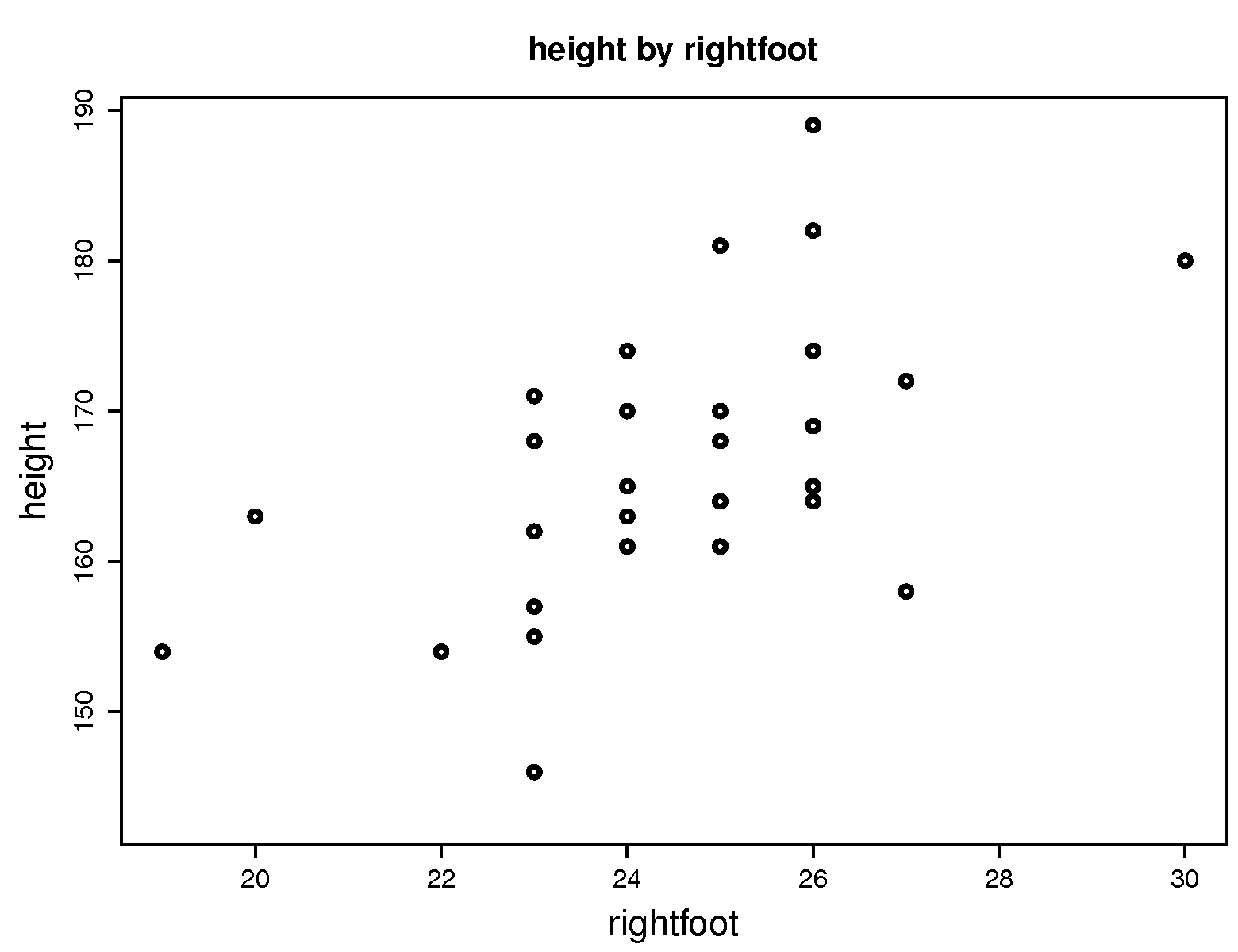
For an individual these variables may be their height, income, gender, etc. For a tree it may be its girth or type. Research often produces data with two pieces of information about each object or individual. Data which has two variables to be studied is called **bivariate data.**

Examples could be:

* the age of kauri trees and their diameter
* house prices and population density
* radioactivity in the ground water and distance from Fukushima nuclear power plant.

**A Scatterplot** is a tool for displaying two pieces of information about an individual on one graph.

Somewhere in your past you might have investigated the relationship between height and foot length of the people in your class. You probably drew a scatter plot similar to the one below. You may have concluded that students with longer feet **tend to** be taller.



Now we want to look further into the relationship between two variables from the same unit. We will be investigating only data which is **quantitative.** The variable on the vertical axis must be **measured** (continuous), the variable on the horizontal axis may be **counted or measured.**

The purpose of drawing a scatterplot is usually to establish whether or not a relationship exists between the two variables.

