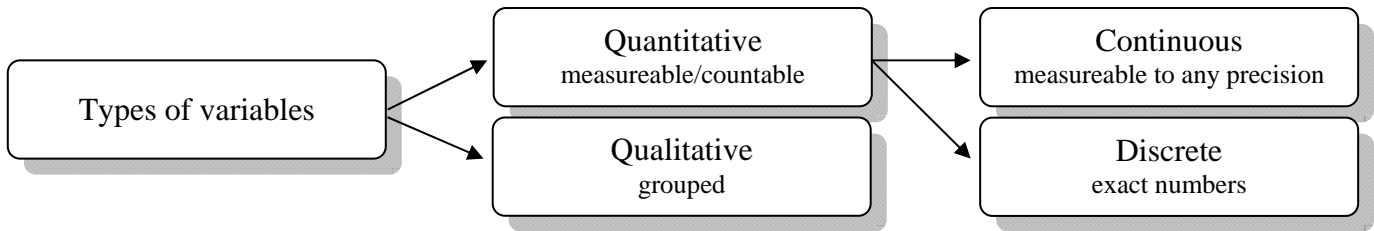
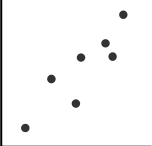


Bivariate Data Analysis (with answers filled in)

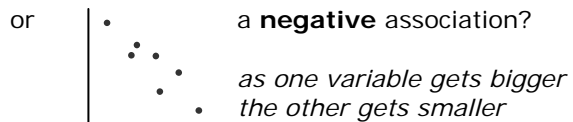
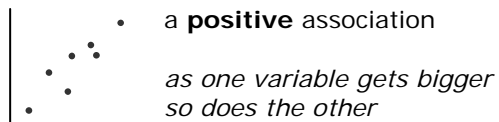
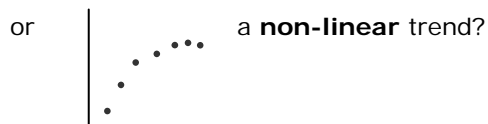
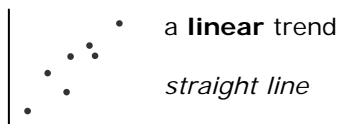
This is adapted from University of Auckland Statistics Department material. The original can be found at <http://www.stat.auckland.ac.nz/~teachers/2003/regression.php>

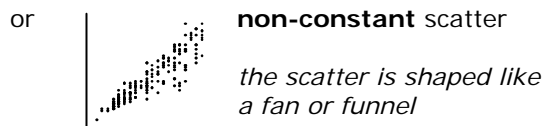
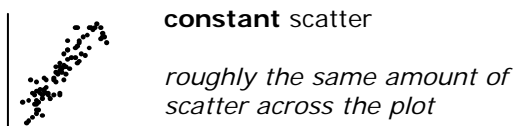
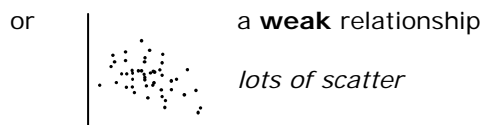
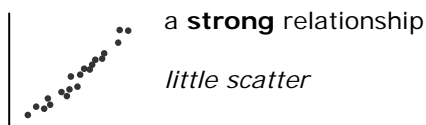
The **scatter plot** is the basic tool used to investigate relationships between two **quantitative** variables.

What do I look for in scatter plots?

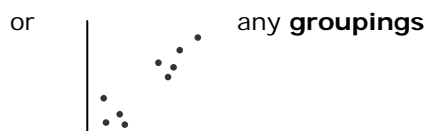
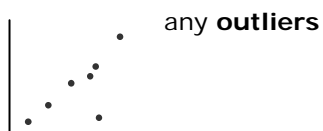
Trend



Scatter

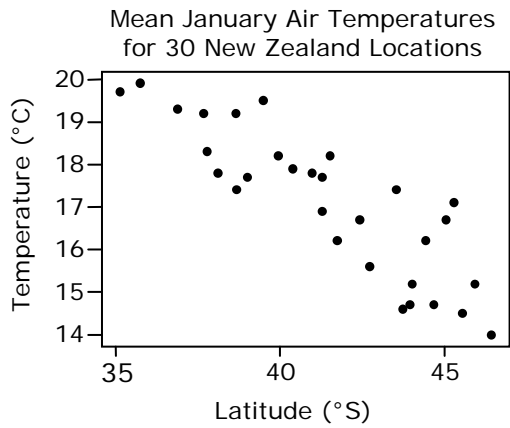


Anything unusual



Exercise:

What do I see in these scatter plots? Try to say as many (correct, useful) things as you can:

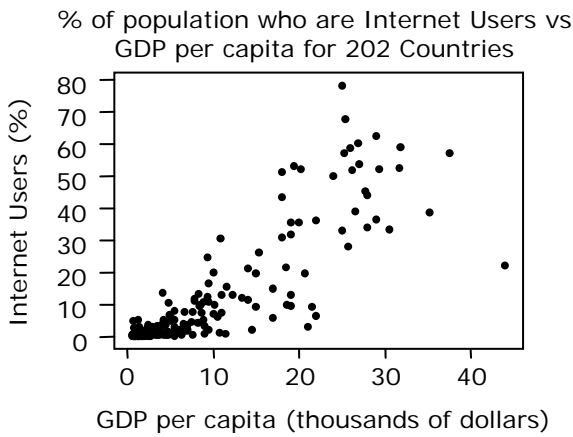


There appears to be a linear trend.

There appears to be moderate constant scatter about the trend line.

Negative association.

No outliers or groupings visible.

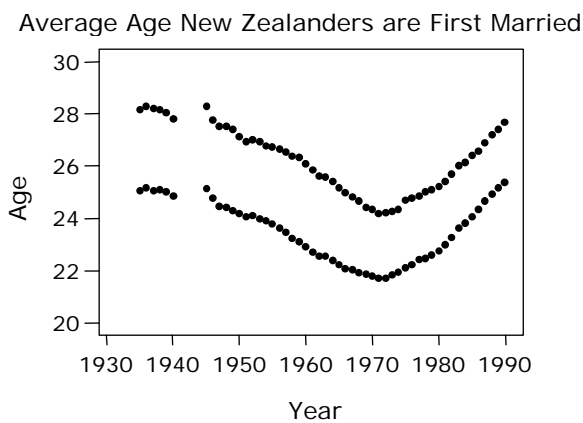


There appears to be a non-linear trend.

There appears to be non-constant scatter about the trend line.

Positive association.

One possible outlier (Large GDP, low % Internet Users).



Two non-linear trends (male and female).

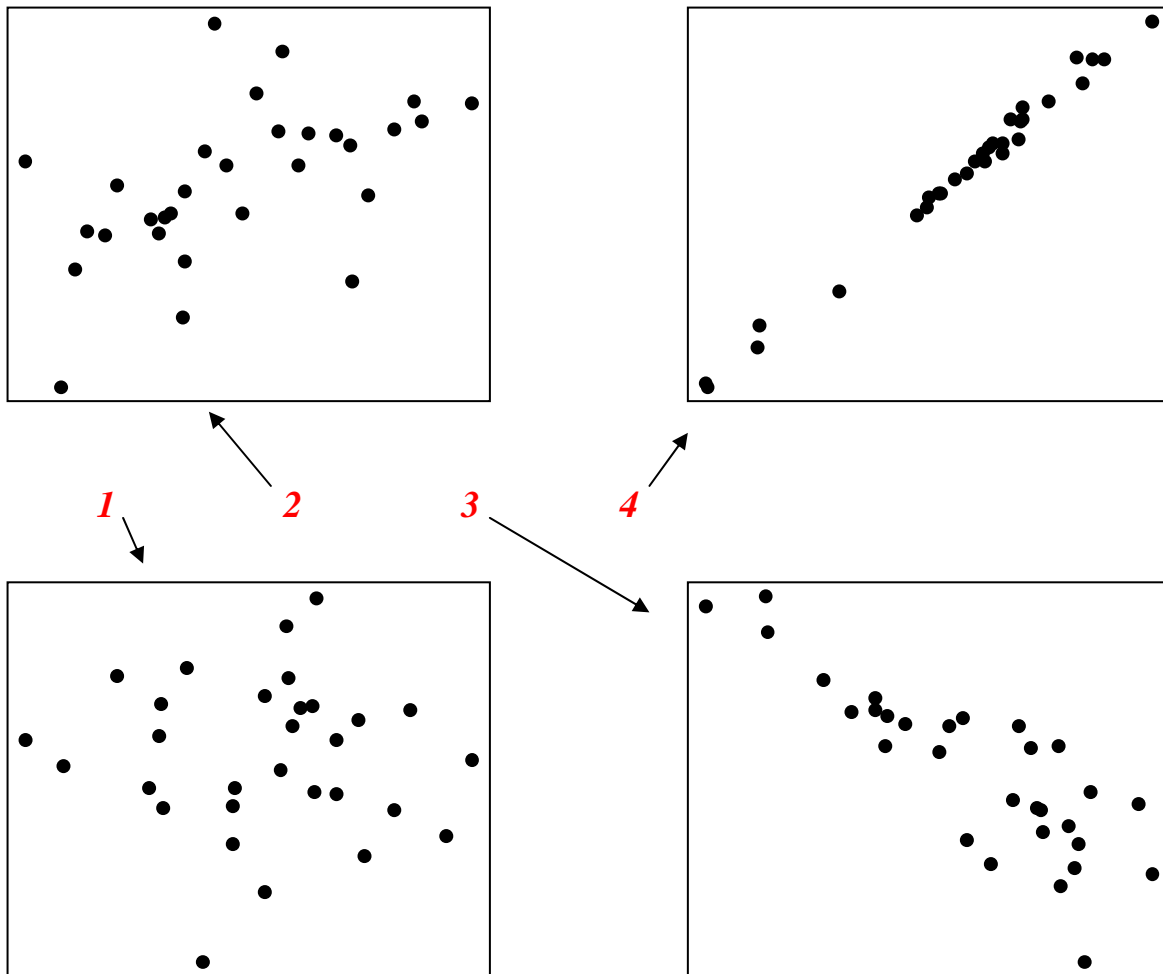
Very little scatter about the trend lines

