Statistics (Term 1)

"There are three kinds of lies:

lies, damned lies, and statistics."

Benjamin Disraeli (1804 -1881),
Mark Twain (1835 -1910)

Course Overview

Course Info

Lecturer	Dr Ian R Vernon
Email	i.r.vernon@durham.ac.uk
Office	CM317
Lectures	Tue 17:00 PH8 & Thur 16:00 PH8
Homework	Weekly
Tutorial Classes	Fortnightly starting next week
Computer Practical Classes	Weekly starting next week
Office Hour	TBA
Webpage	DUO
Notes by	Dr Ian R Vernon

Course Outline

- Design of Experiments
- Descriptive Statistics
- Bivariate Data
- Prediction from linear association: simple linear regression
- Multiple regression
- Exploring differences between many groups
- Exploring differences between two factors

1 Design of Experiments

An Example Experiment

Consider the following passage, extracted from an article contained in the Guardian newspaper, September 29th 1994, which exhibits many of the features which we shall address.

> In 1988, what was intended to be an eight-year, 22,000-person trial of the effectiveness of aspirin was brought to a sudden halt. It was the usual, classic, double blind sort of trial: 11,000 US physicians took 325 milligrams of acetylsalicylic acid every day, and the other 11,000 took a placebo. No-one knew which he or she was taking. Four years later, five of the aspirin takers had died of cardiac infarction, and 99 had had non-fatal heart attacks. But 18 of the placebo group were dead, and 171 had been ill with infarctions. That meant that regular doses of the stuff had lowered the risk by 47%.
>
> These results were so unambiguous and so unexpected that the trial was halted half-way through. With results like that, there didn't seem to be any point in risking the health of the control group. Three-quarters of the doctors in the placebo group immediately switched to taking aspirin, and 99% of the others announced that they would continue to reach for the stuff.

Controlled experiments

- We carry out experiments to examine the effects of different *treatments* on a *response*.
- Many books use other words, especially *plot* for agricultural experiments, instead of response.
- Experiments such as these are *comparative* (or *relative*) and we will concentrate on this type of experiment.
- In general, we will be comparing a treatment effect to the effect with no treatment, or the *control*.
- This type of experiment is called a *controlled experiment*.
- We measure the *response* for the treatment for a number of *subjects*, which then form the *treatment group*; and we measure the response for the control for another group of subjects, which constitute the *control group*.
- Typically we measure effects on several subjects. This is because of a statistical truism: the more measurements you make, the more accurate your results.
- A simple rule-of-thumb is that accuracy is proportional to the square-root of the number of measurements.
- Before we apply the treatment, we try to ensure that the treatment group and the control group are so similar that any difference between them at the end of the experiment is due solely to the treatment.

Confounding

- We must be constantly on guard for factors whose effects are *confounded* (mixed in) with those of the treatment.
- How do we ensure that our treatment group and control group are broadly similar beforehand?
- One method is the *randomised* controlled experiment, where subjects are assigned at random to one of the groups.
- A simple random sample (SRS) of size n is a set of n individuals chosen in such a way that every set of n individuals in the population has an equal chance to be the sample actually chosen.
- A key issue is whether the sample chosen can be considered to be a SRS of the population actually chosen.
- Randomisation does not guarantee similarity of the groups. Confounding because of a *lurking variable* remains possible, but becomes very unlikely as we make measurements on more subjects.
- If something like this ever does happen, we might notice and make corrections or issue interpretational caveats.
- Many experiments do mislead when we fail to notice such shortcomings. We cannot guarantee that experiments are free from confounding.
- We can frequently avoid problems with confounding by *stratifying* beforehand according to factors deemed potentially highly correlated with differences in response.
- To select a stratified random sample, divide a population into groups of similar individuals (strata) and then take a SRS within each stratum.

Placebo effect

- It has become clear over the years that the so-called *placebo* effect is very real, and so we usually try to take it into account by administering a placebo to subjects in the control group.
- Notice that this is one aspect of trying to ensure that the treatment and control groups are dealt with similarly.
- The experiment is *blind* whenever the subjects do not know the group to which they belong.
- The experiment is *double-blind* when both the subject *and* those taking measurements for the experiment do not know the group to which the subject belongs. Usually, this implies that the trial statistician must carry out the randomised designation of subjects to the different groups.

Table 1: Clofibrate trial for treatment of Cholesterol

	Clofibrate		Placebo	
	Number	Deaths	Number	Deaths
Adherers	708	15%	1813	15%
Nonadherers	357	25%	882	28%
Total	1103	20%	2789	21%

Clofibrate trial for Cholesterol

- Considering the Clofibrate drug-taking group alone, it seems as though people who continue to take Clofibrate (the adherers) have lower mortality rate, 15% to 25% deaths.
- However, once we see the corresponding figures for the placebo group, we realise that differences in death rate are not due to the drug (death rate of 15% for both placebo and drug), but perhaps to differences between adherers and non-adherers.
- (We have no clue to what these differences are, or as to why the death rates should differ.)

Interpretation of differences

- Suppose that we have found a difference between the treatment group and the control group. How can we tell whether the difference is real?
- This depends upon our making an assumption about the state of the world as follows: we *assume* that the only differences between the treatment and control groups arise *purely by chance*.
- Under this scenario, we can determine the probability that, by chance, we get a difference as large as the one we see. If the probability that we calculate is small, we conclude that the difference is genuine.
- The calculations and theory required for such *inference* is deferred to lectures on *hypothesis testing*, second term of the Statistics course.

Historical controls

- Randomised experiments are expensive to organise and carry out, and so some experimenters prefer to use *historical* controls.
- The problem here is that the treatment group and the historical control group may be dissimilar in unknown ways, and that we wrongly attribute differences to a treatment effect rather than to underlying dissimilarity of the groups.

Observational Studies

- We can carry out observational studies when it is impractical to undertake a controlled experiment.
- In this type of study we collect subjects into two groups, the *treatment* group and the *control* group as before, where subjects in the treatment group all possess the property we are studying, and where none of the subjects in the control group possess this property.
- Because we are unable to randomize the study, we must especially beware the potential confounding and bias that might arise. In particular if we see a difference between the two groups, this is evidence of *association*, but not of *causation*.
- Sometimes a difference between the treatment and control groups is due not to the treatment, but to some underlying factor which essentially defined the two groups.
- As for controlled experiments, a principal way of avoiding confounding is stratification: we try to make the treatment and control groups as *homogeneous* as possible by taking into account explicitly the confounding factors that we know about

Sex Bias in Postgrad Admissions

Table 2: Sex bias in postgraduate admissions, UC Berkeley

	men		women		percentage
Course	number of applicants	percentage admitted	number of applicants	percentage admitted	women applicants
A	825	62%	108	82%	12%
B	560	63%	25	68%	4%
C	325	37%	593	34%	64%
D	417	33%	375	35%	47%
E	191	28%	393	24%	67%
F	373	6%	341	7%	48%
All	2691	44%	1835	30%	41%

Kidney Stone Example

The following data are extracted from Charig CR, Webb DR, Payne SR, Wickham OE. *Comparison of treatment of renal calculi by operative surgery, percutaneous nephrolithotomy and extracorporeal shock wave lithotripsy.* BMJ; 1986: 292, 879-882.

Intervention	Success	Failure	Total	Success Rate
Open surgery	273	77	350	78%
Percutaneous nephrolithotomy	289	61	350	83%

- The authors compare historical series (1972-1980 for open surgery, 1980-1985 for percutaneous nephrolithotomy) of success rates in removing kidney stones.
- There is small but useful increase in success rate for percutaneous nephrolithotomy), compared to open surgery, and the former is a cheaper intervention.
- How do you interpret this data?

Kidney Stone Data in More Detail

Stones smaller than 2cm diameter

Intervention	Success	Failure	Total	Success Rate
Open surgery	81	6	87	93%
Percutaneous nephrolithotomy	234	36	270	87%

Stones larger than 2cm diameter

Intervention	Success	Failure	Total	Success Rate
Open surgery	192	71	263	73%
Percutaneous nephrolithotomy	55	25	80	69%

1.1 Summary

Summary

- Controlled experiments are preferable to experiments without controls. Historical controls can be very misleading.
- Where possible, subjects should be assigned randomly to treatment and control groups, so as to avoid confounding and bias. Otherwise, for observational studies, care must be taken to guard against confounding.
- Double-blind experiments are preferred to blind or non-blind experiments in that they guard against bias.
- In the following table 1.1, we see that only 4 out of 51 experiments were randomised, and these tended to show that the new technique was not very different from other techniques, 32 of the experiments contained no controls at all, and the remaining 15 contained unrandomised controls.
- These 47 experiments tended to be (most likely spuriously) strongly in favour of the new technique.

Classification of 51 Surgical technique experiments:

Experimental design	Efficacy of new technique		
	much better	better	no better
No controls	24	7	1
Controls, not randomised	10	3	2
Randomised controlled	0	1	3

2 Descriptive Statistics

2.1 Variables and measurements

- Variables, and their values, are the objects of fundamental study in Statistics. Together, they constitute data, and the usual goal of Statistics is to gain understanding from data.
- Characteristics of individuals, in particular, values attached to them, are called *variables*.
- For example, Cities have populations, people have height, planets have mass, bushes have leaves, and so forth. For example, the population of Durham is about 75,000, Venus weighs about 1 Earth, and so forth.
- Variables can be *categorical* or *quantitative*. Any variable that we measure in terms of numbers is a quantitative variable. Variables which we measure in terms of words are categorical (synonym - qualitative).
- For example we measure gender as being either "male" or "female"; and we measure houses as being "terraced", "semi-detached", and so on.
- Variables can be *discrete* or *continuous*. *Continuous* variables are those variables which could be measured arbitrarily finely on some scale.
- For example, the weather temperature in degrees celsius.
- *Discrete* variables are those which take only a finite number of values
- For example the number of christmas cards you receive this year. Sometimes the distinction is obscure: discrete variables can have so many possible values that they resemble continuous variables; and continuous variables are frequently measured in discrete ways.
- Clearly, all categorical variables are also discrete, whereas quantitative variables can be either discrete or continuous. We will be working mostly with quantitative variables.
- Determination of values of variables requires that they be measured in some way. We will need to remember what units of measurement are being used. Remember that *coding* the data is a valuable way of easing the computational burden.
- For example average: 27003, 27012, 27015. The answer is clearly $27000 + 30/3 = 27010$.
- The collection of individuals concerned is called the *population*, which may be infinite, finite, or hypothetical.
- The totality of the values of the variable for all possible individuals (i.e. the population) is called the *distribution* of the variable. This recognises the fact of life that characteristics vary from individual to individual. Much of statistics aims at summarising the shapes of these distributions.
- Usually we see the values for a variable for a small *sample* of the population. It is u7seful to portray the distribution of the sample values yo convey its *shape*.

2.2 Stem and leaf plots

- We now come to ways of portraying the shapes of distributions. For this we use histograms, and the first kind of histogram that we discuss is the *stem and leaf plot* (synonym - stemplot).
- The stem and leaf plot is valuable not only in forming a simple histogram of the data, but also by retaining the data values.
- Each row is called a stem; the digits to the left are called stem labels, and the digits to the right are called leaves.
- As an example, suppose that you tell me the age of your next of kin. I will form a stem and leaf plot as you give me the data. Do a back to back plot for males and females, start with me (51). Have ages from 10-90.
- For our next example, consider the data in table 4. It shows 100 measurements of the lengths of cuckoo eggs. We can form a stem and leaf plot for this data too, but for the stem labels we take the values 19, 19.5, 20, ..., rather than the integers 19, 20, 21, ..., so that the plot we get is more spread out, and more informative. (This is a matter of judgment and expertise.) The plot is shown in table 5.

Table 4: Lengths in millimetres of 100 cuckoo eggs

22.5	20.1	23.3	22.9	23.1	22.0	22.3	23.6	24.7	23.7
24.0	20.4	21.3	22.0	24.2	21.7	21.0	20.1	21.9	21.9
21.7	22.6	20.9	21.6	22.2	22.5	22.2	24.3	22.3	22.6
20.1	22.0	22.8	22.0	22.4	22.3	20.6	22.1	21.9	23.0
22.0	22.0	22.1	22.0	19.6	22.8	22.0	23.4	23.8	23.3
22.5	22.3	21.9	22.0	21.7	23.3	22.2	22.3	22.8	22.9
23.7	22.0	21.9	22.2	24.4	22.7	23.3	24.0	23.6	22.1
21.8	21.1	23.4	23.8	23.3	24.0	23.5	23.2	24.0	22.4
23.9	22.0	23.9	20.9	23.8	25.0	24.0	21.7	23.8	22.8
23.1	23.1	23.5	23.0	23.0	21.8	23.0	23.3	22.4	22.4

- The cuckoo data raises other issues. Presumably the reported lengths are roundings of the actual measurements, i.e. we assume that the reported value of 22.5 represents some measurement between 22.45 and 22.55.
- Our stem and leaf plot subdivides the total range of values into intervals, 19.0 to 19.4, 19.5 to 19.9, and so forth. Such subdivisions are called *class intervals*.
- The *class frequency* is the number of values which fall into a given interval. For example, The class frequency in the class interval 20.0 to 20.4 is 4, and the frequency in the class interval 22.0 to 22.4 is 27.