**Exercise : Describe what you see in these scatter plots.**

22

23

24

25

26

27

28

29

150

160

170

180

190

200

Foot size (cm)

Height (cm)

Height and Foot Size   
for 30 Year 10 Students

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

45

40

35

20

19

18

17

16

15

14

Latitude (°S)

Mean January Air Temperatures   
for 30 New Zealand Locations

Temperature (°C)

Internet Users (%)

% of population who are Internet Users vs   
GDP per capita for 202 Countries

0

10

20

30

40

GDP per capita (thousands of dollars)

0

10

20

30

40

50

60

70

80

wife ($)

Average Weekly Income for Employed New Zealanders in 2001

husband ($)

0

200

400

600

800

1000

1200

0

200

400

600

800

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

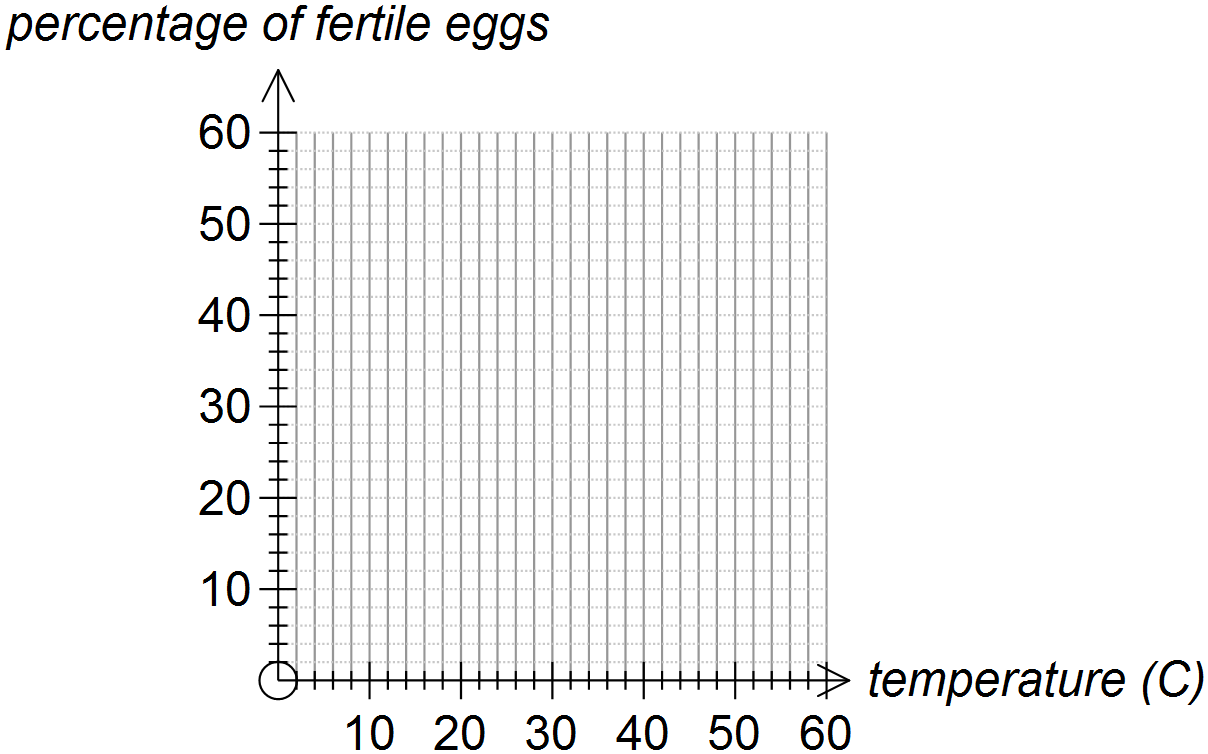
**Example for discussion**

The conservation department at Rotorua raise trout for release into surrounding lakes and rivers in order to prevent their numbers declining. They are investigating the effect of the surrounding water temperature on the percentage of fertile eggs that hatch.

The table below shows the temperature of the surrounding water and the percentage that hatch from randomly selected lots of fertile Rainbow Trout eggs.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Temperature (oC): | 10 | 20 | 30 | 40 | 50 | 60 |
| Percentage of fertile eggs hatching: | 12 | 21 | 26 | 35 | 41 | 52 |

1. Draw a scatter plot of these data.



1. Comment on what the graph reveals about the data.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. How could the hatchery use the table and graph?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Playing with rectangles**

Two hundred Year 9 students were asked to draw a rectangle with sides up to 10cm in length. They were asked to draw and measure sides to the nearest millimetre so that a variety of rectangles would be produced.

They then calculated the perimeter of their rectangle and the area of their rectangle.

Here is a scatterplot of the **length** versus the **width** of each rectangle. Write a description of the scatterplot.





Here is a scatterplot of the **area** of the rectangle versus the **length** of the rectangle. What do you notice?



Here is a scatterplot of the perimeter of each rectangle versus the length of each rectangle. What do you notice?



Here is a scatterplot of the area of each rectangle versus its perimeter. What do you notice?