Negative association until about 1970, then a positive association.

Gap in the data collection (Second World War).

Exercise:

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



What features do you see?

The bottom left plot has no discernable relationship: it could fit a vertical line as easily as a horizontal one. It is the weakest.

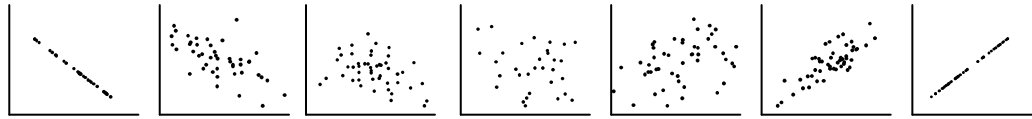
The top left appears to have a weak positive relationship.

The bottom right plot has a moderate negative relationship, but also a funnel shape: the relationship is much stronger at the top left than the bottom right.

The top right plot has a strong positive relationship.

Correlation

- Correlation measures the **strength** of the **linear association** between two **quantitative** variables
- Get the correlation coefficient (r) from your calculator or computer
- The correlation coefficient has no units
- r has a value between -1 and $+1$:

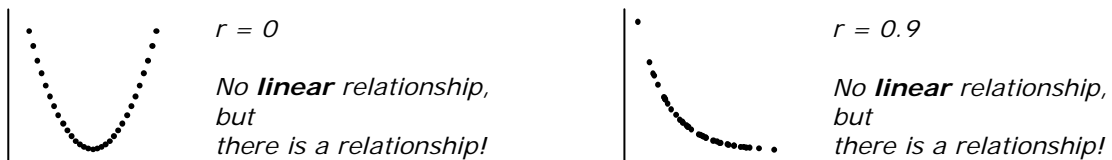


$r = -1$ $r = -0.7$ $r = -0.4$ $r = 0$ $r = 0.3$ $r = 0.8$ $r = 1$

Points fall exactly on a straight line *No linear relationship (uncorrelated)* *Points fall exactly on a straight line*

What can go wrong?

- 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!
- The variables should be **continuous** (or nearly continuous) for an accurate r value.
- Use correlation only if the relationship is **linear**
- Beware of outliers!
- Always** plot the data **before** looking at the correlation



Causation

Two variables may be strongly associated (as measured by the correlation coefficient for linear associations) but may not have a cause and effect relationship existing between them. The explanation maybe that both the variables are related to a third variable not being measured – a “**lurking**” or “**confounding**” variable.

These variables are positively correlated:

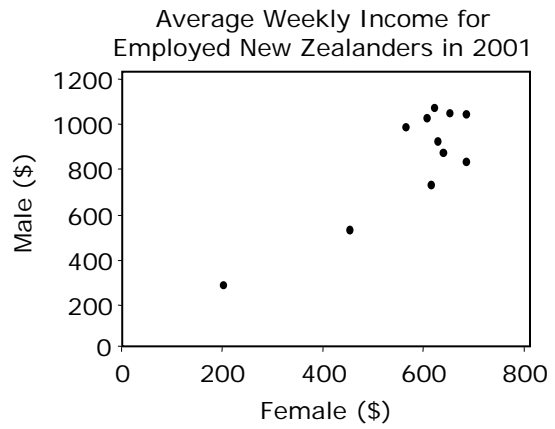
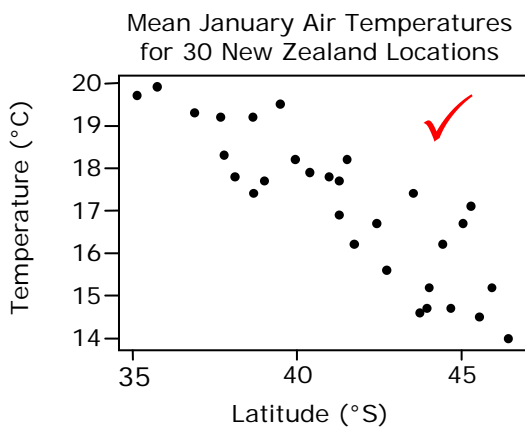
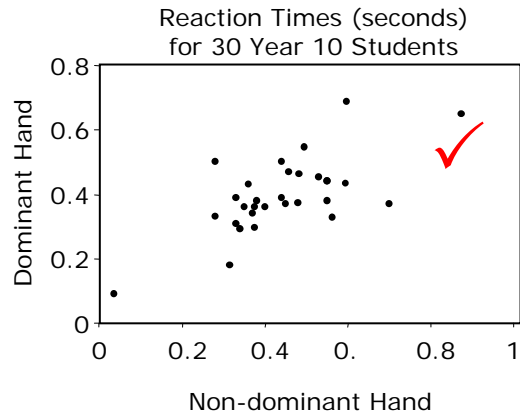
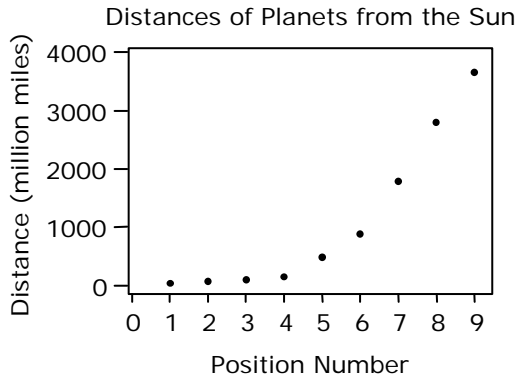
- Number of fire trucks vs amount of fire damage
- Teacher’s salaries vs price of alcohol
- Number of storks seen vs population of Oldenburg Germany over a 6 year period
- Number of policemen vs number of crimes

Only talk about causation if you have well designed and carefully carried out **experiments**. That way confounding variables can be excluded.

If you do suspect that one variable is causing the other, then that variable should go along the x axis and is called the “explanatory” variable. If you suspect a causal link, but don’t know which way, then the one that you are controlling in your experiment, the “control” variable is placed along the x and the measured variable is the y . In other cases it makes no difference.

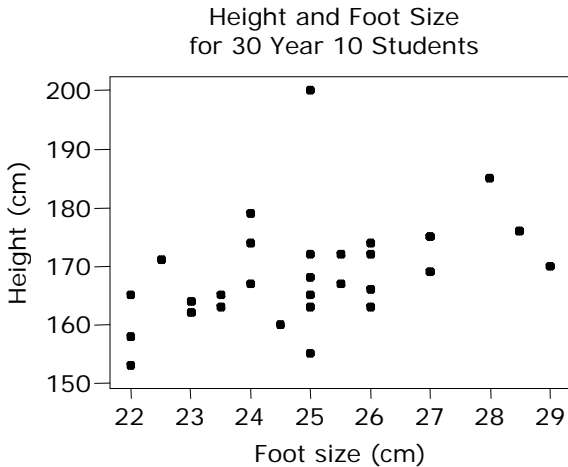
Exercise:

Tick the plots where it would be OK to use a correlation coefficient to describe the strength of the relationship:



Exercise:

What do I see in this scatter plot?



Appears to be a linear trend, with a possible outlier (tall person with a small foot size.)