Table 5: Lengths in millimetres of 100 cuckoo eggs (stem and leaf plot)

19																										
1·	6																									
20	1	4	1	1																						
2·	9	6	9																							
21	3	0	1																							
2·	7	9	9	7	6	9	9	7	9	8	7	8														
22	0	3	0	2	2	3	0	0	4	3	1	0	0	1	0	0	3	0	2	3	0	2	1	4	0	4	4
2·	5	9	6	5	6	8	8	5	8	9	7	8														
23	3	1	0	4	3	3	3	4	3	2	1	1	0	0	0	3										
2·	6	7	8	7	6	8	5	9	9	8	8	5														
24	0	2	3	4	0	0	0	0																		
2·	7																									
25	0																									

2.3 Histograms

The essential purpose of histograms is to allow us to visualise the *shape* of a *distribution*. Most importantly, we convey shape via the *area* of the blocks in the histogram, and not by the *height* of the block. As an example, consider the frequency distribution shown in table 6.

2.3.1 Methodology

- Decide upon the *class intervals*, the term for the different ranges of values for the distribution. Often, this will have been done for you. Here, the smallest class interval is \$1000.
- The *class width* is the width of the class interval.
- It is sometimes worth combining one or more classes if you feel that this would better convey the shape of the distribution.
- Decide what your *endpoint convention* will be. For example, is the left endpoint included in the interval? Is the right endpoint included? In our example, the left endpoint is included, and the right endpoint is excluded. Note that most discrete distributions don't need an endpoint convention; most continuous distributions do.
- Standard notation, for example $0 - 1000^-$, is used to indicate that a left endpoint is included, and that a right endpoint is excluded.
- Decide upon your vertical scale (for the *y-axis*): usually your vertical scale will be the original frequency, or the percentage frequency, or - if the class intervals are unequal - a density. In this example we use a density because the classes are of unequal length.
- Decide what to do with open-ended classes. In this example, we will ignore the 1% of families with incomes exceeding \$50000. It might have been possible arbitrarily to assume that the right endpoint was \$100000, or some other number.

Table 6: Distribution of families by income, USA 1973. Class intervals include the left endpoint, but not the right endpoint.

Income level	Percentage of families	Class width	Percentage per $1000
$0-$1000	1	1	1
$1000-$2000	2	1	2
$2000-$3000	3	1	3
$3000-$4000	4	1	4
$4000-$5000	5	1	5
$5000-$6000	5	1	5
$6000-$7000	5	1	5
$7000-$10000	15	3	5
$10000-$15000	26	5	5.2
$15000-$25000	26	10	2.6
$25000-$50000	8	25	0.32
$50000 and more	1	-	-

- Determine what each block's area should be: remember that we are trying to convey *density* rather than raw frequency. Therefore the height of each block must be chosen so as to guarantee that the area matches the percentage frequency in the class. For example, The $10000-$15000 class is 5 times the smallest class interval width, and the total frequency in the class is 26%. This is, on average, 5.2% for each of the 5 classes. That is, we divide the frequency by the length of the interval.
- When we use the density scale, the total area in the histogram is 100%; and the percentage frequency within any block is the area of the block: i.e. the height of the block multiplied by the length of the interval.
- For the horizontal axis (*x-axis*), decide where to centre the blocks that make up the histogram. For discrete measurements, we usually centre the block at the relevant value. For continuous measurements, there is less to think about: we draw the block over the range of the class (sometimes leaving a small space, so that the histogram appears as a series of bars).
- Label the axes clearly: the histogram conveys no information until both axes have been labelled, and a title given.

2.3.2 Remarks

- The choice of class interval, when this has not been fixed in advance, can make a big difference to the apparent shape revealed. There are ways of handling this. The method used by the R package is the Sturges method, which bases the interval size on an estimate of spread.
- Choice of width and height of scales can make a dramatic difference to perceptions of underlying distributions.

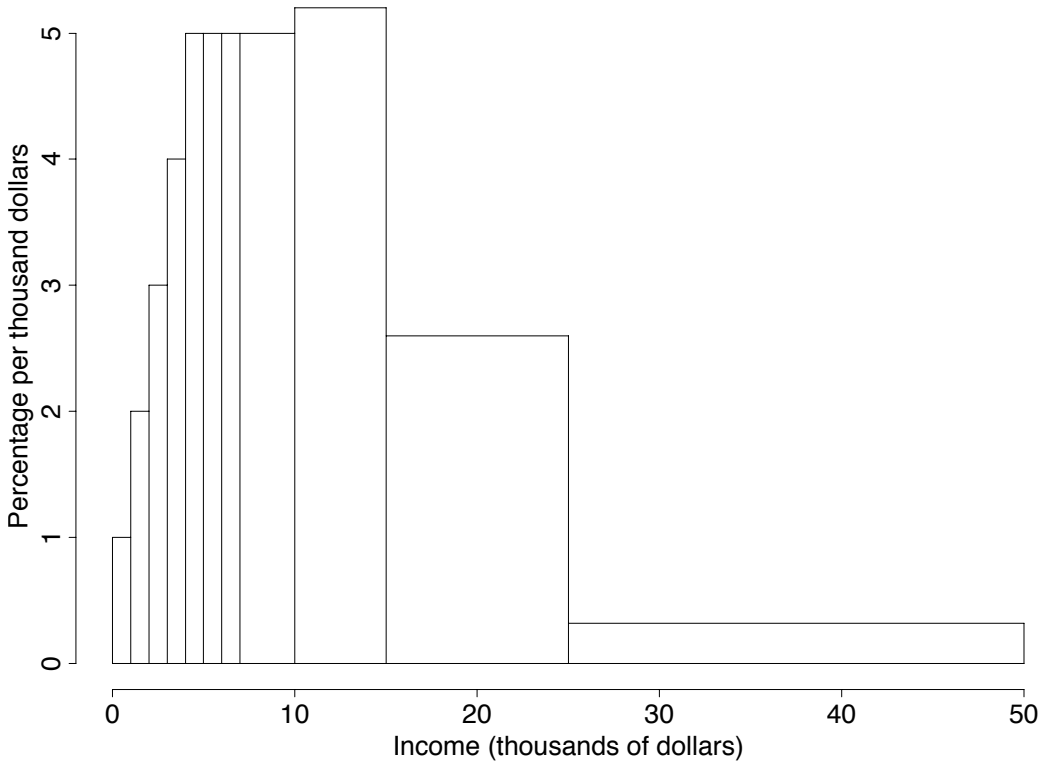
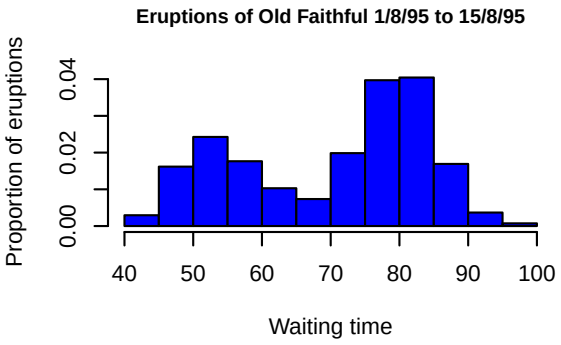
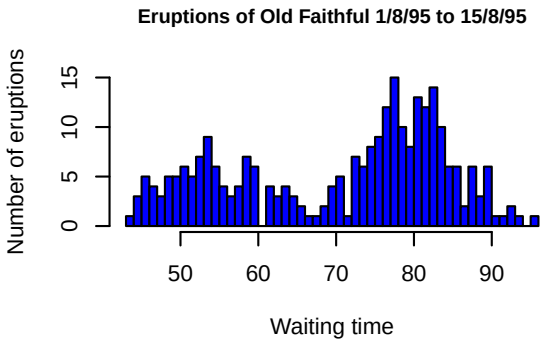
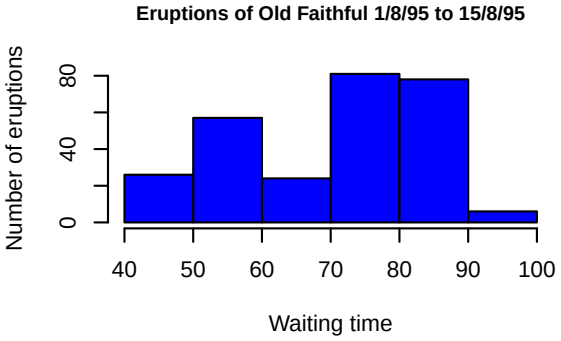


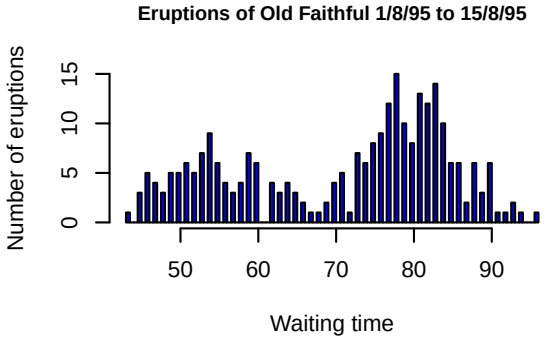
Figure 1: Distribution of families by income, USA 1973

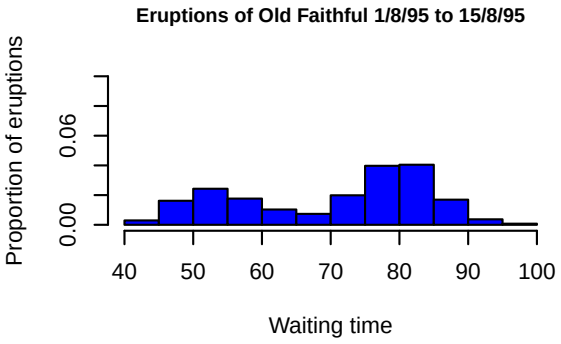
- We can control for a variable by *cross-tabulation*: we separate out entries according to the different possible values of the confounding variable, and tabulate each of these separate distributions.
- For example, we can show a histogram which shows the distributions of singers' heights in inches, classified by voice part. Obviously sex of singer is a confounding factor, but the further classification into voice part is revealing.