**Beginning a bivariate investigation: Knowing your data and the context.**

You need to understand what the variables are, the units they are measured in, what is ‘normal’, the setting or reason that this data was collected, who collected it or why it is important.

This will involve you doing some research. Always acknowledge your source of information.

In learning about investigating bivariate data we will work through an example about food.

Do: An Investigation: McDonalds Foods, Step 1 Understanding the context.

**Writing a purpose statement.**

Think about which variables might there be a relationship between (and which there would be no sense in looking at a relationship).

Write down four of five possibilities.

**What makes a good purpose statement?**

* It is written as a question.
* It is written as a relationship question.
* It can be answered with the data available.
* The variables of interest are specified.
* It is a question whose answer is useful or interesting.
* The question is related to the purpose of the investigation.
* Think about the population of interest. Can the results be extended to a wider population?

NB The source of the data should be stated

Types of questions:

Correlation style:

* Is there a relationship between …. and….
* What sort of relationship is there between ….. and … /What is the nature of the relationship between…..

Regression style:

* Can (variable 1) be used to predict (variable 2) for ……

Better questions will incorporate a purpose for the investigation. Why are you interested in investigating this relationship?

e.g. What is the nature of the relationship between actual weight and ideal weight for Year 1 Auckland University maths students? Is this relationship different for males and females?

The better your question the better your investigation will be!

**The Scatter Plot**

…is the only graph used to display bivariate data

Which variable goes on which axis?

… it depends on the variables and on the question …..

The variable we **want to predict** is called **the response variable**. This is plotted on the **y axis**.

The variable we **want to predict from** is called the **explanatory** variable. This is plotted on the **x axis**.

Sometimes, but not always, the data may suggest that one of the variables can “explain” the other: this is called the **explanatory** variable. If this is the case this variable goes on the X axis.

If one variable cannot “explain” the other, we say the two variables are just **associated.**

* If one of the variables is calories and the other variable “% fat content”, “% fat content” would be put on the X axis. Why?
* If the data is gathered from an experiment, with one variable being controlled and measured, and the other variable’s response measured, the controlled variable goes on the X axis. Why?
* If one of the variables is time, this always goes on the X axis, and the data is called a time series. We will be considering this type of data separately later.

Sometimes it does not matter which variable goes on which axis.

Eg Is there a relationship between the time a person can balance on their left leg with their eyes closed and the time a person can balance on their right leg with their eyes closed?

Whatever way round you decide to put the variables, **you should explain why** you put them the way you did.

**Using iNZight to produce a scatter plot::**

Run iNZight. →**Data in/out** → **import data** → **browse for file** → **OK** →click and drag the column name to ‘**drop name here**’. Variable 1 appears on the x-axis. Use the save button to save a copy of your graph (save as jpeg or you may choose to print screen).

Do: An Investigation: McDonalds Foods, Step 2 Deciding on an investigation and writing a question/purpose statement.

**Describing a scatter plot. What do I look for in scatter plots?**

**Trend**

Do you see

a **linear** trend… **or** a **non-linear** trend?

curve

straight line

**Association**

Do you see

a **positive** association… **or** a **negative** association?

as one variable gets bigger, the other gets smaller

as one variable gets bigger, so does the other

**Strength**

Do you see

a **strong** relationship… **or** a **weak** relationship?

little scatter

lots of scatter

**Groupings**

Do you see

any **groupings**?

TASGU

**Anything unusual**

Do you see

any **outliers**?

unusually far from the trend

Does the amount of scatter change over the range of values?

roughly the same amount of scatter as you look across the plot

the scatter is greater in one part of the graph than another

**Only talk about what you do see - not what is not there!**

**Measuring the strength of a relationship**

Having drawn the scatterplot you should print it out. Now draw by hand a ‘band’ (2 lines/curves) that includes most of the scatter. The width and shape of that band tells us about the strength and type of the relationship.

Are your bands straight or curved? This tells us if the relationship is linear or non-linear.

How far apart are the lines? This gives an indication of the strength of the relationship. The closer together the lines are, the stronger the relationship.

On each of the following graphs draw:

(i) a trend line or curve

(ii) a band on each side of the trend that encloses most of the scatter.

Rank these relationships from weakest (1) to strongest (4):

How did you make your decisions?

Do: An Investigation: McDonalds Foods, Step 3 The Graph.

The strength of a **linear** relationship may also be measured using the **CORRELATION COEFFICIENT, little r.**

* The correlation coefficient is a number between -1 and 1.
* A positive sign indicates that y increases as x increases, a negative sign indicates that y decreases as x increases.
* The closer r is to zero the weaker the relationship.
* If r is 1 or -1, there is perfect correlation, that is, the points lie in one single line.
* r has no units

This table gives a ‘rule of thumb’ for how a relationship with a particular value of r may be described.

|  |  |
| --- | --- |
| **value of r** | **description of relationship** |
| 1 | A perfect correlation |
| | r | >0.8 | strong correlation |
| 0.5 < | r | < 0.8 | Moderate correlation |
| 0.3< | r | < 0.5 | Weak correlation |
| | r |< 0.3 | Very weak or no correlation |

In summary: - What does the value of *r* tell us?