Appears to be constant scatter.

Positive association.

Foot size will not cause height, but it seems likely that they have a common underlying cause.

What will happen to the correlation coefficient if the tallest Year 10 student is removed? Tick your answer:

(Remember the correlation coefficient answers the question: "For a linear relationship, how well do the data fall on a straight line?")

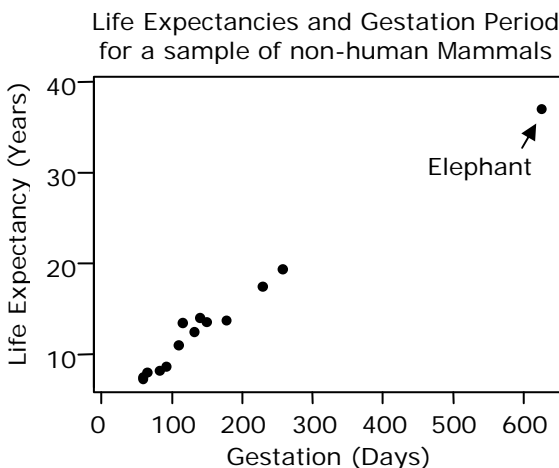
It will get smaller

It won't change

It will get bigger

Exercise:

What do I see in this scatter plot?



Appears to be a strong linear trend.

Outlier in x (the elephant).

Appears to be constant scatter.

Positive association.

What will happen to the correlation coefficient if the elephant is removed? Tick your answer:

It will get smaller

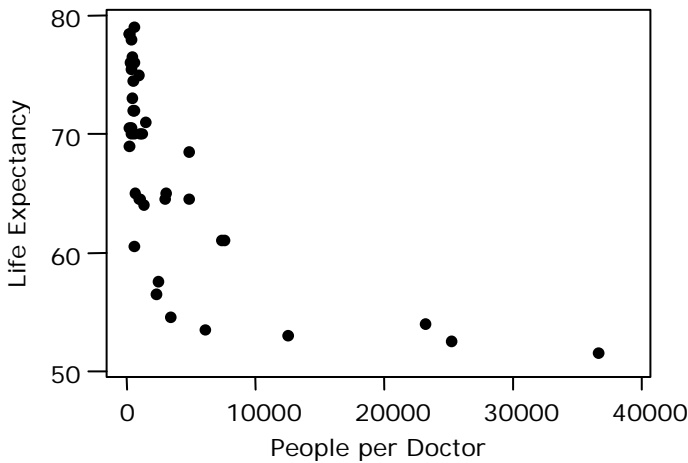
It won't change

It will get bigger

(Though not a lot smaller)

Exercise:

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



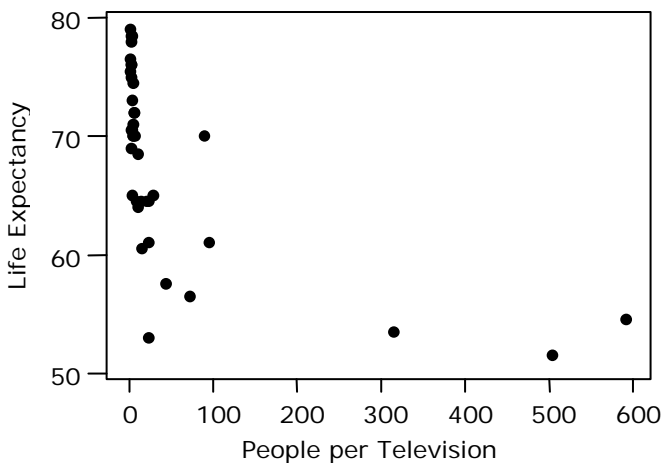
Using the information in the plot, can you suggest what needs to be done in a country to increase the life expectancy? Explain.

Perhaps more doctors in a country raises life expectancy.

Alternatively there may be some criterion (probably wealth) that is strongly associated with life expectancy and the number of doctors.

(Technically the graph would suggest that killing lots of people would also lower the ratio of people to doctor and so raise life expectancy, but that is not really a useful suggestion.)

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



Using the information in this plot, can you make another suggestion as to what needs to be done in a country to increase life expectancy?

It looks like if you decrease the number of people per television (i.e., have more TVs per person), then the life expectancy will increase!

More likely wealth is strongly associated with both, so that raising wealth increases life expectancy (and TV numbers too).

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)?

Some measure of wealth of a country, e.g., average income per person or GDP.

Note: to use correlation you need the wealth to be quantitative – that is given a number.

Data Sources

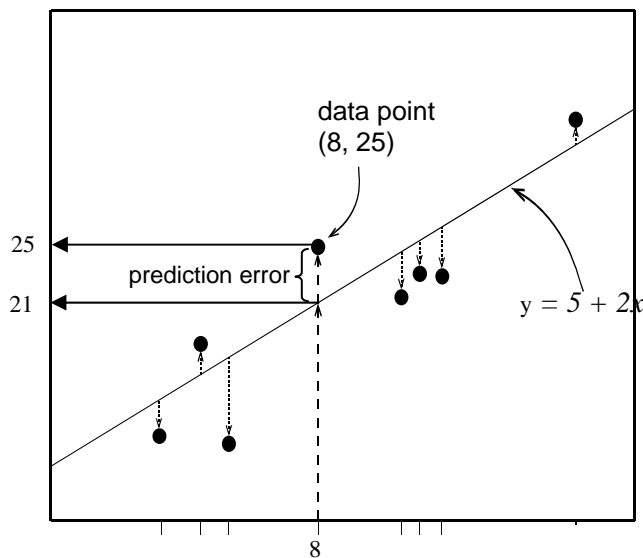
- <http://www.niwa.cri.nz/edu/resources/climate>
- <http://www.cia.gov/cia/publications/factbook>
- <http://www.stats.govt.nz>
- <http://www.censusatschool.org.nz>
- [http://www.amstat.org/publications/jse/jse data archive.html](http://www.amstat.org/publications/jse/jse%20data%20archive.html)

Regression

(The theory on this page is **not** required for this unit, but helps explain what calculations are done.)

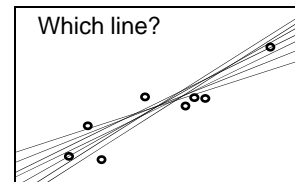
Regression relationship = trend + scatter

Observed value = predicted value + prediction error



The Least Squares Regression Line

Choose the line with smallest sum of squared prediction errors.



Minimise the sum of squared prediction errors

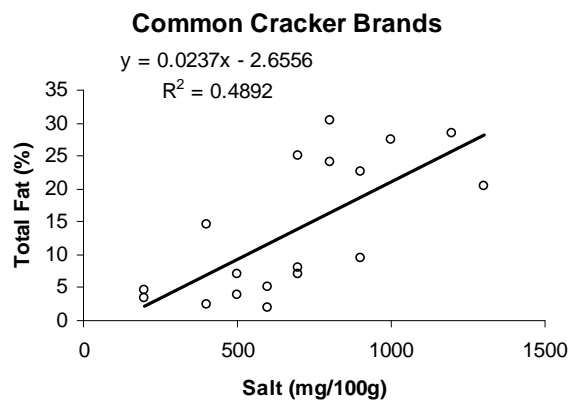
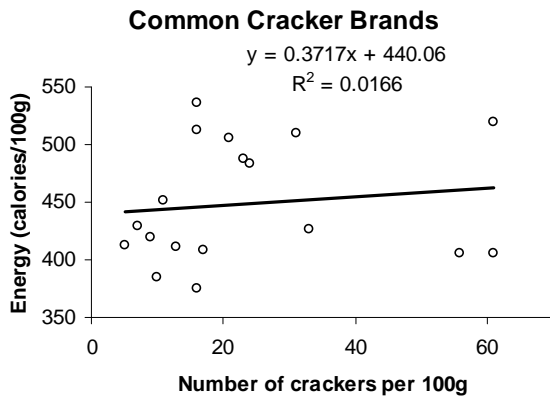
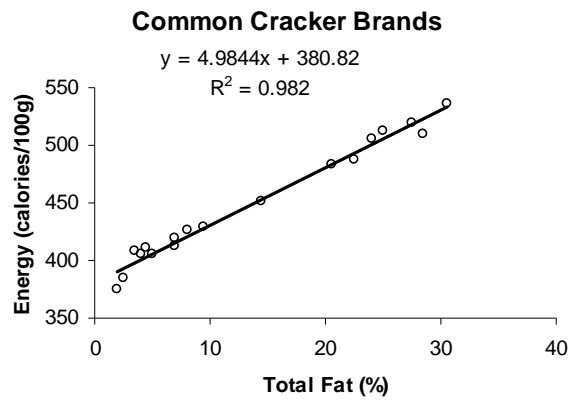
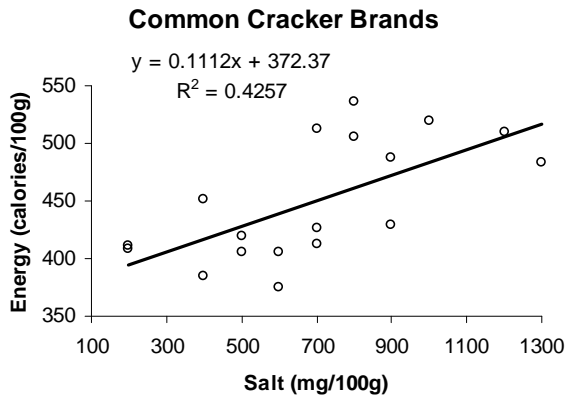
$$\text{Minimise } \sum (\text{prediction errors})^2$$