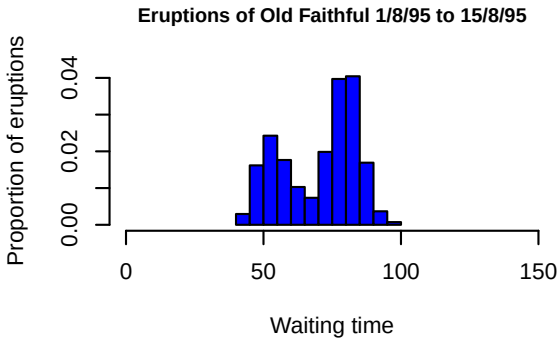


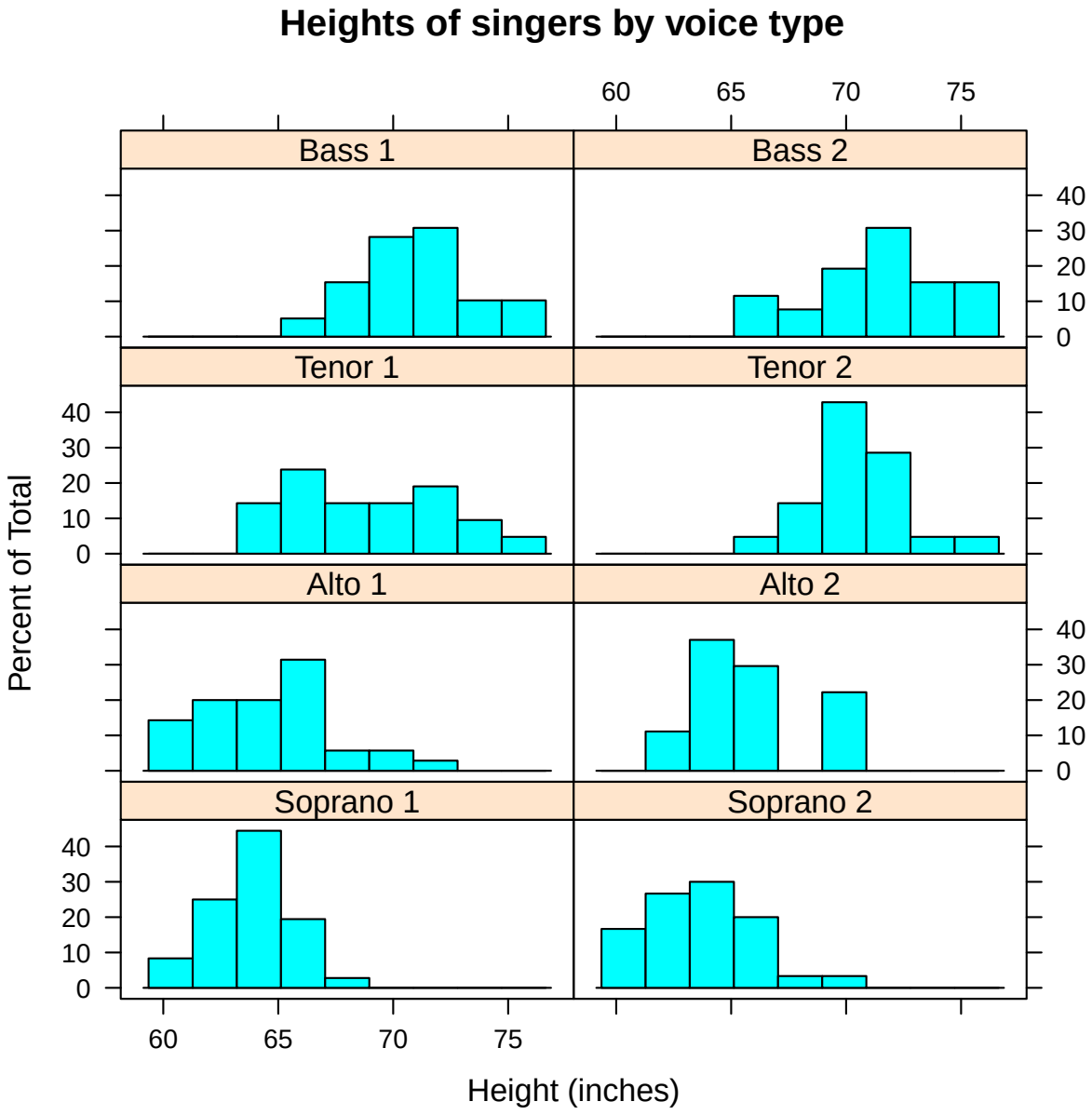











Heights of singers by voice type
Bass 1
Bass 2
Tenor 1
Tenor 2
Alto 1
Alto 2
Soprano 1
Soprano 2
Percent of Total
Height (inches)
0
10
20
30
40
60
65
70
75

2.4 Summarising distributions

Summarising distributions: The Mean

- We have considered the *shape* of a distribution already, and have seen that histograms are a handy way of showing shape. We come now to several *summary statistics*.
  - the *average* and *median* summarise the *centre* of the distribution.
  - the *standard deviation* and *inter-quartile range* summarises *spread around the centre* of the distribution.
- We suppose that there are n values of a variable, and that they are labelled $x_1, x_2, \ldots, x_n$. There is no special significance to the labels, they just help us distinguish between different values.
- The average (synonym: *mean*) of n values is given by

$$\bar{x} \;=\; \frac{x_1 + x_2 + \ldots + x_n}{n} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Remember to take into account the frequency of values if you are calculating an average from a frequency table.

Resistance and extreme values or shape

- The average, or mean, is the centre of mass of a distribution.
- One way to imagine this is that the distribution exactly balances on a pivot at the mean value.
- The average is the most used (probably overused) statistic. It is useful especially where distributions are reasonably symmetric. For distributions with long tails, it can give misleading signals.
- For example, average household income overstates the income that you or I are likely to see, as it includes households like Buckingham palace.
- Technically, we term this property low resistance to extreme values.

The median

- The median is a measure of the midpoint of a distribution. It can be calculated as follows.
  1. Suppose that there are n observations altogether. Arrange all n the observations in order of size, from smallest to largest.
  2. If n is odd, then the median is the middle observation, i.e. the $\frac{n+1}{2}$th observation.
  3. if n is even, the median lies halfway between the two middle observations, i.e. the $\frac{n}{2}$th and $(\frac{n}{2}+1)$th observations.
- If the ordered observations are ranked from 1 to n, the median is the value of the observation corresponding to the average rank.

- The median, unlike the mean, is quite resistant to extreme values.
- In statistical practice we commonly see standard and resistant measures reported side-by-side. They are not usually much different.
- If they do differ, we might need to think more carefully about what kind of summary we want.
- Often, if we want a summary which suggests a *typical* member of the distribution, we should report the median.
- Many news reports quote averages (like average income is £30000 per annum) which most people wrongly interpret as a median: average people tend to have median incomes, not average incomes.
- Note that the median is technically not uniquely defined:
  - for data 4,7,7,9 median is exactly 7,
  - but for data 4,7,8,9 median is 7 or 8 or any number in the interval $[7, 8]$.

Measuring spread: the standard deviation

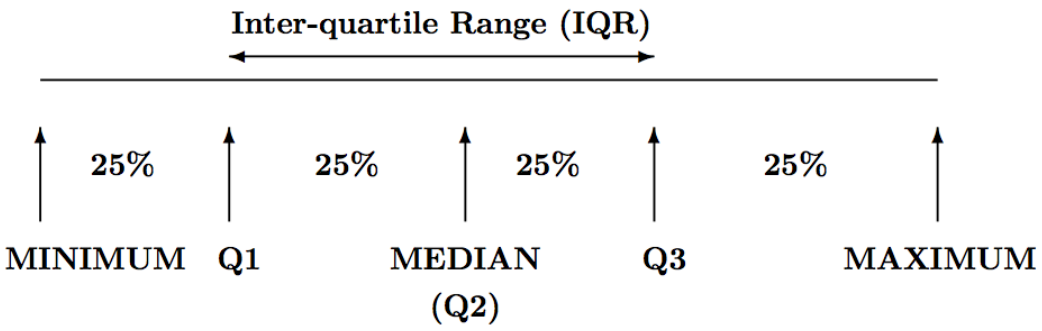
- The most common measure of spread is the *standard deviation*. A possible, BUT INCORRECT, definition of the standard deviation of a sample is

$$\widehat{s} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- Think about what this formula represents: we subtract the average from each value, giving a *deviation* from the average. The deviations are then squared, added, and then divided by n, giving an average squared deviation.
- Consequently, the standard deviation of the sample is a measure of the "average" distance of the values from the centre of the sample. (We must resort to squaring because the average deviation from the mean is zero.)
- For *all* practical work, we will use a slightly different (I.E. CORRECT!) version of the standard deviation of a sample which we define as

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- We will use s and not $\widehat{s}$, and when we say *standard deviation*, we will ALWAYS mean s.
- The only difference between the two lies in the divisor: n versus $n-1$. It turns out that s is a better estimator of the *standard deviation in the original population from which our sample was drawn.*
- $n-1$ are the number of *degrees of freedom* of the variance or standard deviation.



- Make sure that you know which version of the standard deviation is being calculated by your calculator, R, and so forth.
- We will sometimes use the word *variance* to mean s^2.
- You can use the formula

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i^2) - n\bar{x}^2$$

  if you wish - it makes hand calculations easier. However, the first formula is better if you calculate using a computer.
- The standard deviation is very sensitive to extreme values, and so not a resistant measure of spread.

Quartiles and the interquartile range

- The other main summary of spread is given by the interquartile range: IQR, the difference between the first and third quartiles.
- For these we must calculate quartiles as follows. Every distribution has three quartiles.
- The first quartile, Q_1, is a number such that 25% of the values in the distribution do not exceed this number. This is called a lower quartile.
- The second quartile, Q_2, is a number such that 50% of the values in the distribution do not exceed this number. This is the same as the *median*.
- The third quartile, Q_3, is a number such that 75% of the values in the distribution do not exceed this number. This is called an upper quartile.
- The three quartiles divide up a distribution into four quarters, where the number of values in each quarter is about the same.
- Calculate Q_1 as the median of the values falling below the overall median.
- Calculate Q_3 as the median of the values falling above the overall median.

- We define the *interquartile range* IQR to be the difference between the first and third quartiles:

$$IQR = Q_3 - Q_1$$

- The IQR clearly contains the middle 50% of the values in the distribution.
- The IQR is resistant to extreme values in the bottom 25% or top 75% of the values.

Mean and SD plots

- There are a number of ways of visualizing the shape of a distribution - the histogram is probably best.
- However we need nice ways to visualize many distributions in parallel, and nice ways to generate graphs from simple numerical summaries.
- It can be shown that for *any* unimodal probability distribution, 95% of the data values lie within three standard deviations of the mean. This is a very large family of shapes.
- For a smaller number of shapes – generally those that are nearly symmetrical and with not too many outlying values, it can be shown that about 95% of the data values lie within two standard deviations of the mean.
- This suggests the following crude plot.
- Suppose that we take a distribution which has been categorized according to some factor.
- Calculate the mean $\bar{x}$ and s for each grouping.
- For each group, plot the mean $\bar{x}$ and plot a *whisker* extending down to $\bar{x} - 2s$ and a *whisker* extending up to $\bar{x} + 2s$.
- The mean point then shows the centre of the distribution, and the whiskers indicate the range over which falls the bulk of the distribution. The result is a very quick way of comparing many distributions.
- The main difficulties with such a plot are (1) that we know these measures are not resistant; (2) they convey nothing about shape.

Boxplots

- A common five-number summary of a distribution consists of the minimum, Q_1, median, Q_3 and maximum, and these can be displayed graphically as a *boxplot*.
- The rules for drawing boxplots are as follows.

1. Draw a box with lower boundary at $Q1$ and upper boundary $Q3$.
2. Draw a line inside the box at $Q2$ (median).
3. Mark outliers, if any. Outliers are any values larger than $Q3 + (1.5)IQR$, or any values smaller than $Q1 - (1.5)IQR$.

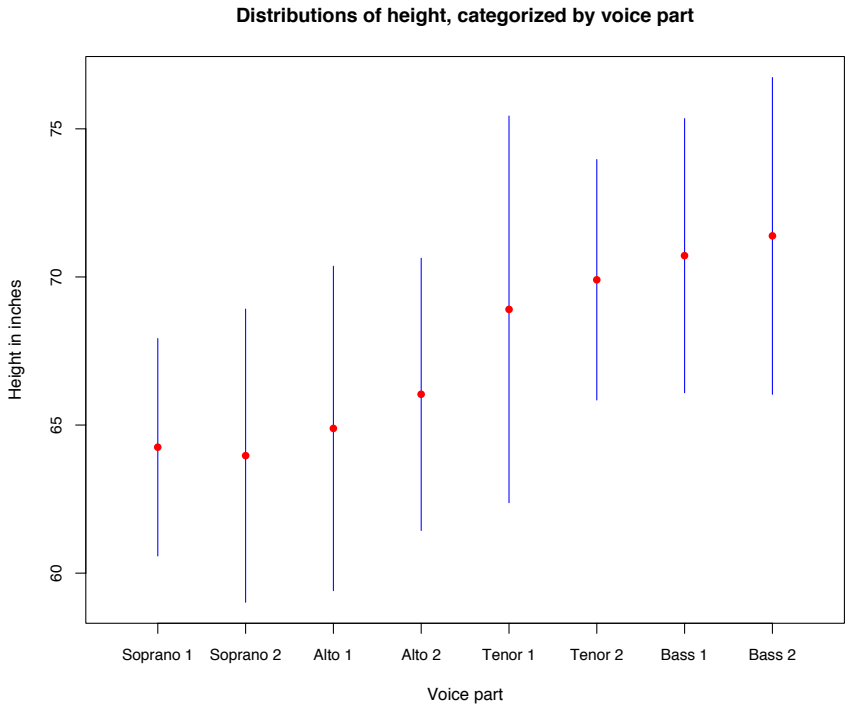


Figure 2: See previous histogram of "Heights of singers by voice type".

4. Excluding the outliers, draw lines from the new minimum value (if any) up to the box, and from the new maximum value (if any) down to the box. These lines are called the whiskers of the box.

- Boxplots are very useful and hence commonly used to compare several distributions at once.

Cuckoo data box plot example: handout only

- Consider the cuckoo data in table 4. For these 100 values, the median is 22.45, the lower quartile is $Q_1 = 22.0$, the upper quartile is $Q_3 = 23.3$. The smallest and largest values are 19.6 and 25.0 respectively.
- We draw a boxplot as shown in the figure. A box is drawn between Q_1 and Q_3, representing the IQR. The position of the median is indicated by a line dividing the box.
- *Whiskers* are then drawn from the box to the smallest and largest values. Exception: we don't want to emphasize outlying values, so it is usual practice to extend the whiskers to largest/smallest values only if they lie within $1.5 \times IQR$ of the upper/lower quartile.
- Here, $Q_1 - 1.5IQR = 22.0 - (1.5)(1.3) = 20.05$ so that we draw the whisker down to the smallest value not lower than 20.05. There are some values at 20.1, so the whisker is extended down to 20.1. Any values smaller than this are plotted as outliers, indicated by bullets, circles, etc.
- For the upper part of the boxplot, all values are within $Q_3 + (1.5)IQR = 25.25$, and the whisker thus extends up to the largest value, 25.0.