*r* measures how close the points in a scatter plot come to lying in a **straight line.** It is a numerical measure of the scatter and therefore of the **strength** of a linear relationship.

The precision of a prediction depends on the strength of the relationship.

The sign on r indicates the **direction** of the **linear** association.

***r* is only used for linear relationships.**

**CAUTION**

* Use correlation only if you have two **quantitative** variables

There is an association between gender and weight but there isn’t a correlation between gender and weight

* Use correlation only if the relationship is **linear**.
* Beware of **outliers** (see later)

Use the following applet to estimate the position of a regression line and the value of the correlation coefficient, r.

<http://www.ruf.rice.edu/~lane/stat_sim/reg_by_eye/index.html>

**Obtaining a model**

Inserting a linear trend line, sometimes called a line of best fit.

Use your ruler to decide where to draw a straight line through the scatterplots you ticked above so that

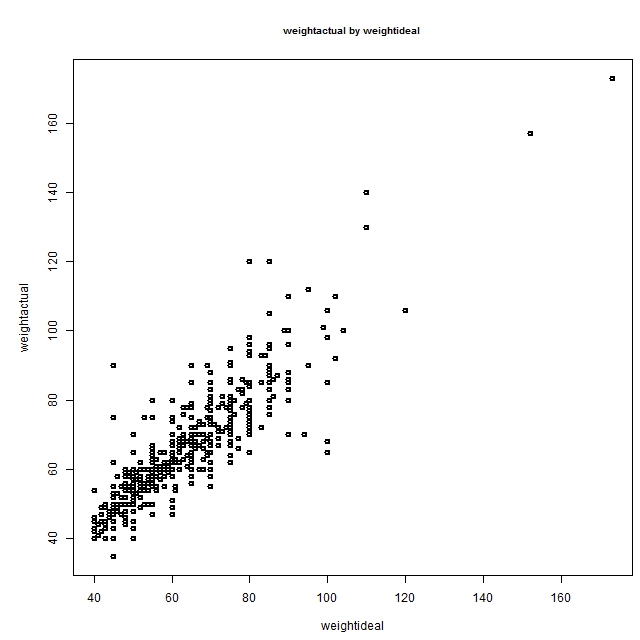
* The gradient of the line matches the overall trend of the data
* The line passes through as many points as possible
* There are as many points on one side of the line as the other

Obtaining the equation of the fitted line

We will now turn our attention to the equation of the fitted line. Remember that to describe a straight line graph we need two pieces of information:

1. the gradient and 2. the y-intercept.

We can estimate the equation of our drawn line from the graph.



This scatterplot shows the relationship between the weight that they would like to be and the actual weight for a group of young adults.

Draw a trendline on the graph.

Estimate the gradient of your line:

**Interpreting the gradient**

A linear trend line suggests that the response variable, Y, changes by a fixed amount for each unit change in X. So if the trend line has equation y = 5x + 7 the gradient is 5, Y changes by 5 units for every change of 1 unit in X. It is important to refer to the units when describing the gradient or rate of change of Y as X changes.

Interpret the gradient of the above graph:

**In iNZight, add to plot**

**Trend curve**

**Tick linear (or quadratic or cubic if the trend looks non-linear)**

**Ok**

Inserting a curve trend.

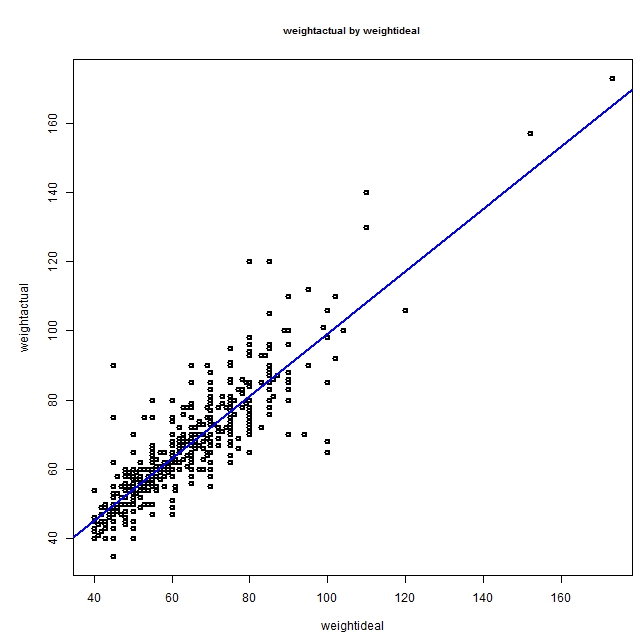
Where the scatter forms an obvious curve, you can draw a freehand curve to give an idea of what the trend looks like. In iNZight, you only have the option of a quadratic or cubic curve. You should try both to see which gives the better fit to the scatter. If neither of these is appropriate you need to comment that this statistical software does not provide an appropriate model.