- There is one and only one least squares regression line for every linear regression
- The **sum** of the prediction errors is zero for the least squares line but it is also true for many other lines
- The line includes the means of x and y i.e. the point (\bar{x}, \bar{y}) is on the least squares line
- Calculator or computer gives the equation of the least squares line

R-squared (R^2)

On a scatter plot *Excel* has options for displaying the equation of the fitted line and the value of R^2 .

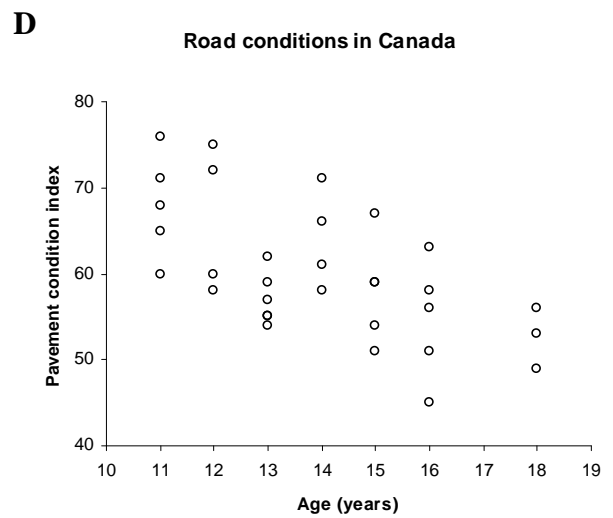
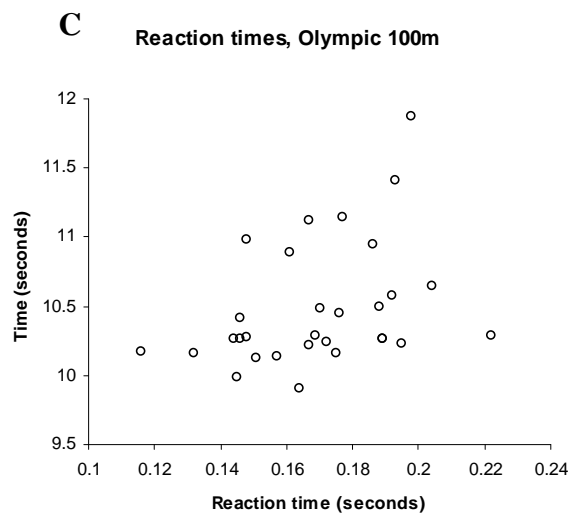
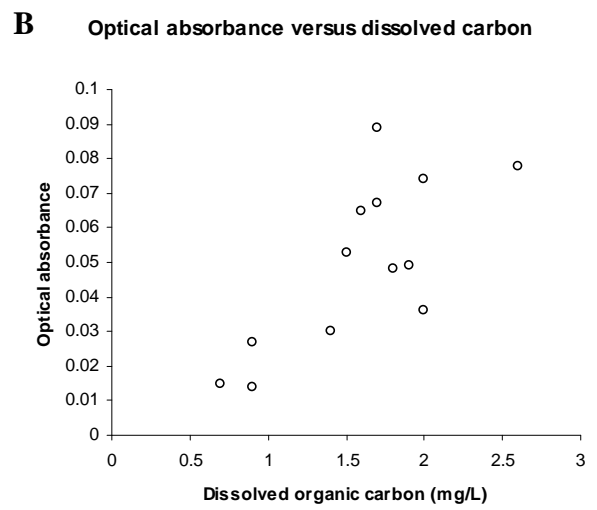
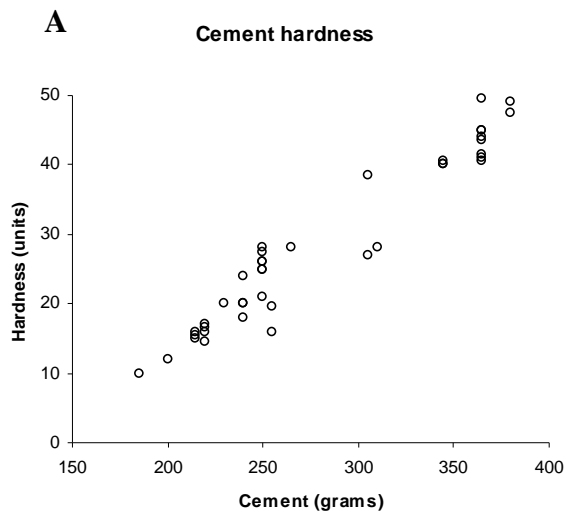
Four scatter plots with fitted lines are shown below. The equation of the fitted line and the value of R^2 are given for each plot. Compare the values for R^2 to the scatter seen.



- R^2 gives the fraction of the variability of the y values accounted for by the linear regression (considering the variability in the x values).
- R^2 is often expressed as a percentage.
- If the assumptions (straightness of line) appear to be satisfied then R^2 gives an overall measure of how successful the regression is in linearly relating y to x .
- R^2 lies from 0 to 1 (0% to 100%).
- The smaller the scatter about the regression line the larger the value of R^2 .
- Therefore the larger the value of R^2 the greater the faith we have in any estimates using the equation of the regression line.
- R^2 is the square of the sample correlation coefficient, r .
- For the above example, the linear regression accounts for 86.6% of the variability in the y values from the variability in the x values.

Exercise:

List the plots from greatest R^2 to least R^2 .



Greatest to least R^2 :

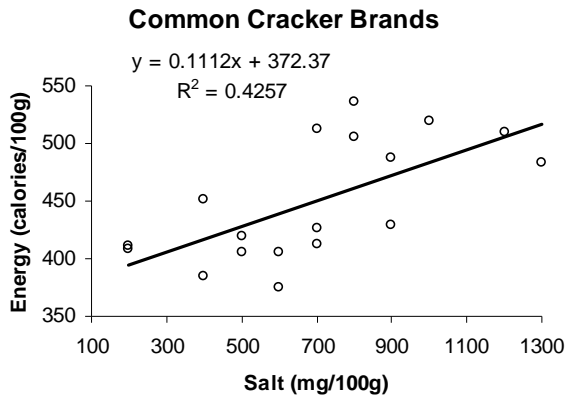
A, B, D, C

(Note: r is negative for D, but R^2 is positive)

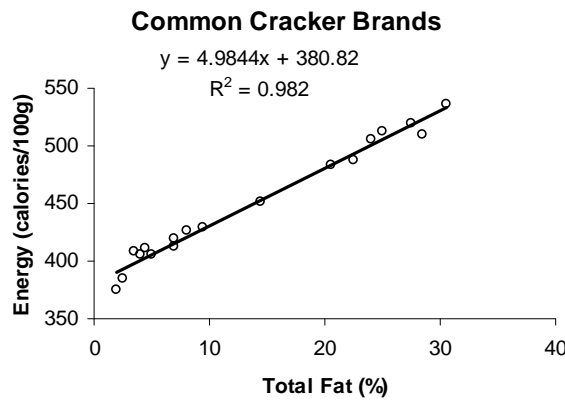
Source: *Chance Encounters: A First Course in Data Analysis and Inference* by Christopher J. Wild and George A. F. Seber

Exercise:

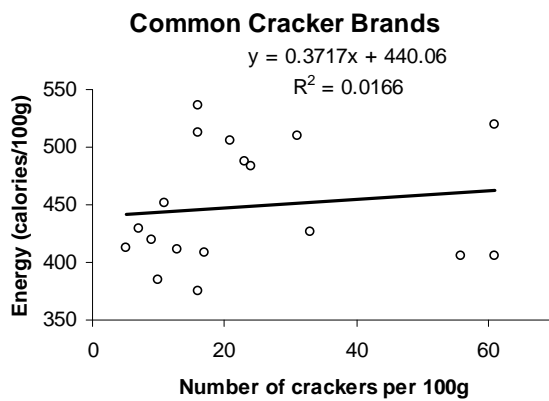
for each scatter plot, use the value of R^2 to write a sentence about the variability of the y-values accounted for by the linear regression.