Boxplots and Outliers

- We seem to have defined *outlier* en passant. This choice of defining an observation as extreme if it lies more than 1.5IQR distant from a lower/upper quartile is made out of experience and practice.
- It leads on average to about 0.7% of (normally distributed) observations being described as extreme for large samples. For small samples, the proportion is higher - up to about 8%.
- Boxplots are a limited means of displaying information about a single distribution - histograms are much better at displaying the clusters in a distribution, for example.
- However, boxplots can be very useful in comparing many distributions.
- A slight refinement is to make the box widths proportional to the square root of the number of observations in the group. This is recommended when using the R package, but not recommended when drawing by hand.

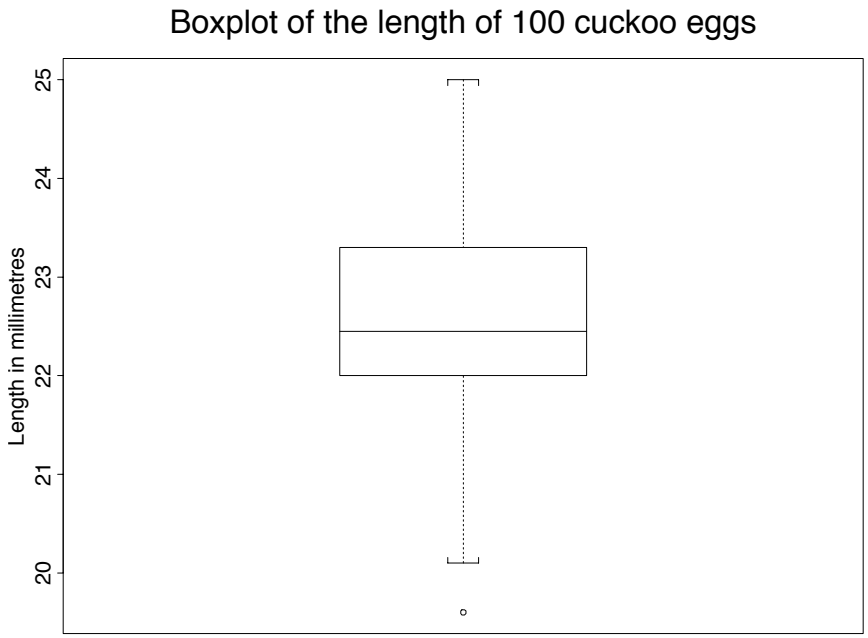


Figure 3: See previous table of "Lengths in millimetres of 100 cuckoo eggs". Boxplot calculation in handout notes.

Interpreting boxplots

- Examine a single boxplot for the following features.
  - Is the median line in the centre of the 50% portion of the distribution? If not, some degree of skew or assymmetry is indicated.
  - Are the whiskers the same length? If not, some degree of skew or assymmetry is indicated.
  - Are there any outliers?
- Examine several boxplots for the following features.
  - Do the groups have similar shapes? Some might seem symmetric, some skewed.
  - Are the median lines about the same, or do there seem to be differences in location between groups?
  - Are the IQRs about the same, or do there seem to be differences in spread between groups?
  - Are there some groups with more outliers than others?
- Such inspection can reveal important differences between groups.
- Note that with computer packages like R you can add further detail, such as colour, to aid in interpretation of the underlying statistical story.

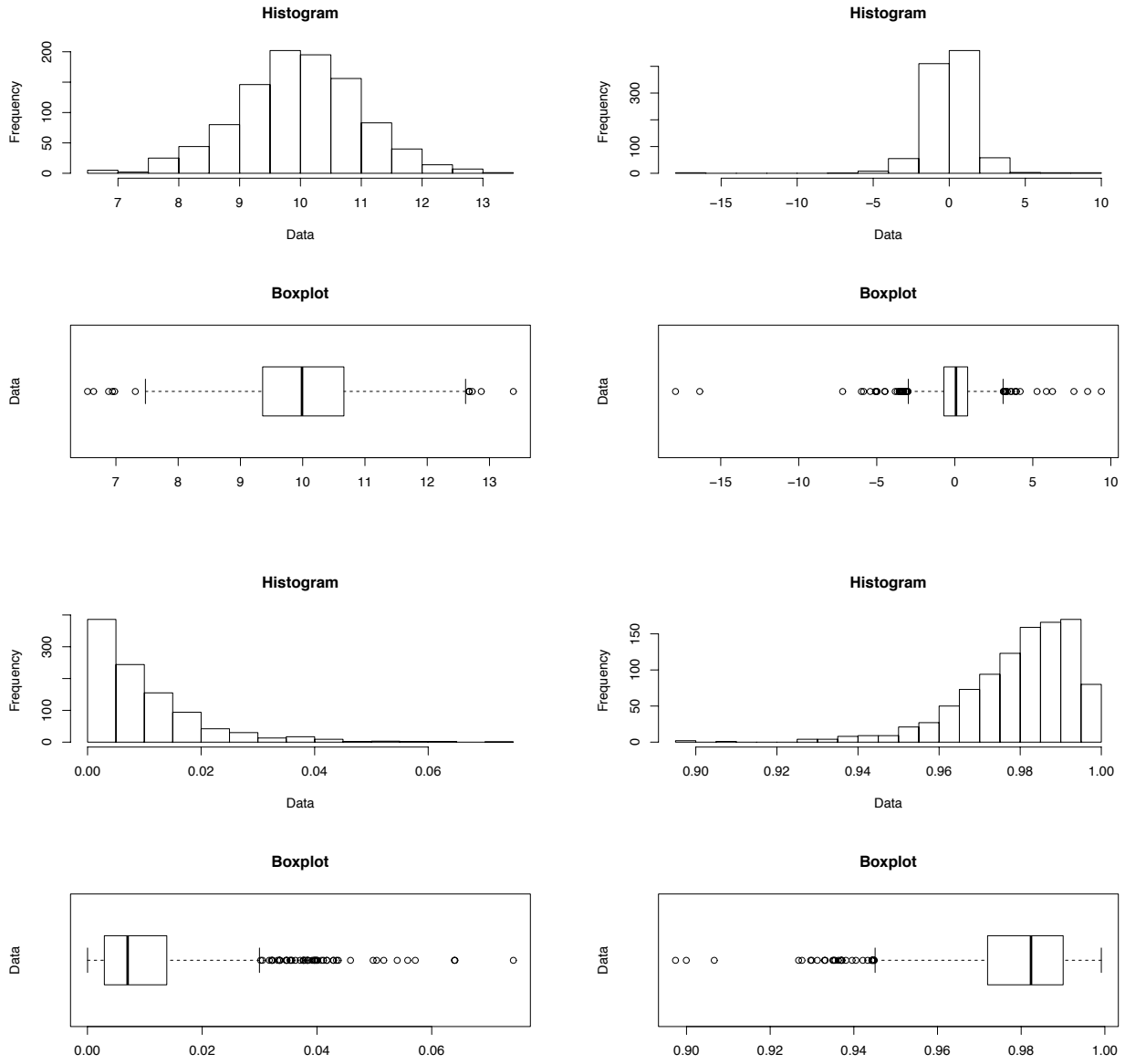
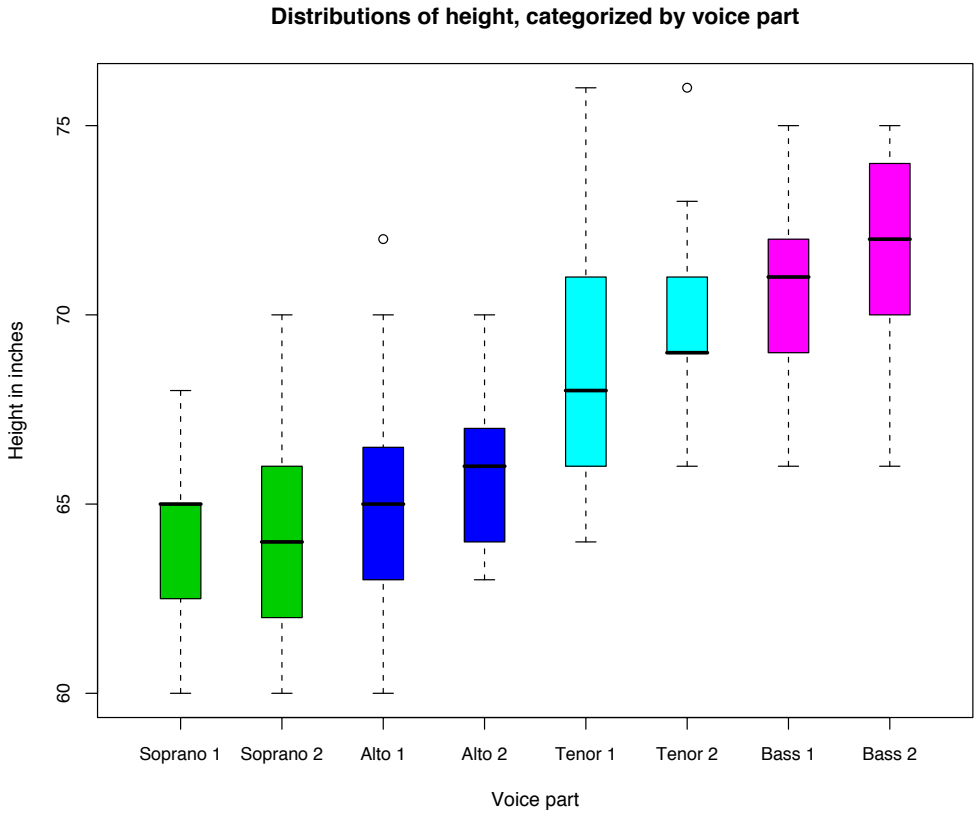


Figure 4: Top left: histogram and boxplot of a symmetric distribution (actually N(10,1)). Top right: histogram and boxplot of a symmetric distribution with long tails each side (t_3). Bottom left: histogram and boxplot of a distribution showing skew to the right (Gamma). Bottom right: histogram and boxplot of a distribution showing skew to the left (negative Gamma).


Distributions of height, categorized by voice part
Height in inches
60
65
70
75
Soprano 1
Soprano 2
Alto 1
Alto 2
Tenor 1
Tenor 2
Bass 1
Bass 2
Voice part

Linear transformations of data

- Take a set of n data values $x_1, x_2, \ldots, x_n$.
- A linear transformation of the data is one which defines a new set of data values $z_1, z_2, \ldots, z_n$, where

$$z_i = a + bx_i$$

  for some values a, b.
- How are summary statistics for the original and transformed values related?
- For the original data x_i we have summaries $\bar{x}, s_x, m_x, Q1_x, Q3_x$ and IQR_x.
- For the new linearly transformed data z_i we have summaries $\bar{z}, s_z, m_z, Q1_z, Q3_z$ and IQR_z.
- They are related via:
  - Mean: $\bar{z} = a + b\bar{x}$.
  - standard deviation: $s_z = bs_x$.
  - Median: $m_z = a + bm_x$. As linear transformations preserve the ordering of the data values.
  - Quartiles: $Q1_z = a + bQ1_x$ and $Q3_z = a + bQ3_x$.
  - IQR: $IQR_z = b \times IQR_x$.
- So, the location summaries change in the same way as the data.
- However, the spread summaries only reflect the change in scale (b) and not the change in location (a).
- Principle: making linear transformations to data should not affect the conclusions we draw from data.

Linear transformations of data: proof of sd

- Proof of standard deviation: $s_z = b\, s_x$. Quartile proofs are trivial.

$$\begin{aligned} s_z^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}((a + bx_i) - (a + b\bar{x}))^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(bx_i - b\bar{x})^2 \\ &= \frac{b^2}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= b^2 s_x^2 \\ \Rightarrow \quad s_z &= b\, s_x \end{aligned}$$

Standardisation

- We *standardise* a list of data $x_1, \ldots, x_n$, by re-expressing each value in terms of *standard units* as follows:

  1. calculate the mean $\bar{x}$ for the distribution;
  2. calculate the standard deviation s for the distribution;
  3. For each value in the distribution, subtract the average and then divide by the standard deviation.

- Hence the formula we use is that the standardised value

$$v_i \quad = \quad \frac{x_i - \bar{x}}{s}$$

- Each transformed value v_i tells us how far the value is from the average in terms of standard deviations.
- The sign of each v_i tells us whether the value is larger or smaller than the average.
- This is a linear transformation $a + bx_i$ with $a = \frac{-\bar{x}}{s}$ and $b = \frac{1}{s}$.
- The transformed data has the following properties:
  - the mean is zero: $\bar{v} = 0$,
  - the standard deviation is one: $s_v = 1$.
- The main purpose is to transform all data to a common scale for purposes of comparison.