**Making a prediction**

A prediction can be made by reading from the graph.

A prediction can be made by substituting a suitable value into the trend equation, but with reference to the amount of scatter and the width of the band drawn around the data. Unless there is almost no scatter, a prediction should be given as an interval. This interval would be determined by the width of the band.

Predictions should only be made for values within the data range.



Band that encloses scatter

weightactual = 0.9 \* weightideal + 9.07

correlation = 0.88

trendline

weightactual = 0.9 \* weightideal + 9.07

Correlation = 0.88

To predict a young adult’s actual weight if their ideal weight is 70kg:

Substitute into the trend equation weightactual = 0.9 70 + 9.07 = 72.07

Looking at the scatter I predict that a young adult’s actual weight would be between 55kg and 90kg if their ideal weight was 70kg. The trend predicts an average weight of 72kg.

**Note: Round your answer sensibly in context!**

Because the relationship is moderately strong (the points are close to the trend) my estimate of 72kg is reasonably precise.

Use **precise, reasonably precise** or **not very precise** to discuss your prediction

To add a trendline using iNZigt:

**add to plot** → **add trend curves** → **linear/quadratic/cubic**

To get the equation of the trendline and the value of the correlation coefficient:

**Get summary** →[ right click, select all, copy] so that you can paste this into a word document.

Do Exercise 1 Drawing scatterplots

Do: An Investigation: McDonalds Foods, Step 4 Obtaining a model and making a prediction.

**Taking Care with Predictions**

The data in the scatter plot below were collected from a set of heart attack patients. The response variable is the creatine kinase concentration in the blood (units per litre) and the explanatory variable is the time (in hours) since the heart attack.

Source: *Chance Encounters: A First Course in Data Analysis and Inference* by Christopher J. Wild and George A. F. Seber, p514.



Comment on the scatter plot.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Suppose that a patient had a heart attack 17 hours ago.

Predict the creatine kinase concentration in the blood for this patient.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

In fact their creatine kinase concentration was 990 units/litre.! Comment!

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The complete data set is displayed in the scatter plot below.



**So …beware of extrapolating beyond the data.**

* A fitted line will often do a good job of summarising a relationship for the range of the observed *x*-values.
* Predicting *y*-values for *x*-values that lie beyond the observed *x*-values is dangerous. The linear relationship may not be valid for those *x*-values.

**Further factors to consider:**

**1. Outliers and unusual points**

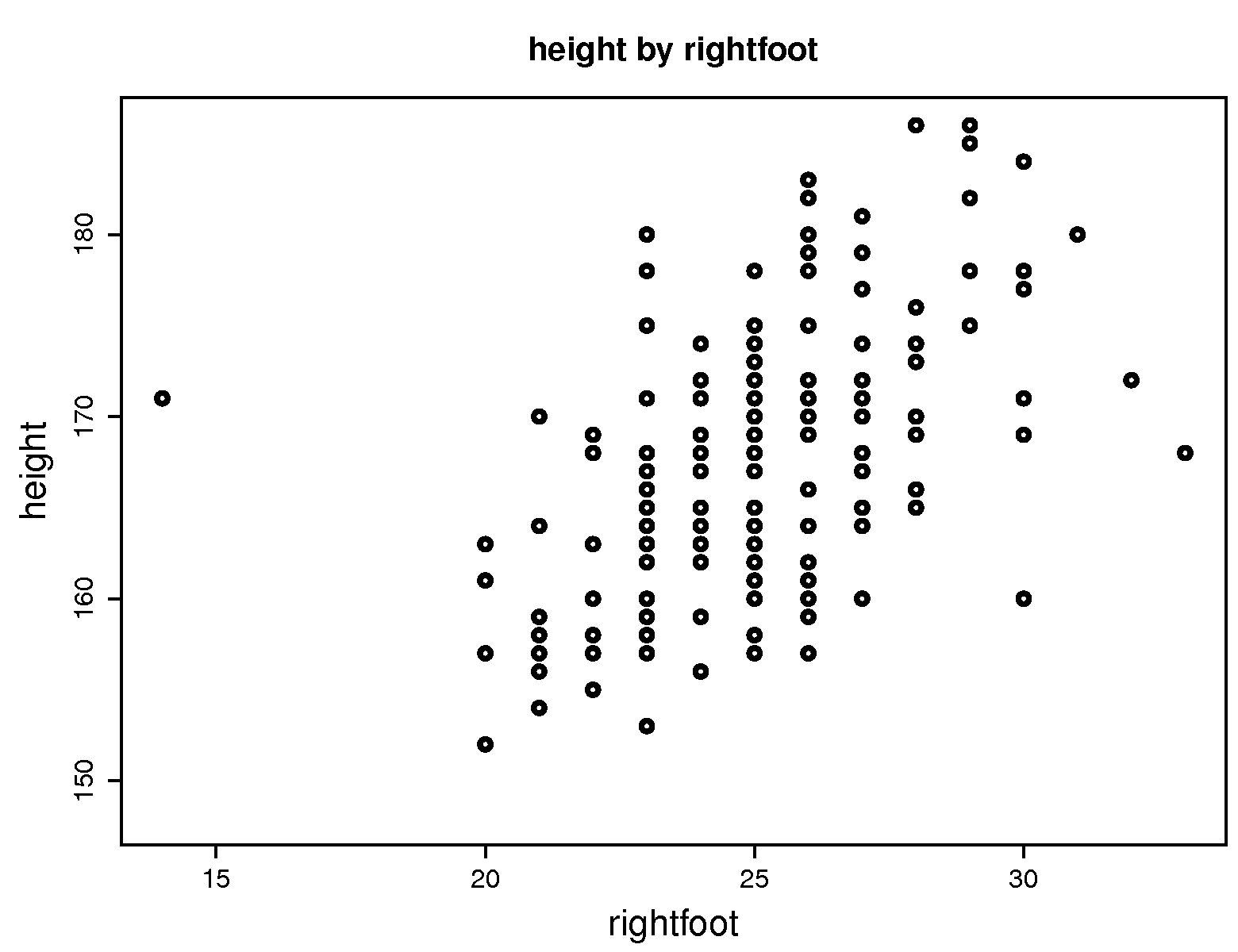
An outlier, in a regression context, is a point that is unusually far from the trend. It would lie a distance from the band of points.