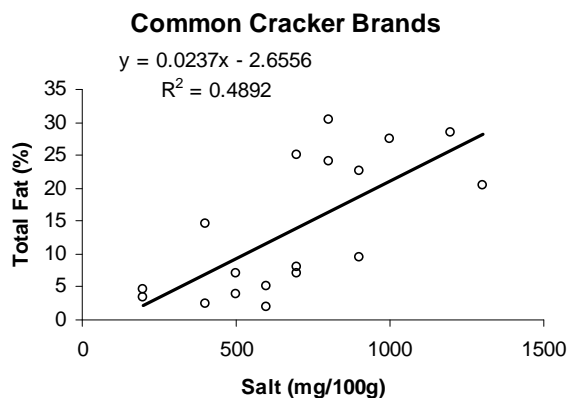
The linear regression accounts for 42.6% of the variability in the energy values from the variability in the salt content values.



The linear regression accounts for 98.2% of the variability in the energy values from the variability in the percentage total fat values.



The linear regression accounts for 1.7% of the variability in the energy values from the variability in the number of crackers per 100g.



The linear regression accounts for 48.9% of the variability in the percentage total fat values from the variability in the salt content values.

Outliers in a regression context

An outlier, in a regression context, is a point that is unusually far from the trend.

The effect can be quite different for outliers in the y dimension compared to the x dimension.

Outliers in y

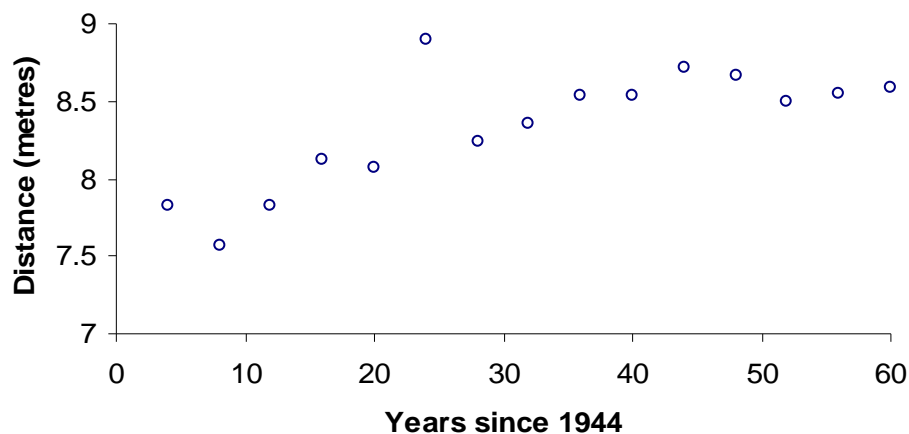
This is when the data point is a great distance above or below the trend line. An example:

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/trackandfield.asp

**Men's Long Jump Winning Distances,
Olympic Games, 1948-2004**



Exercise:

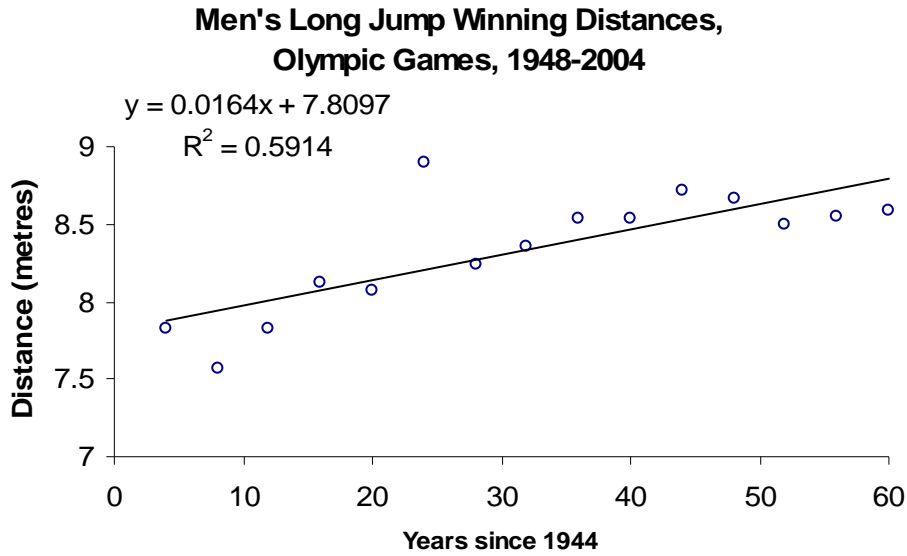
Comment on the scatter plot.

**The data suggests a linear trend. (Alternative: The data suggests a trend with a slight curve.)
Positive association.**

The data suggests constant scatter.

There appears to be a reasonably strong relationship with one y-outlier.

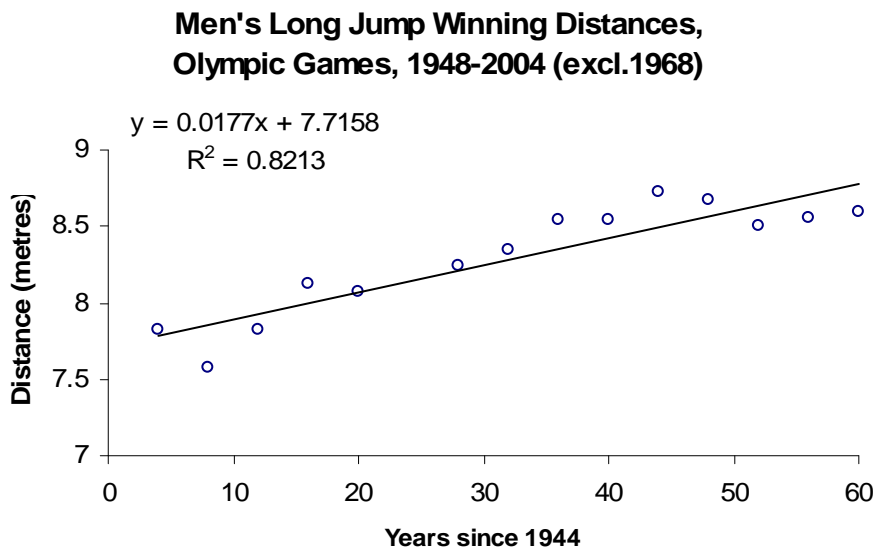
Excel output for a linear regression on all 15 observations



We estimate that for every 4-year increase in years (from one Olympic Games to the next) the winning distance increases by **0.0656m** , on average.

Using this linear regression we predict that the winning distance in 2004 will be **8.79m** .

Excel output for a linear regression on 14 observations (with the 1968 observation removed)



We estimate that for every 4-year increase in years (from one Olympic Games to the next) the winning distance increases by **0.0708m** on average.

Using this linear regression we predict that the winning distance in 2004 will be **8.78m** .

What effect did the 1968 observation have on the:

a) fitted line?

The fitted line was pulled towards the outlier by a small amount.

b) predicted winning distance in 2004?

Not much effect (a difference of only 1cm).

c) value of r ?

When the 1968 observation was removed R^2 increased from 0.59 to 0.82, so r went from 0.77 to 0.91.

So what does this mean, altogether?

Removing the outlier does not affect our prediction much but does increase the firmness of our prediction, because the r value is much higher.

We see how an outlier in y can affect the R^2 value a lot although not necessarily the trend line itself.

Such 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.
- 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 line and discuss the differences.