The Normal distribution approximation for data

- Many of the histograms that we have examined so far have held certain similarities.
- In particular, many of them have a lot of frequency in the centre of the distribution, and not much in the tails.
- That is, many of them look bell-shaped, or would if we drew a curve through the histogram.
- See for example the histogram of the heights of male singers.
- These heights have been standardized by subtracting the mean and dividing by the standard deviation.
- The added line is the function

$$f(x) = \frac{1}{\sqrt{2\pi}} exp(-x^2/2)$$

  which we can refer to as the *standard Normal curve.*

- Powerful mathematical techniques (for example, the central limit theorem) can be used to show that very many distributions of measurements have, or approximately have, this sort of underlying curve.

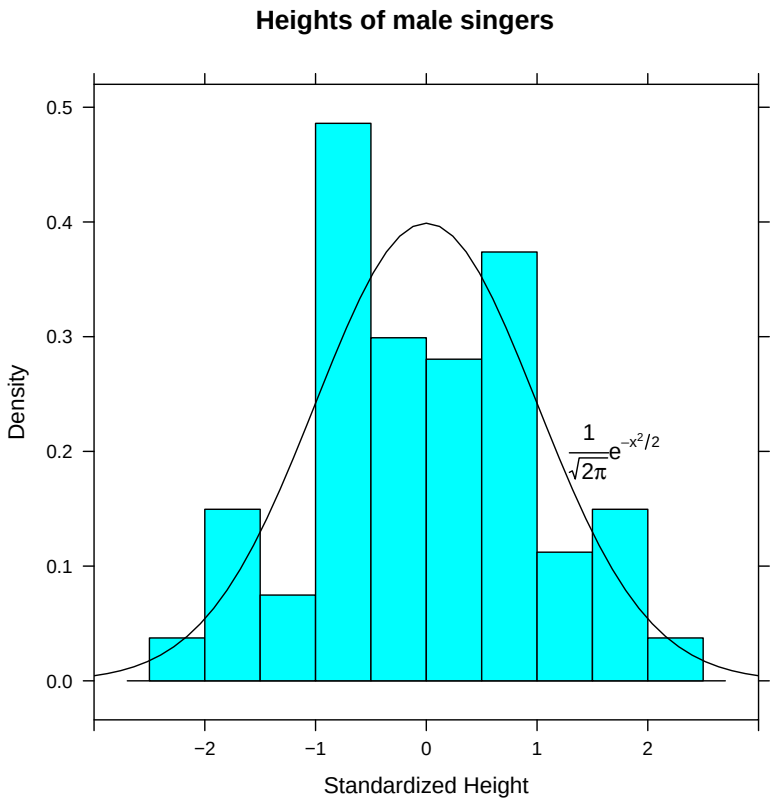


Figure 5: The standardised heights of male singers with normal approximation.

- The Belgian mathematician Adolphe Quetelet (1796-1874) introduced the idea of using the Normal distribution as an "ideal histogram" representing what would happen if sample sizes became very large.
- The Normal distribution is thus a *mathematical mode* for many distributions.
- Mathematical statistical theory shows that if the attributes for an individual can be thought of as being formed from many independent influences, the distribution of the measurement should be approximately Normal.
- For example, adult height might be thought of as being influenced by parents' genes, diet, health and diet of parents, various environmental factors, and so forth.
- The function of the Normal curve with mean μ and standard deviation σ is

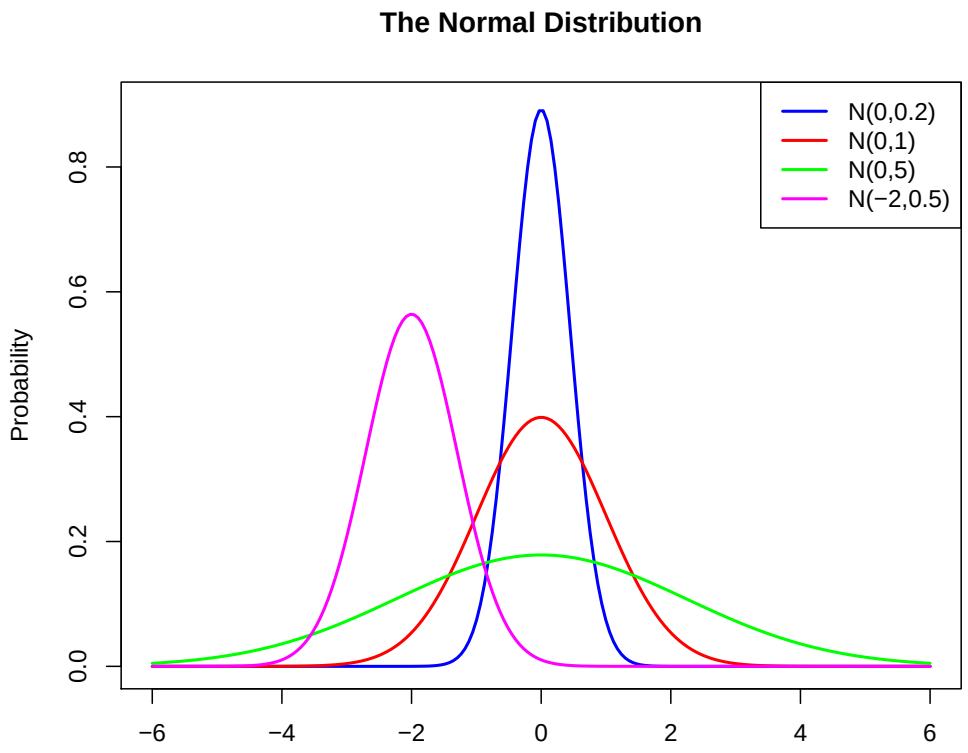
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2\sigma^2}(x-\mu)^2), \quad -\infty < x < \infty$$

- The shape of the curve depends on its mean μ and standard deviation σ
- The curve is symmetric about its mean μ, whilst the standard deviation σ flattens or sharpens the curve.

- Our notation for stating that a variable x has a Normal distribution with mean μ and standard deviation σ is

$$x \sim N(\mu, \sigma^2)$$

- This probability distribution is also known as the Gaussian distribution, named after Carl Friedrich Gauss (1777−1855), who used it to analyze astronomical data.



- We will be interested in finding the area under this curve between two values a and b, as this relates to the probability of finding the variable x there, which we write as $P(a < x < b)$. (More rigour next term.)

- Unfortunately, the Normal function has no definite integral, i.e. there is no way to simplify

$$P(a < x < b) \quad = \quad \int_a^b f(x)dx \quad = \quad \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

- Although it is possible to show that the total area under the curve is 1 for every μ, σ, i.e.

$$P(-\infty < x < \infty) \quad = \quad \int_{-\infty}^{\infty} f(x)dx \quad = \quad 1$$

- Instead, we must resort to tables of the function: see the Normal Table handout.

The Standard Normal distribution

- Fortunately, we need just one table, as all Normal distributions can be transformed to the *standard normal* by *standardising*: subtracting the mean, μ and then dividing by the standard deviation σ.
- If $x \sim N(\mu, \sigma^2)$ then we define
$$z \;=\; \frac{x-\mu}{\sigma}$$
- Then it is easy to show that $z \sim N(0,1)$ i.e. that z has a Normal distribution with mean zero and standard deviation 1.
- This distribution is called the standard Normal distribution, with
$$f(z) = \frac{1}{\sqrt{2\pi}} exp(-z^2/2), \;\; -\infty < z < \infty$$
- In reverse, if $z \sim N(0,1)$ and if x is defined as $x = \sigma z + \mu$, then this implies that $x \sim N(\mu, \sigma^2)$.

Areas under the standard Normal curve

- So if $z \sim N(0,1)$, we say that the probability of finding the variable z between values a and b is given by the area under the standard Normal curve between a and b:
$$P(a < z < b) \;=\; \int_a^b f(z)dz \;=\; \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$
- The area beneath the whole standard Normal curve is 1, or 100%.
- To find the area to the left or to the right of a certain value, we need to use *Normal Tables* and logic.
- For any positive value z, the standard Normal table given out gives the area to the left of z under the Normal curve.
- The usual notation for the area to the left of z under a standard Normal curve is $\Phi(z)$.
- The definition of $\Phi(z)$ is (more detail next term):
$$\Phi(z) \;=\; \int_{-\infty}^{z} f(z')dz'$$
- For example, $\Phi(1.0) = 0.8413$ and $\Phi(1.96) = 0.975$.
- We use logic, especially the symmetry about 0, to find areas between two values, or when z is negative, or when you need to find an area to the right. Be encouraged to draw pictures when doing so.
- Notice that the area between -1.96 and $+1.96$ is $2\Phi(1.96) - 1 = 0.95$.
- That is, 95% of the values lie within roughly 2 standard deviations of the mean, zero. About 68.26% of the values lie within 1 standard deviation of the mean, and 99.74% of the values lie within 3 standard deviations.

Areas under other Normal curves

- Suppose that a distribution is Normal with mean μ and with standard deviation σ.
- Then the area under this Normal curve to the left of x is simply

$$\Phi\left(\frac{x-\mu}{\sigma}\right)$$

- For example, the heights of male singers shown in the histogram have a mean μ of about 70.36 inches, and a standard deviation s of 2.72 inches.
- Let us suppose that a Normal distribution gives a reasonable fit for this data.
- What proportion of these men are less than 74 inches tall?
- The answer is $\Phi(\frac{x-\mu}{\sigma}) = \Phi(\frac{74-70.36}{2.72}) = \Phi(1.34) = 0.9099$, or about 91%.
- The argument can be reversed as follows.
- Find out the height x such that 75% of male singers are no taller than x.
- From the Normal table, we discover that $\Phi(0.675) = 0.75$ approximately.
- Therefore we must have $\frac{x-70.36}{2.72} = 0.675$, i.e. $x = 72.20$ inches.
- Notice that if 2.72 is the standard deviation, then we can say instantly that about 95% of male singers have heights within about 2 standard deviations = 5.44 inches of the mean.

Assessing Normality - the Normal quantile plot

- We can only use the Normal approximation when it is reasonable, but how can we tell whether the approximation is reasonable?
- By using *Normal quantile plots*.
- The basic idea is that these plots have the property that plotted points for Normally distributed data should fall roughly on a straight line.
- Systematic deviations from the straight line indicate non-normality.
- We construct a Normal quantile plot as follows.
- First, order the n values of the data. Sometimes the notation for the ordered values is $x_{(1)}, \ldots, x_{(n)}$.
- Next, calculate the values $\frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}$. Clearly, this gives $n+1$ equal divisions of the interval $(0, 1)$.
- Use Normal tables to find approximately the value z_k such that $\Phi(z_k) = \frac{k}{n+1}$ for the values $k = 1, 2, \ldots, n$.
- In words, we find the number z_k such that the area under the standard Normal curve to the left of z_k is $\frac{k}{n+1}$.

- This gives us n values $z_1, z_2, \ldots, z_n$. Here z_k is called the $\frac{k}{n+1}$ *quantile* of the Normal distribution.
- The quantile plot consists of the n pairs of points $(x_{(1)}, z_1), (x_{(2)}, z_2), \ldots, (x_{(n)}, z_n)$.
- For example, when $n = 18$ and $k = 11$ we calculate $\frac{11}{19} = 0.5789$.
- By using standard Normal tables, we find that the area to the left of 0.200 is 0.5793, and the area to the left of 0.190 is 0.5753.
- Hence the $\frac{11}{19}$ quantile of the Normal distribution is about $z_{11} = 0.199$.
- For our drawing purposes, high accuracy isn't required and we'd settle for $z_{11} = 0.20$ rather than 0.199.
- The Normal table supplied only gives areas for $z \geq 0$ with areas hence exceeding 0.5, so we must use symmetry for smaller areas (for which z will be negative).
- To find the $\frac{8}{19} = 0.4211$ quantile, we want the value z such that an area to the left of $-z$ is 0.4211.
- Because of symmetry about zero, this must be the same as requiring that the area to the *right* of $+z$ be 0.4211, i.e. that the area to the left of $+z$ be $(1 - 0.4211) = 0.5789$.
- But *we already know the answer to this* from the previous calculation which gave $+z = 0.20$, and it now follows that the $\frac{8}{19}$ quantile is about $z_8 = -0.20$.
- In fact, when we draw quantile plots, we need calculate only for the positive side of the distribution, as we can then deduce the negative side values.
- For example, notice in the Lean Body Mass example table (next) the symmetric values -1.28 and $+1.28$, -0.84 and $+0.84$, and so forth.
- As we said before, the quantile plot consists of the n pairs of points $(x_{(1)}, z_1), (x_{(2)}, z_2), \ldots, (x_{(n)}, z_n)$.
- It is usual to put the z's on the y-axis (R does the opposite and puts them on the x-axis, but this can be altered).