An unusual point is far from the trend and may impact on the model

* An outlier should be checked out to see if it is a mistake or an actual unusual observation.
* If it is a mistake then it should either be corrected or removed (clean the data).
* If it is an actual unusual observation then try to understand why it is so different from the other observations.

If it is an actual unusual observation (or we don’t know if it is a mistake or an actual observation) then carry out two linear regressions; one with the outlier included and one with the outlier excluded. Investigate the amount of influence the outlier has on the fitted trend and discuss the differences. If the trendlines are so different that the outlier can be considered to have an undue influence on the regression, then the line without it can be considered as a more reliable basis for prediction. If the two trendlines are similar, then the outlier does not have an undue influence and the prediction with it included is a more reliable basis for prediction.

eg foot size compared to height for Year 10 students.



There is one student with a foot length of less than 15cm, but the student is quite tall (more than 170cm). This does not seem to fit in with the usual pattern of foot sizes for teenage students and this point lies well outside the band of points. Maybe the student was measured using different units (inches rather than centimetres). Maybe the data has been recorded incorrectly. If possible the sensible thing to do would be to go and check these possibilities out. However we do not have access to the students whose data this is.

In either case, the data point does not fit the normal pattern of teenage foot sizes and we might be justified in omitting this point in our analysis.

Whatever you do, explain and justify your actions!

Example: Life expectancies and gestation period for a sample of non-human mammals.

elephant

Should the elephant be regarded as an outlier?

Certainly this point is well away from the rest of the data. However, it could be that other animals do exist with gestation periods and life expectancies that would fill in the gap, but they have not been included in the sample. Be cautious about discarding points that lie away from the rest of the scatter.

The two examples above are called x-outliers.

* An x-outlier is a point with an extreme x-value.
* An x-outlier can alter the position of the fitted line substantially, i.e. it can influence the position of the fitted line.

The fitted line may say more about the x-outlier than about the overall relationship between the two variables.

Unusual y-values may also be pull the trend line away from the middle of all the points in the scatterpot.

If a data set has an outlier/unusual point then carry produce two graphs; one with the outlier included and one with the outlier excluded. Investigate the amount of influence the outlier has on the fitted line/curve and discuss the differences. Also consider the effect on predictions.

Example:

The following table shows the winning distances in the men’s long jump in the Olympic Games for years after the Second World War.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Year** | **Winner** | **Distance** | **Year** | **Winner** | **Distance** |
| 1948 | Willie Steele (USA) | 7.82m | 1980 | Lutz Dombrowski (GDR) | 8.54m |
| 1952 | Jerome Biffle (USA) | 7.57m | 1984 | Carl Lewis (USA) | 8.54m |
| 1956 | Gregory Bell (USA) | 7.83m | 1988 | Carl Lewis (USA) | 8.72m |
| 1960 | Ralph Boston (USA) | 8.12m | 1992 | Carl Lewis (USA) | 8.67m |
| 1964 | Lynn Davies (GBR) | 8.07m | 1996 | Carl Lewis (USA) | 8.50m |
| 1968 | Bob Beamon (USA) | 8.90m | 2000 | Ivan Pedroso (Cuba) | 8.55m |
| 1972 | Randy Williams (USA) | 8.24m | 2004 | Dwight Phillips (USA) | 8.59m |
| 1976 | Arnie Robinson (USA) | 8.35m |  |  |  |

Source: <http://www.sporting-heroes/stats_athletics/olympics_trackandfield/trackfield.asp>



Comment on the scatter plot.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The following summary information was found for this plot:

y= 0.0164x + 7.8097 r = 0.77

We estimate that for every 4-year increase in years (from one Olympic Games to the next) the winning distance increases by \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, on average.