Outliers in x (or x -outliers)

The effect of an outlier that is distant along the x axis can be quite different. An example:

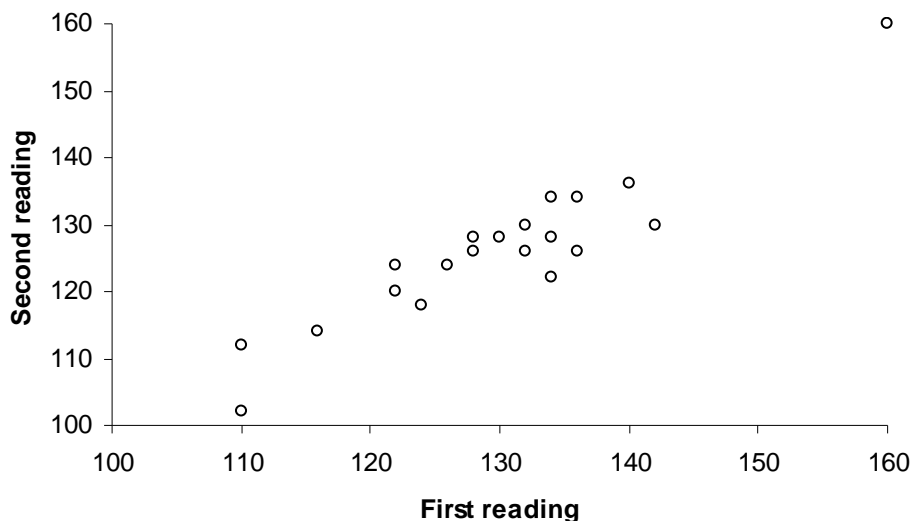
We often talk about a person's "blood pressure" as though it is an inherent characteristic of that person. In fact, a person's blood pressure is different each time you measure it. One thing it reacts to is stress. The following table gives two systolic blood pressure readings for each of 20 people sampled from those participating in a large study. The first was taken five minutes after they came in for the interview, and the second some time later.

Note: The systolic phase of the heartbeat is when the heart contracts and drives the blood out.

Source: *Chance Encounters: A First Course in Data Analysis and Inference* by Christopher J. Wild and George A. F. Seber (Exercise for Section 3.1.2., Question 3, p113).

Observation	1	2	3	4	5	6	7	8	9	10
1st reading	116	122	136	132	128	124	110	110	128	126
2nd reading	114	120	134	126	128	118	112	102	126	124
Observation	11	12	13	14	15	16	17	18	19	20
1st reading	130	122	134	132	136	142	134	140	134	160
2nd reading	128	124	122	130	126	130	128	136	134	160

Blood Pressure



Exercise:

Comment on the scatter plot.

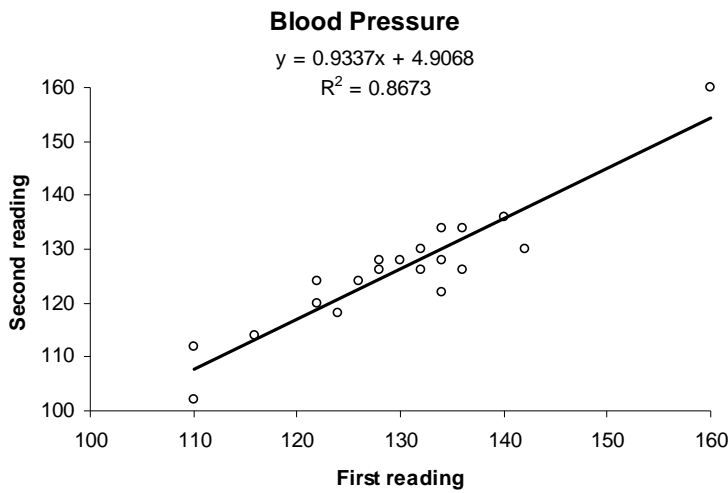
The data suggests a linear trend.

Positive association.

The data suggests constant scatter.

There appears to be a reasonably strong relationship with one observation having a much higher first and second reading than the other observations.

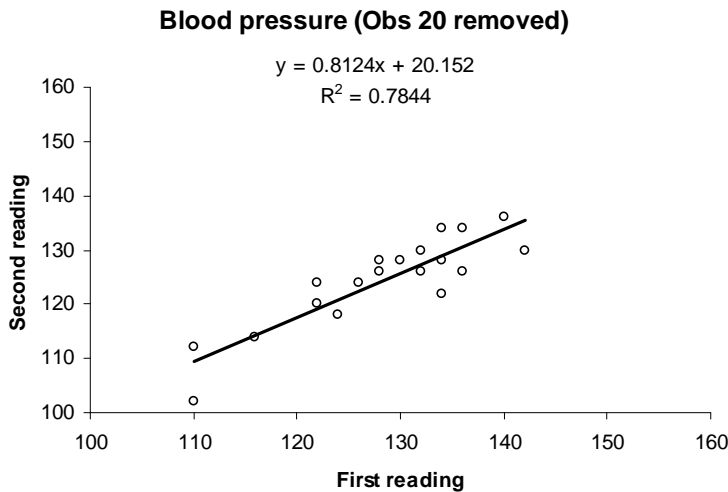
Excel output for a linear regression on all 20 observations



We estimate that for every 10-unit increase in the first blood pressure reading the second reading increases by **9.337 units**, on average.

For a person with a first reading of 140 units we predict that the second reading will be **135.6 units**.

Excel output for a linear regression on 19 observations (#20 removed)



We estimate that for every 10-unit increase in the first blood pressure reading the second reading increases by **8.124 units**, on average.

For a person with a first reading of 140 units we predict that the second reading will be **133.9 units**.

What effect did observation 20 have on:

a) the fitted line?

The fitted line was pulled towards observation 20, increasing the slope.

b) the predicted second reading (for a first reading of 140)?

When observation 20 is included the prediction is slightly higher.

c) the value of R^2 ?

When observation 20 is removed R^2 decreased from 0.87 to 0.78

So what does this mean overall?

Our original prediction was likely both wrong and underestimated the error.

We see that an extreme x value can have the effect of artificially making the R^2 value too high.

The fitted line may say more about the x -outlier than about the overall relationship between the two variables. Such an outlier is sometimes called a high-leverage point.

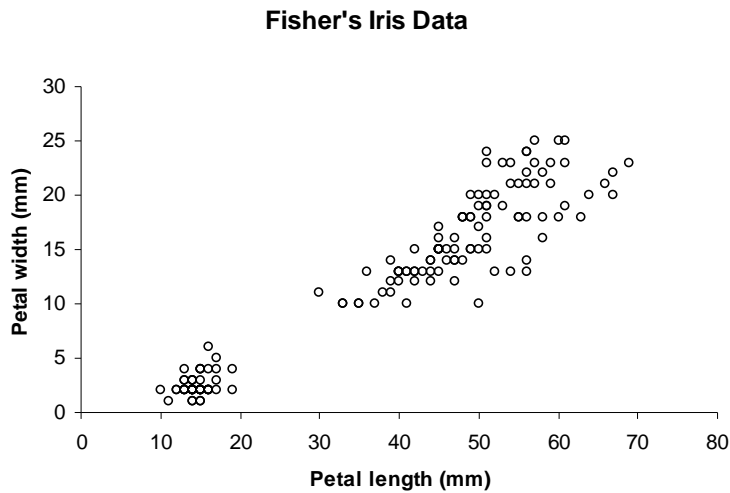
- If it is a mistake then it should either be corrected or removed.
- If it is an actual unusual observation then try to understand why it is so different from the other observations.

If a data set has an x -outlier which does not appear to be in error then carry out two linear regressions; one with the x -outlier included and one with the outlier excluded. Investigate the amount of influence the outlier has on the fitted line and discuss the differences.

Groupings

In the 1930s Dr. Edgar Anderson collected data on 150 iris specimens. This data set was published in 1936 by R. A. Fisher, the well-known British statistician.

This is sourced from: <http://lib.stat.cmu.edu/DASL/Stories/Fisher'sIris.html>



Exercise:

Comment on the scatter plots.

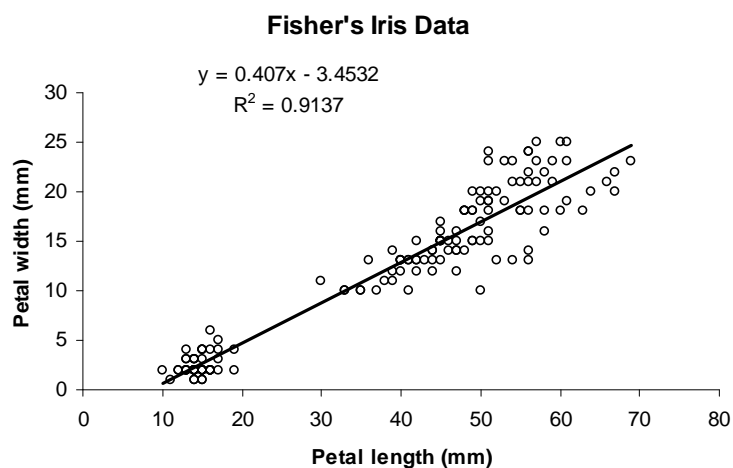
The data suggests a linear trend.

Positive association.