- For a second example, consider the hamster lifespan data set, which gives the lifespan in days of 11 hamsters.

k	1	2	3	4	5	6	7	8	9	10	11
Hamster Lifespan $x_{(k)}$	116	314	364	496	545	562	579	612	711	744	760
$\frac{k}{11+1}$	0.083	0.167	0.250	0.333	0.417	0.5	0.583	0.667	0.75	0.833	0.917
z_k s.t. $\Phi(z_k) = \frac{k}{11+1}$	-1.39	-0.96	-0.67	-0.43	-0.21	0	0.21	0.43	0.67	0.96	1.39

- Note that there are alternative ways of calculating the quantiles.
- The method given here is perhaps simplest - it just chops up the distribution into $n+1$ equal portions.
- Computer packages e.g. R offer this and other methods.

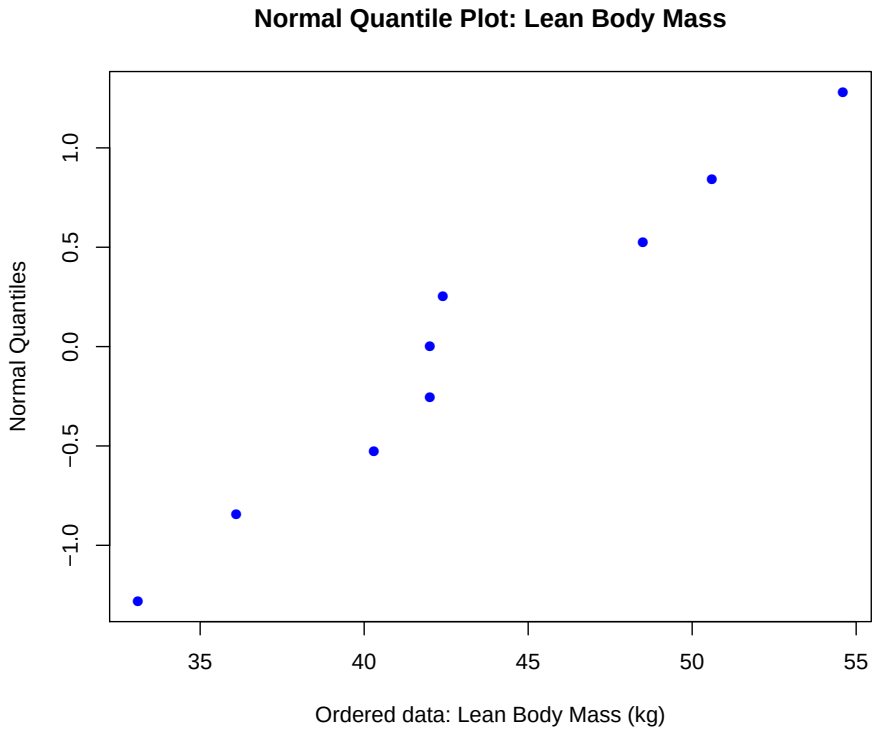


Figure 6: Quantile plot with original unstandardised data. This is fine when doing quantile plots by hand, or in an exam! Note the cluster of 3 points.

Interpreting Normal quantile plots

- The points for Normally distributed data should fall approximately on a straight line: use the fat pen test.
- Systematic deviations from the straight line indicate non-normality.
- Outliers appear as points that are far from the overall pattern of the plot.
- Vertical piles and horizontal jumps in the plot show *clustering (granularity)* by indicating gaps in the histogram.
- Example: the Lean Body Mass quantile plot shows up a large gap between the cluster at about 42.0, and 48.5.
- If the largest observations are to the right of a straight line drawn through the *main body of the data*, the distribution is skewed to the right (long tail of large values).
- Some computer packages misleadingly draw a line between the lefthandmost and righthandmost points, rather than through the main body of the data.

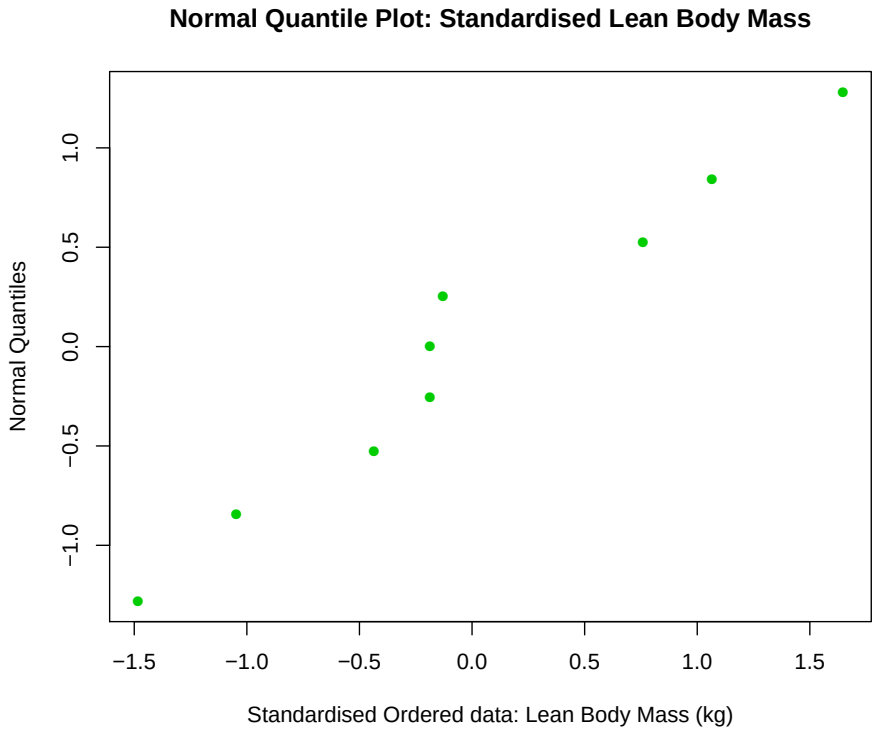


Figure 7: Quantile plot using standardised data. Now both axis represent standardised quantities: the y-axis has the theoretical Normal model, and the x-axis has the standardised values.

- If the smallest observations are to the left of a straight line drawn through the main body of the data, the distribution is skewed to the left (long tail of small values).
- There is some evidence of skew to the left in the Hamster Lifetime quantile plot.
- Bear in mind that real data usually show at least some departure from theoretical Normality.
- It is often possible to transform the data so that the transformed data are approximately more Normal.

Univariate Transformations

- Sometimes a Normal quantile plot will show that the data is clearly not Normally distributed.
- In this case we can often fix this problem by transforming the data using a carefully chosen transformation.
- We will consider these transformations in more detail later, when we deal with bivariate data.
- The kinds of transformation of x we might employ are simple one-to-one *power* transformations, such as $x^4, x^3, x^2, \sqrt{x}, x^{1/3}, \ln x, 1/\sqrt{x}, 1/x$, etc.

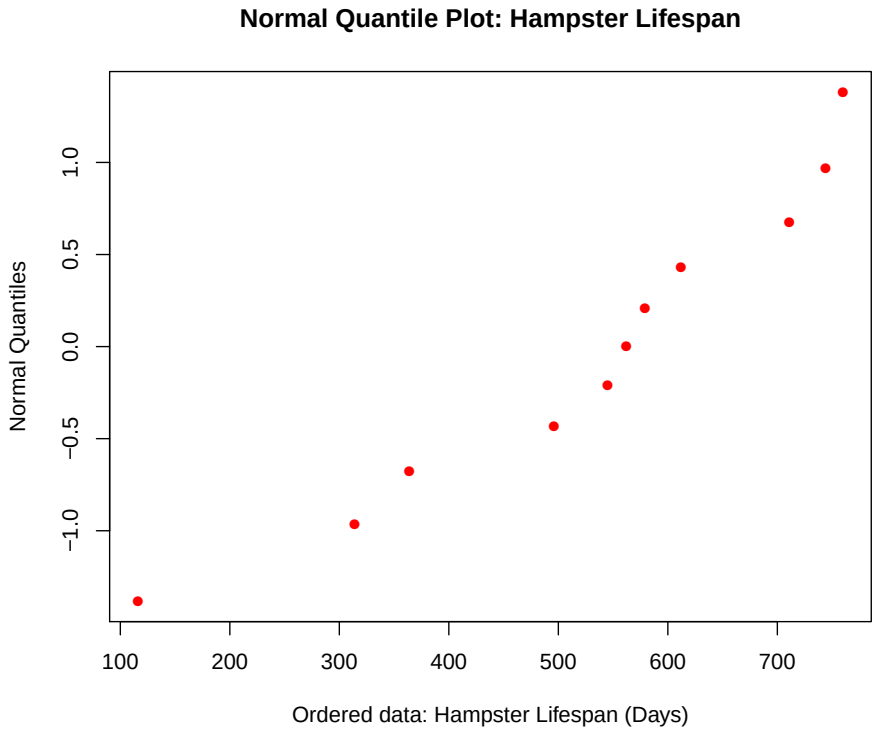


Figure 8: Quantile plot using original Hamster data. Note the clustering and the possible skew to the left (long tail to left).

- We can decide informally whether a transformation is useful by examining a Normal quantile plot of the transformed data.
- How do we decide which transformations to try?
- Imagine a *ladder of powers*, with $\log x$ taking the place of power zero.
- Start with power 1, which corresponds to the original data x.
- If the data are right-skewed (long tail to the right) move down the ladder of powers, so try square root: $\sqrt{x}$, cube root: $x^{1/3}$, logarithmic: $\ln x$, inverse square root: $1/\sqrt{x}$, inverse: $1/x$ etc.
- If the data are left-skewed (long tail to the left) move up the ladder of powers so try square: x^2, cube: x^3, quartic: x^4 etc.

Univariate Transformations: Warp Breaks example

- Consider the following data giving the number of warp breaks per loom, where a loom corresponds to a fixed length of yarn.

10	12	13	14	15	15	15	16	16
17	17	18	18	18	19	19	20	20
21	21	21	21	24	24	25	26	26
26	26	27	28	28	28	29	29	29
29	30	30	31	35	36	36	39	39
41	42	43	44	51	52	54	67	70

- These have a long tail to the right, suggesting we need to move down the ladder, so let us try powers of 1/2, 0 (log), and −1/2.
- Histograms for the original data and three possible transformations are shown in the next slides along with quantile plots of the untransformed and log transformed data.
- Note how the quantile plot for the original data reveals skew to the right, together with some granularity.
- The quantile plot for the log of the data displays no obvious problem with skew, but the granularity is still – of course – present.

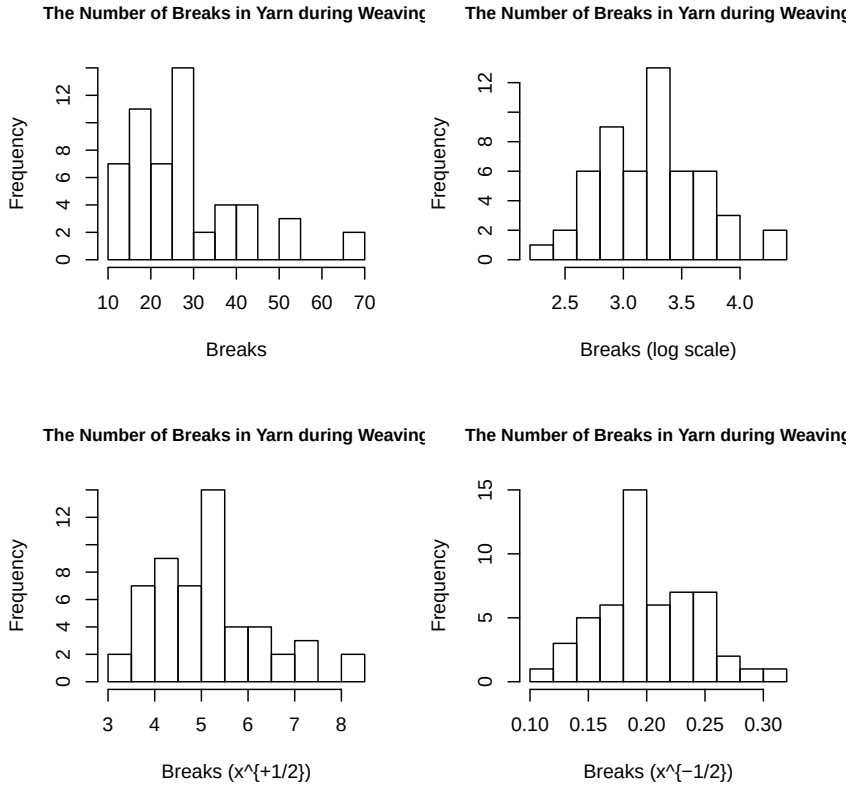


Figure 9: Number of warp breaks per loom: original data and three possible transformations, $\log x$, $x^{1/2}$ and $x^{-1/2}$.