Using this linear regression we predict that the winning distance in 2004 would be \_\_\_\_\_\_\_\_\_\_\_.

With the 1968 winning distance removed the following summary information is found:

y = 0.0177x + 7.7158 r = 0.906

We estimate that for every 4-year increase in years (from one Olympic Games to the next) the winning distance increases by \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, on average.

Using this linear regression we predict that the winning distance in 2004 will be \_\_\_\_\_\_\_\_\_\_\_.

What effect did the 1968 observation have on the:

fitted line?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

predicted winning distance in 2004?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

value of r?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

22

23

24

25

26

27

28

29

150

160

170

180

190

200

Foot size (cm)

Height (cm)

Height and Foot Size   
for 30 Year 10 Students

What will happen to the position of the model trend line if the tallest Year 10 student is removed?

Tick your answer:

It will get smaller It won’t change It will get bigger

600

500

400

300

200

100

0

40

30

20

10

Gestation (Days)

Life Expectancy (Years)

Life Expectancies and Gestation Period   
for a sample of non-human Mammals

Elephant

What will happen to the correlation coefficient if the elephant is removed?

Tick your answer:

It will get smaller It won’t change It will get bigger

**Investigating unusual points using iNZight:**

Points can easily be removed and restored.

**Filter data Select cases Filter** by (probably you will use row number- this removes the entire row of data)

You will now be presented with a blank page on which to draw a new scatterplot. (Your previous one is on another tab)

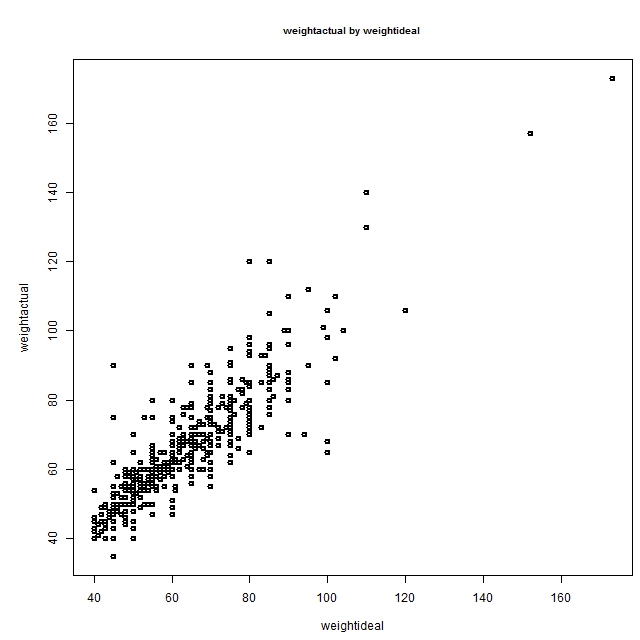
To replace the data: **Filter data Restore Dataset**

Do Exercise 2: Investigating outliers.

**2. Groups**

Sometimes a scatterplot will show distinct groups of points. Do the groups represent different subcategories of the variables? If so, does it make sense to analyse the data using two (or more) separate scatterplots? Sometimes you will have categorical information about the data that warrants investigating the categories separately (eg male, female).

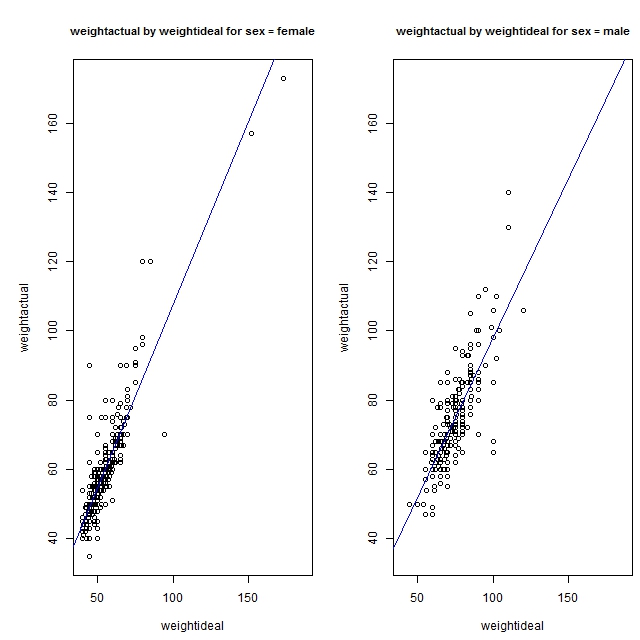
Complete the analysis as one whole set of data first. Try to explain why there are apparently separate groups. Do the analysis again, separating the groups. Does this give less scatter of points from the line of best fit? Which model is better? The prediction based on which model is likely to be more believable?



weightactual = 0.9 \* weightideal + 9.07

Correlation = 0.88

This scatterplot shows the relationship between the weight that they would like to be and the actual weight for a group of young adults, both female and male.