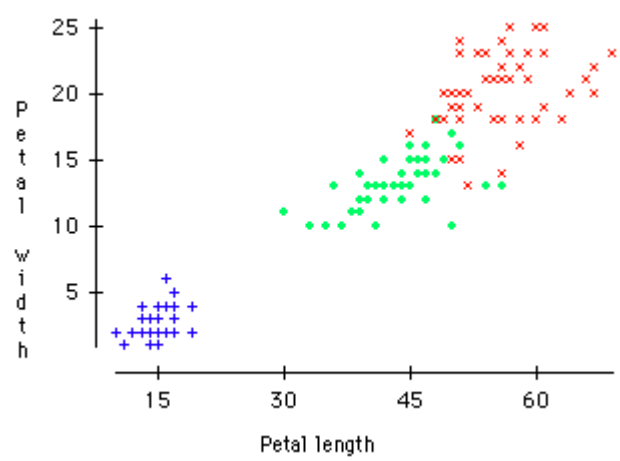
The data suggests non-constant scatter.

Moderate relationship.

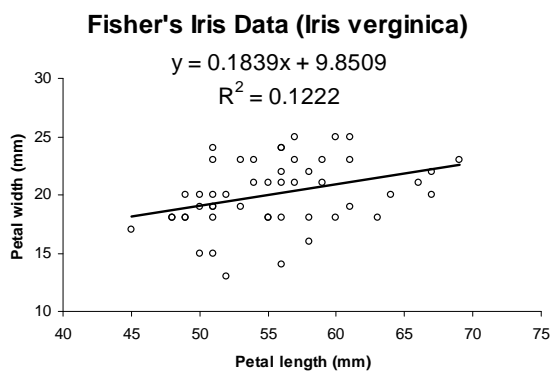
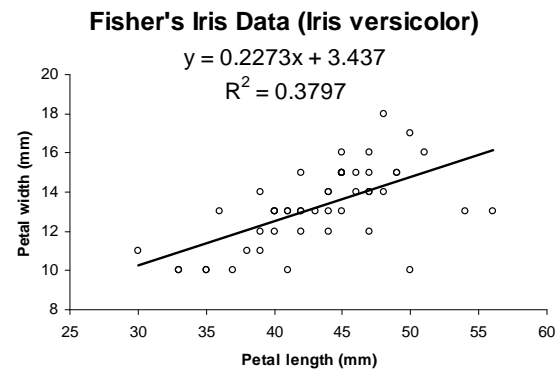
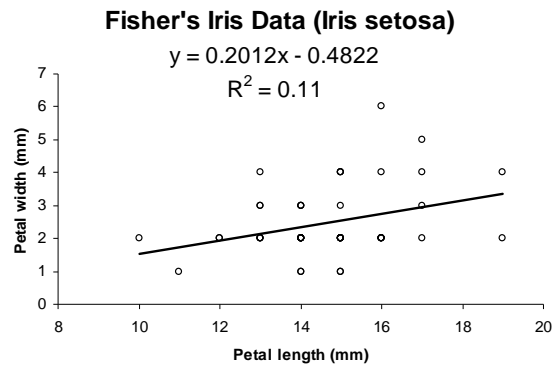
Appears to be two groupings.



The data were actually on fifty iris specimens from each of three species; *Iris setosa*, *Iris versicolor* and *Iris virginica*. The scatter plot below identifies the different species by using different plotting symbols (+ for *setosa*, • for *versicolor*, × for *virginica*).



Let's see what happens when we look at the groups separately.



Exercise:

Comment.

The equations of the 3 fitted lines are quite different.

R^2 is quite small in all 3 cases – partly as a result of smaller samples (50 compared to 150) and partly by the reduced range for the x-values.

Watch for different groupings in your data.

- If there are groupings in your data that behave differently then consider fitting a different linear regression line for each grouping.

Conclusions about R^2 and outliers/groupings.

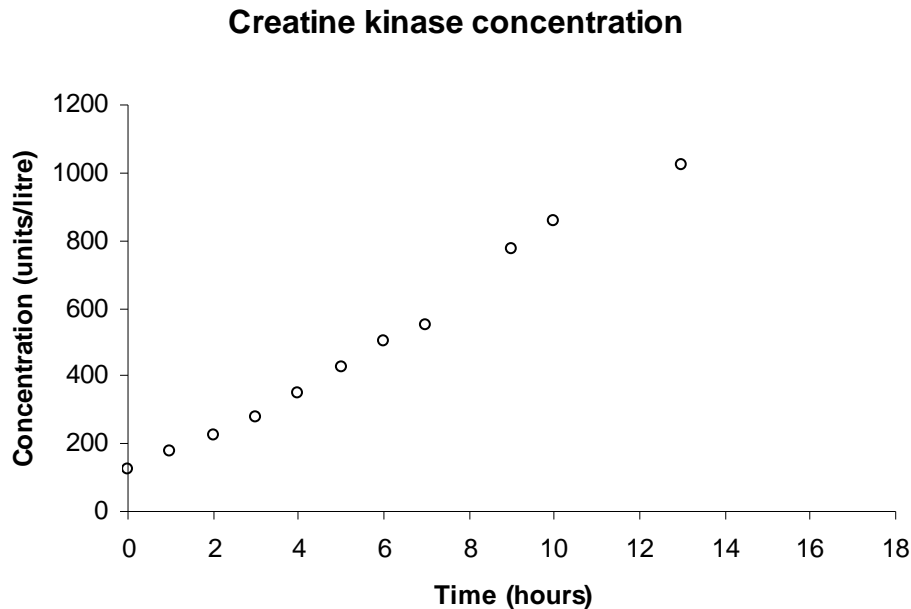
- A large value of R^2 does not mean the linear regression is appropriate.
- An x -outlier or data that has groupings can make the value of R^2 seem large when the linear regression is just not appropriate.
- On the other hand, a low value of R^2 may be caused by the presence of a single y -outlier and all other points have a reasonably strong linear relationship.
- Groupings of different kinds of objects may give a value of R^2 that does not hold for the individual types.

Prediction

The purpose of a lot of regression analyses is to make 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.



Exercise:

Comment on the scatter plot.

The data suggests a linear trend.

Positive association.

The data suggests constant scatter.

Appears to be a strong relationship.

Suppose that a patient had a heart attack 17 hours ago. Predict the creatine kinase concentration in the blood for this patient.

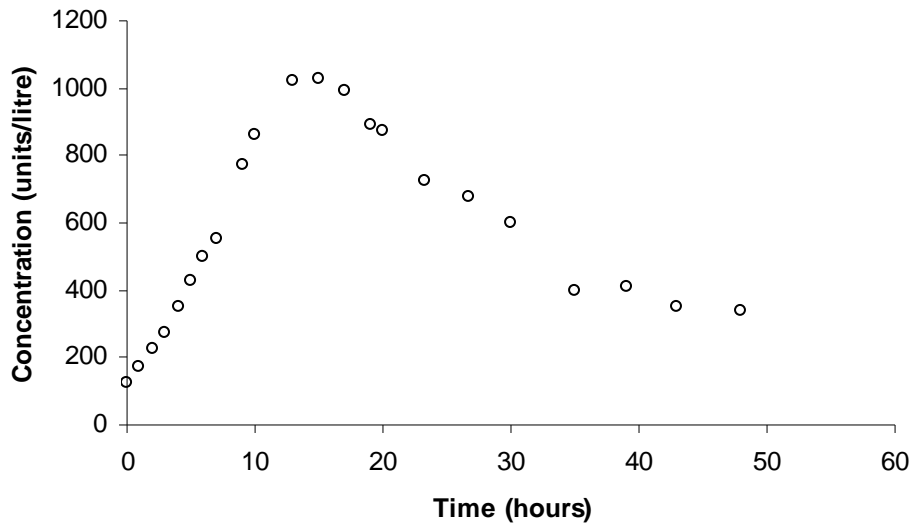
The regression line would suggest about 1200 units/litre.

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

This value is much lower than that predicted by the fitted line. The random error expected from a graph with that high a correlation would be very low, suggesting the model of a straight line is at fault (rather than this being a random fluctuation).

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

Creatine kinase concentration



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.

The removal of an x outlier will mean that the range of observed x values is reduced. This should be discussed in the comparison between the two linear regressions (x outlier included and x outlier excluded). It is possible that the supposed "outlier" may actually indicate the start of a change in the pattern.