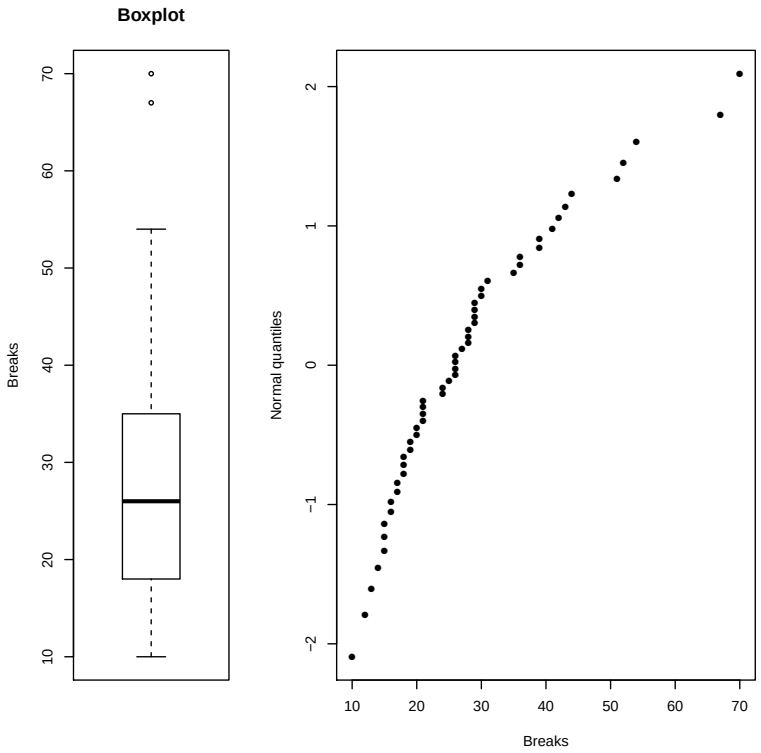


Figure 10: Boxplot and quantile plot for untransformed Warp Breaks data. Note the clear right skew and clustering/granularity.

More complex Transformations

- Log transformations are not helpful when there are non-positive values.
- An alternative is to transform to $\log(x+1)$. This will depend on how close the data are to one.
- A better alternative is to transform to $\log(\frac{x}{\bar{x}}+k)$ where $k << 1$ is a small constant, e.g. $k = 0.01$.
- In this transformation, the mean of x will be transformed to something close to zero, but positive and k functions as a shape factor: small k will make the transformed data more left-skewed, larger k will make it less so.
- Also changing the units of measure will not change the shape of the distribution.

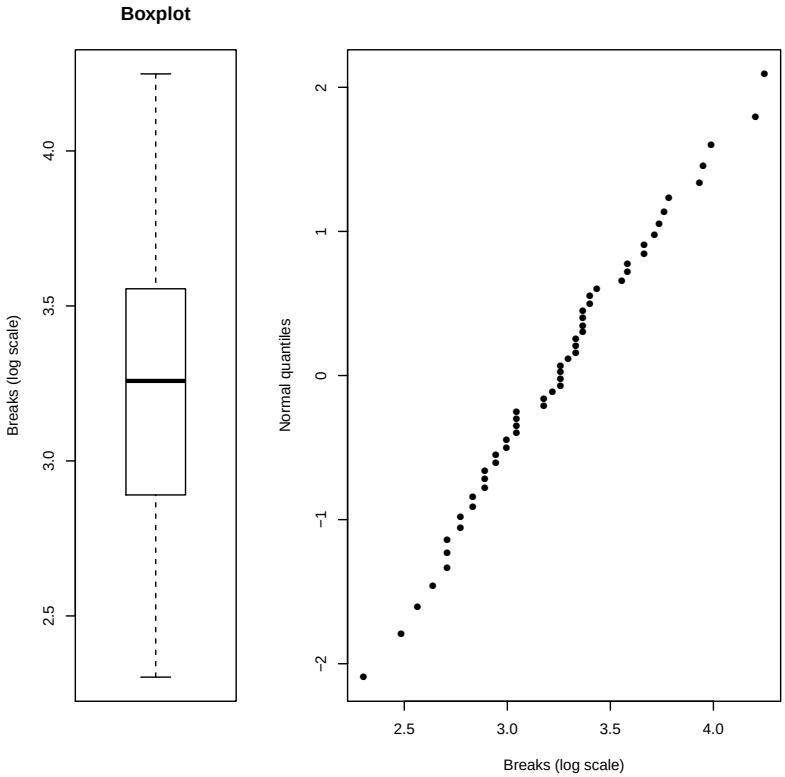


Figure 11: Boxplot and quantile plot for log-transformed Warp Breaks data. Looks a much better fit to a normal distribution. Clustering/granularity still present.

2.5 Measurement and Error

- We have previously argued that the more measurements we make, the more accurate our results become. Why?
- One view of the nature of the world is that when we measure something, the length of a pencil, say, there is some true exact figure that we are trying to measure; but that for various reasons we can't pin down precisely what this exact value is, and the measurements that we make include (hopefully small) *chance errors* (often called *random errors* and *random fluctuation*). That is,

$$\text{individual measurement} = \text{unknowable exact value} + \text{unknowable chance error}$$

- These chance errors arise (a) because to reach a precise figure, we need an infinitely accurate measure; (b) because the circumstances when we measure may change. For example, measuring the length of a pencil could give different results according to the fineness of measure used; and we might get different results from different people at different times of the day. Further, the length of the pencil itself might be fluctuating minutely in response to changes in temperature, etc. If you think hard about it, it is clearly impossible even to replicate something as simple as tossing a coin: do you throw it to the same height? Are your fingers sweaty to the same extent? Whatever else, the *time* at least will be different.

- Essentially this means that we can never measure precisely enough; and we never know exactly what we are measuring: there will always be random fluctuation about some "true" value, partly because the unknowable true value may itself be changing minutely.
- The purpose in taking many measurements, called *replications*, lies in hoping that the chance errors will cancel each other out, so that the *average chance error* is about zero. This means that the average of a number of measurements should be closer to the true exact value.
- At the same time, we can examine the differences between the measurements and their average, and roughly *estimate* the individual chance errors. Taken together, this is what the standard deviation of the measurement achieves: a rough estimate of the probable size of chance error in a single measurement.
- Sometimes, knowing the typical chance error can be useful when *calibrating* similar processes. (For example, scientific instruments should be calibrated.)
- We detect "extreme" values by expressing all values in terms of standard units. Then, values which are more than about 2 standard deviations from the mean are suspiciously extreme, and values more than about 3 standard deviations distant are distinctly unusual. We call these values *outliers*. At the very least we make a point of noting them as potentially a nuisance. Often we make our calculations with and without the outliers, so as to gauge the sensitivity of our results to the outliers.
- Recall that the average and standard deviation especially are poorly *resistant* to outliers, and the Normal curve provides a poor fit in such circumstances. (The median and inter-quartile range are more resistant.)
- *Bias* or *systematic error* is an error which affects measurements in the same direction: all up or all down, unlike chance error. An example is age, often measured downwards to completed years, and Electricity accounts, measured downwards to the nearest complete unit. This sort of bias, whose exact value is, like the other terms, also unknowable, is more common than you might think.
- The two major aspects of making measurements are thus the degree of bias in them, and the size of the typical chance error. Measurements with large chance errors are said to be highly variable, and the value of s calculated from such measurements will be large.
- Finally, you will all have seen opinion polls reported along with the caveat *accurate to plus or minus 3%*. What are the issues here?

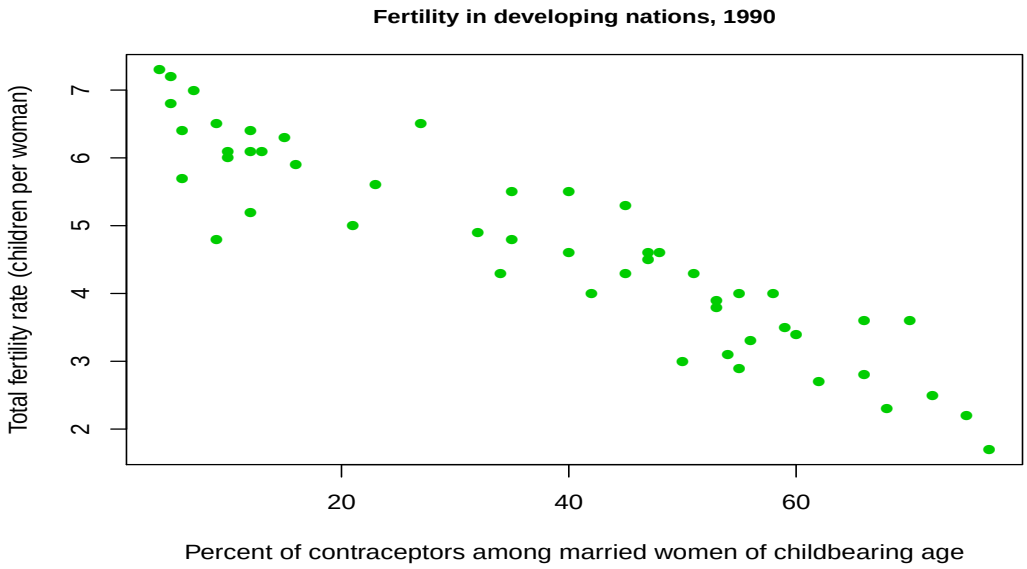
    1. Chance errors affect the measurements for various reasons. For example, the way of making measurements might differ (different questioners, for example); and underlying support (the "true" value being measured) fluctuates from minute to minute.
    2. An estimate of the accuracy of the procedure is about 3%. Actually, the 3% quoted represents about 2 standard deviations according to a fairly reasonable statistical model, so the implication is that the opinion poll could be wrong by as much as 3%, but *probably* not by more. Note *probably* and not *certainly*.
    3. It is less frequently reported that this figure of 3% depends upon the *size* of the sample of voters, and that the sample was representative of the population as a whole.

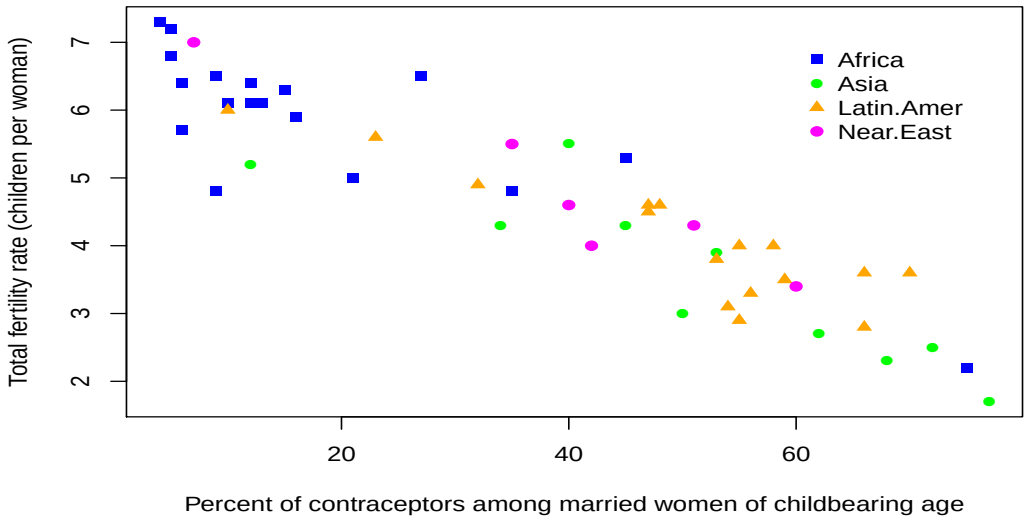
3 Bivariate Data

- We come now to exploring and summarising *bivariate* (two-variable) data.
- We do this by calculating summary statistics for each quantity, as before, and also:
  - graphically, using a scatter plot;
  - numerically, by calculating the correlation coefficient;
  - inferentially, by calculating regression lines.
- We shall be dealing largely with quantitative data.
- Bivariate categorical data is typically harder to deal with - some methods are shown at the end of the Statistics course.
- We have already encountered most of the terminology to be used.
- We suppose that we will obtain n observations on two or more variables.
- Mostly we will be dealing with a random sample of observations.
- We label each pair of observations by a case number (or observation number or individual number).
- The case numbers have no value except to help us identify a particular pair of observations.