Summary for sex = female

weightactual = 1.06 \* weightideal + 1.29

Correlation = 0.89

Summary for sex = male

weightactual = 0.92 \* weightideal + 5.67

Correlation = 0.8

This scatterplot shows female and male as separate groups. While there is a relationship between actual weight and ideal weight for these young adults as a whole, there is some sense in also analysing the two groups separately.

In iNZight you can show groups in two ways:

1. **add to plot** → **code more variables** →  **colour by levels of** →choose from drop down menu which variable) → **show changes**.

This gives you one graph for all the data but colour coded by category.

2. in left hand box under data columns: →**subset by** → (drop name here)

This gives you separate graphs for each category. The graphs have the same scales on the axes to allow for visual comparison. Get summary will give separate trend equations for each graph.

Do: Exercise 3 Investigating groups.

Do: An Investigation: McDonalds Foods, Step 5 Investigating groups or outliers

**3. Correlation is not Cause!**

When you find that there is a relationship between two variables then you have found a relationship and nothing more than that!

Finding a positive relationship between the number of doctors per 100 000 population and the number of cell phones per 100 000 population does not mean that more doctors will cause there to be more phones. It is more likely that some other economic factor is the cause of both more doctors and more phones. Do not speculate. Simply state that while you have found a relationship between the two variables there is no suggestion that you have any evidence about why there is such a relationship. A scatterplot, even one showing a very strong relationship, does not establish a cause between the two variables. The explanation maybe that both the variables are related to a third variable not being measured – a “lurking” or “confounding” variable.

Note: These variables are positively correlated!

* Number of fire trucks vs amount of fire damage
* Teacher’s salaries vs price of alcohol
* Number of policemen vs number of crimes

There would be a strong positive relationship (correlation) between the damage that a fire does and the number of firemen involved in fighting the blaze. Do we conclude that the firemen cause the damage? Of course, the most likely explanation of this correlation is that the size of the fire (an external variable that hasn’t been included in the study) caused the damage as well as number of firemen needed.

Cause **cannot** be established by plotting a scattergraph.

It is easy to draw scatterplots that are sheer nonsense. For example, the Number of McDonald’s sold vs Number of Medals won in the Sydney Olympics for countries that took part**.** Although the points in this scatter plot may or may not show an association, it is ridiculous to conclude that the Number of McDonald’s is related in any way to the Number of Medals won in Sydney. This is an example of a **spurious correlation**. The two variables may be related because they both depend on a third factor, in this case, population.

600

500

400

300

200

100

0

80

70

60

50

People per Television

Life Expectancy

Life Expectancy and Access

to Television for

a Sample of 40 Countries

40000

30000

20000

10000

0

80

70

60

50

People per Doctor

Life Expectancy

Life Expectancy and Availability of Doctors for a Sample of 40 Countries

Can you suggest another variable that is linked to life expectancy and the availability of doctors (and televisions) which explains the association between the life expectancy and the availability of doctors (and televisions)?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**4. Summary statistics**

Where appropriate you could calculate (and comment on) the mean / median and standard deviation of each of the two variables.

eg The average height of Year 11 students based on this sample is 164cm with a standard deviation of 6.5cm. The average length of armspan is 161.5cm with a standard deviation of 5.7cm. In general a Year 11 student can be expected to have an armspan that is slightly shorter than their height. However there is a greater variation in heights as opposed to lengths of armspans.

**5. Assumptions about the data/potential sources of bias**

Consider the source of the data and any assumptions you have made about it.

The data set should be large enough and (ideally) randomly selected so as to give a good representation of the population to which you are referring.

Bias is the distortion of the outcomes when the sampling method is faulty. For example, if a survey is to be conducted in a city, collecting data from only particular areas of the city may lead to biased results.

Other variables which might affect the variable you are investigating should be kept constant or controlled.

Think about and comment on other variables for which you have no information, which may impact on your findings.

Be clear about the population that you are making conclusions about.

**6. Comparing relationships**

Have you investigated different relationships and now you need to compare them. Compare the strength and direction of the relationships and predictions made from the models you have developed.

**7. The relevance and usefulness of evidence**

Consider how widely your findings can be applied, and in what situations they may be relevant

For whom would your investigation be useful?

Research about the data.

Do: An Investigation: McDonalds Foods, Step 6 Discussion

Resources:

McDonalds data set

Hector’s dolphin data set

Australian Institute of Sport data set

Extension:

Take a look at <http://www.gapminder.org> and explore the way scatter plots look with linear scales and with log scales.