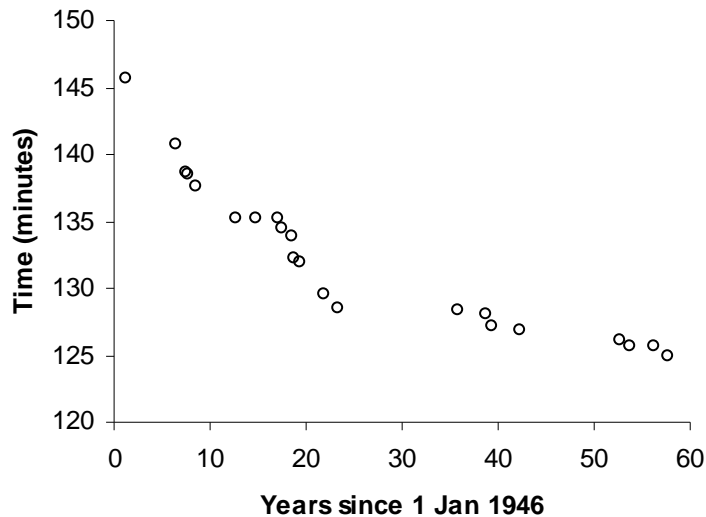
Non-Linearity

Sometimes a non-linear model is more appropriate. An example:

The data in the scatter plot below shows the progression of the fastest times for the men's marathon since the Second World War. We may want to use this data to predict the fastest time at 1 January 2010 (i.e. 64 years after 1 January 1946).

Source: <http://www.athletix.org/>

Men's Marathon Fastest Times



Exercise:

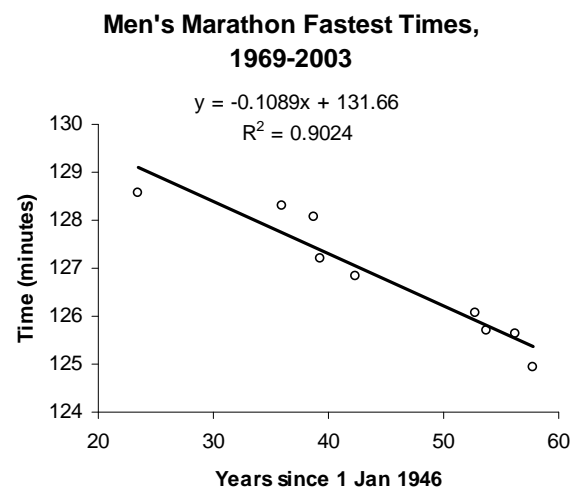
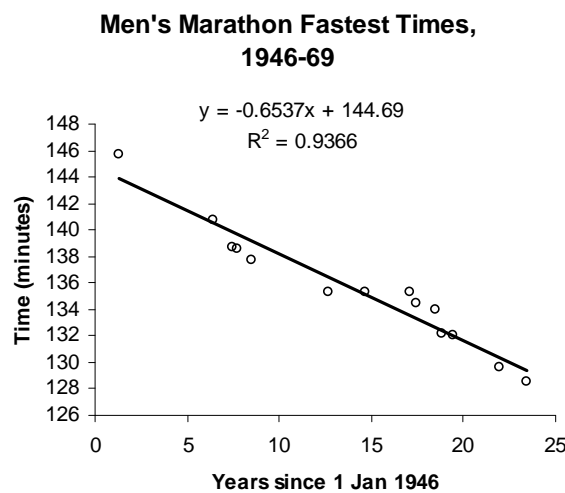
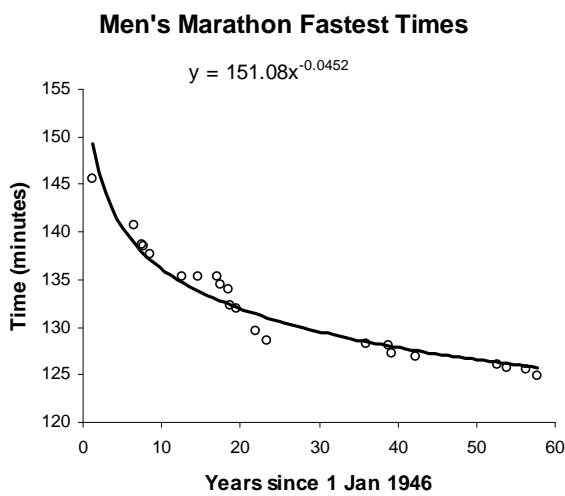
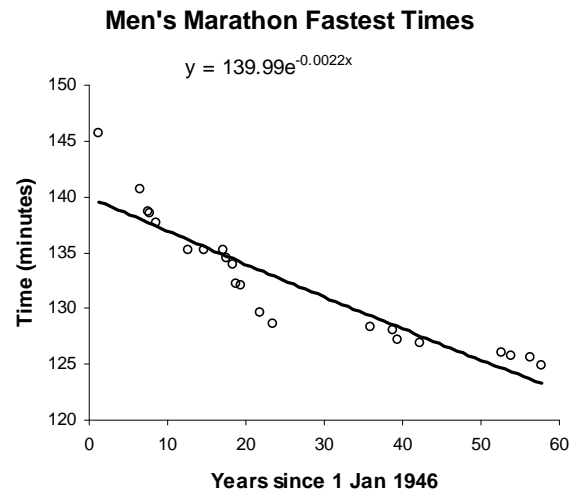
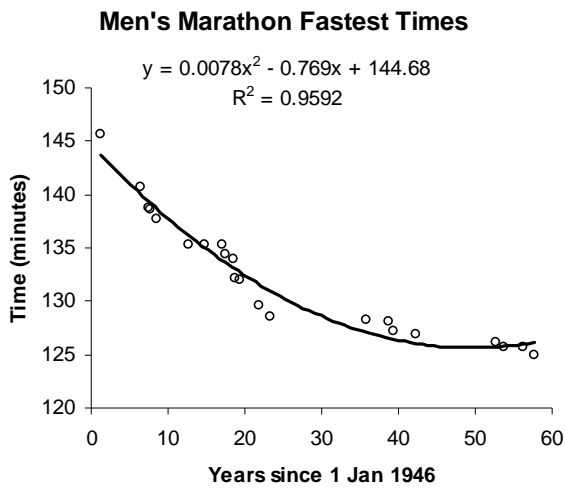
Concerns:

Non-linearity

Possible solutions:

Try fitting:

- 1. an exponential function ($y = ae^{bx}$)**
- 2. a power function ($y = ax^b$)**
- 3. two separate straight lines: one for say 0–23 years and one for say 23–60 years**
- 4. a line for only the later years, say 23–60 years**
- 5. a quadratic ($y = ax^2 + bx + c$)**



Exercise:

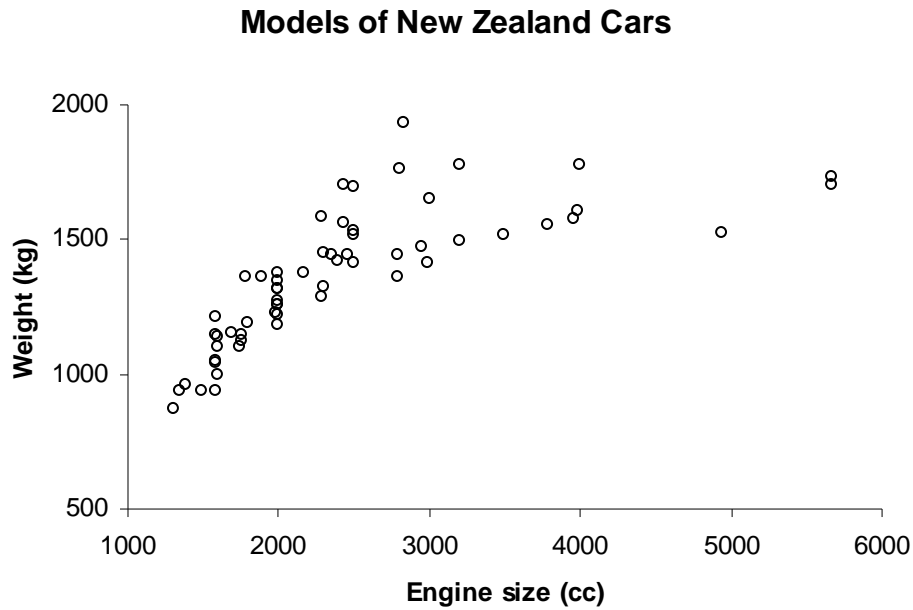
Comments:

Power and piecewise line both give a reasonable fit

Exponential and a single line both give a bad fit

Quadratic: time starts increasing (which is not sensible, and a typical problem to polynomial fit)

The data in the scatter plot below comes from a random sample of 60 models of new cars taken from all models on the market in New Zealand in May 2000. We want to use the engine size to predict the weight of a car.



Exercise:

Concerns:

Non-linearity

Increasing spread in y values as x increases, so that any prediction based on a line might be right for small engines but is unlikely to be correct for large ones.

Possible solutions:

Seems to be linear for engine sizes less than 2500cc.

Very weak or no linear relationship for engine sizes over 2500cc.

Solution: Fit a line for engine sizes less than 2500cc.

Note: The solution need not be to exclude all linear models. It might be to restrict the range of values which the linear model is applied to.