3.1 Identifying patterns

- We display bivariate quantitative data on a scatter plot, plotting each (x_i, y_i) as one point.
- The purpose of drawing a scatterplot is to try to identify patterns; in particular
  - its **form**,
  - its **direction**,
  - and its **strength**.
- Numerical summaries are usually inadequate to this task as they don't show all the data. Instead, we need to interpret scatter plots.
- Note that axes for the scatter plot should be chosen appropriately, and axes should be labelled appropriately. Generally speaking, one axis should be not much larger than the other, and the ranges of the axes should be appropriate to the data so that the data are not compressed into one region.
- In interpreting scatter plots, we might bear in mind some of the following.
  - Is there any pattern?
  - Is there a linear pattern, or one which might be transformed to linear?
  - Is a pattern likely to be explainable with reference to another variable?



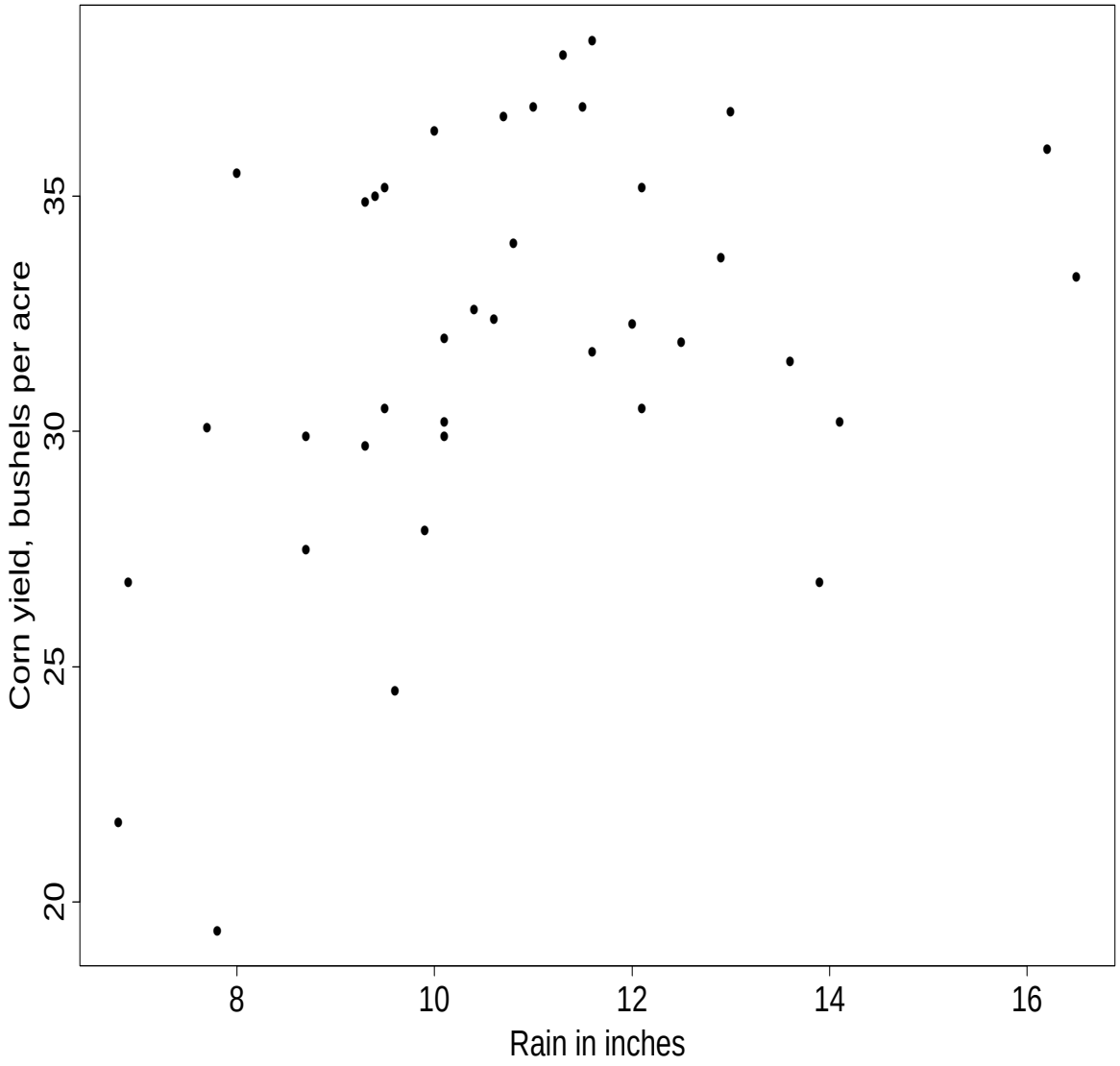


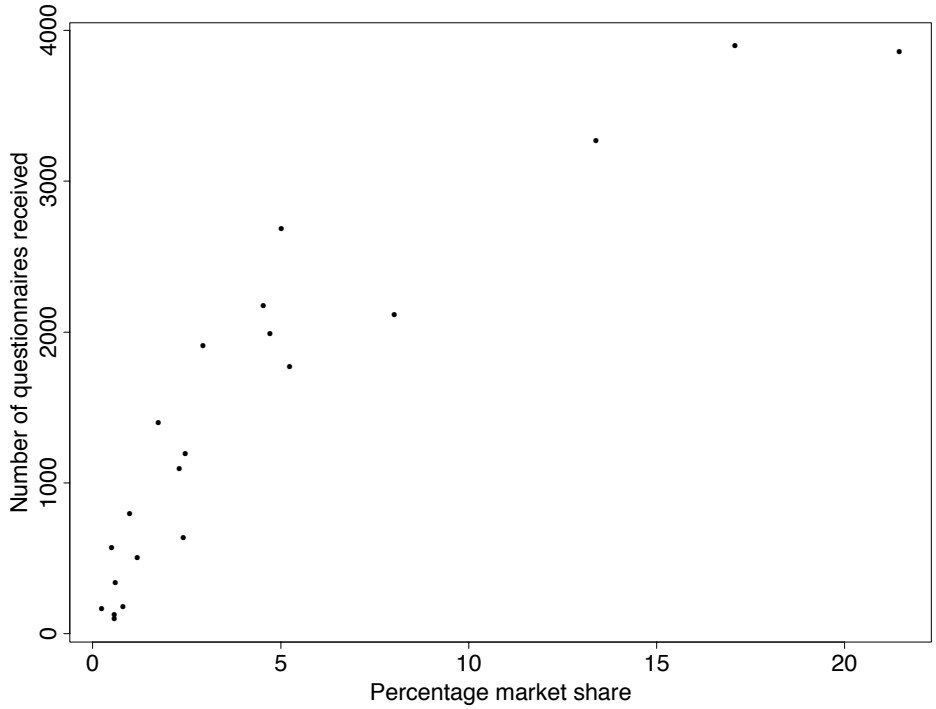
- Consider, for example, a plot of fertility (number of children per mother) versus percentage of contraceptors among women of childbearing age. It seems fairly clear from this plot that we can draw a rough straight line through most of the points on the plot, indicating perhaps a linear association between fertility and contraceptor percentage. There is still a lot of spread about the rough straight line, i.e. the fit is far from perfect. Note that these appear to be averages from countries. On the second plot, geographical region is indicated. This doesn't seem to affect the overall conclusion, in that the pattern seems to be the same, and linear, in each region. Pragmatically, the data seems to agree with the common sense view that fertility declines on average as the percentage of contraceptors in a country rises.

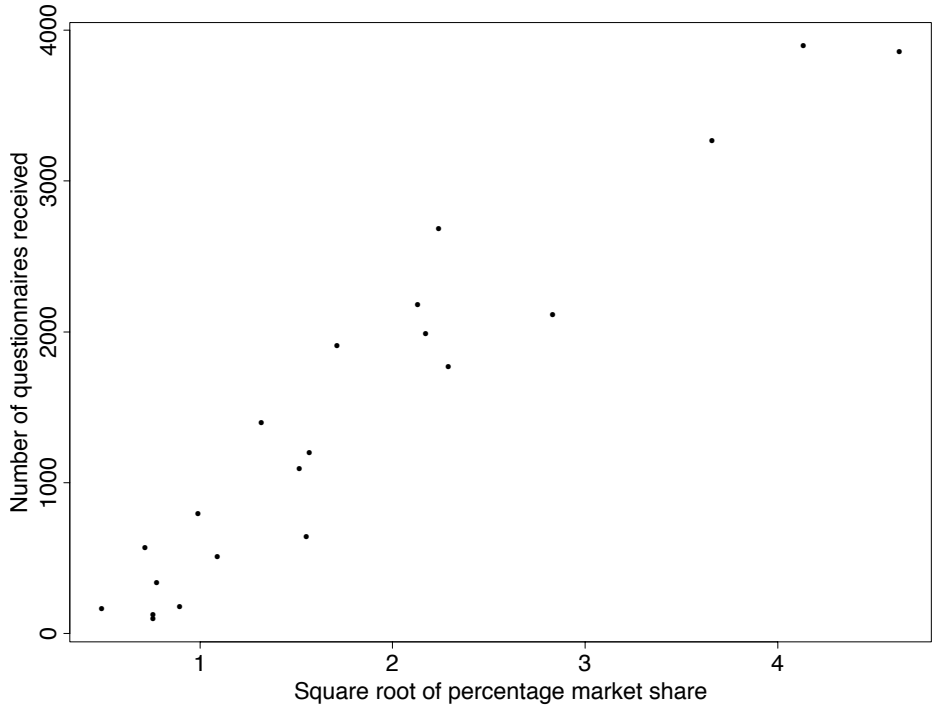
- Note a danger in interpreting these plots. Because averages per country are being reported, a lot of variation at the individual level is being concealed. (Think what would happen if we saw averages per

postal district.) We will deal with this later under the heading of ecological correlations, which are correlations between groups of individuals.

- Another plot is of rainfall versus corn yield in six USA states. The pattern here, if any, is much harder to see. Perhaps we could draw a curve, peaking with rainfall about 11 inches, through the data. This suggests that there might be a quadratic relationship between rainfall and yield. Think about why this might be so - perhaps 8 inches isn't enough rain, perhaps 15 inches is too much.



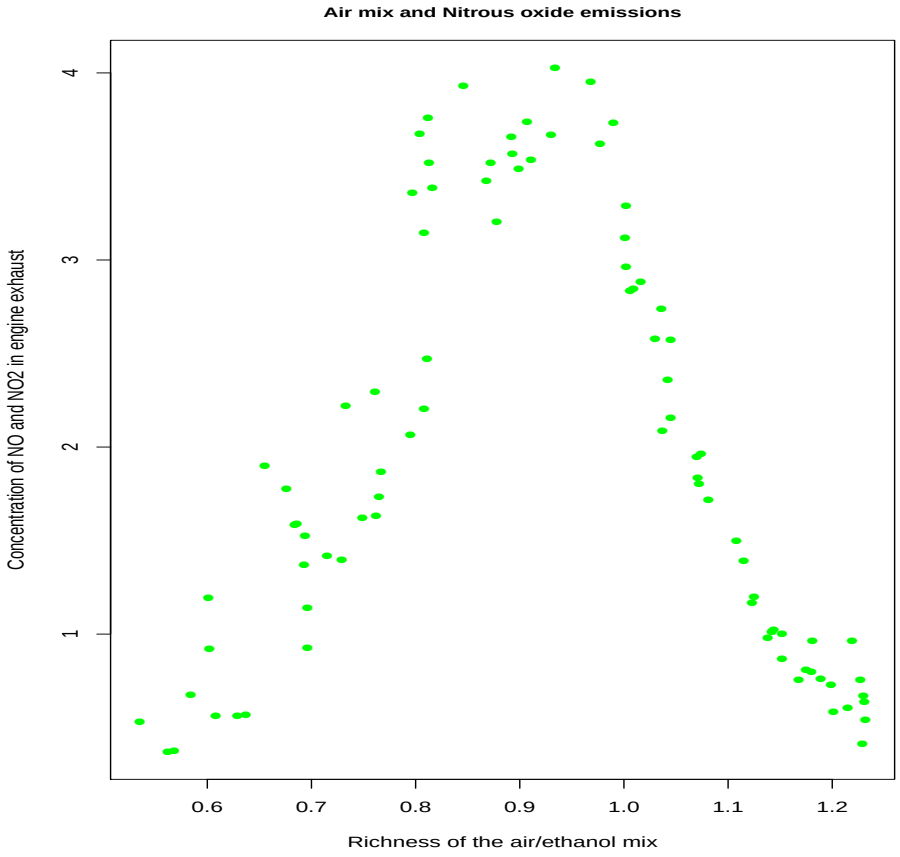




- The next plot shows again the data concerning car manufacturer market share and number of questionnaires per manufacturer returned to Which? magazine in a survey. The plot suggests a curve running through the data: in fact the curve looks rather like what we would get if we plotted $\sqrt{x}$ on

the y axis. This suggests that an appropriate transformation of the data is to replace each value of market share by its square root. Doing so, we see in Table 19 that it now seems possible to draw a straight line through the points, and we could conclude that there seems to be a linear relationship between questionnaire response and square root of market share.

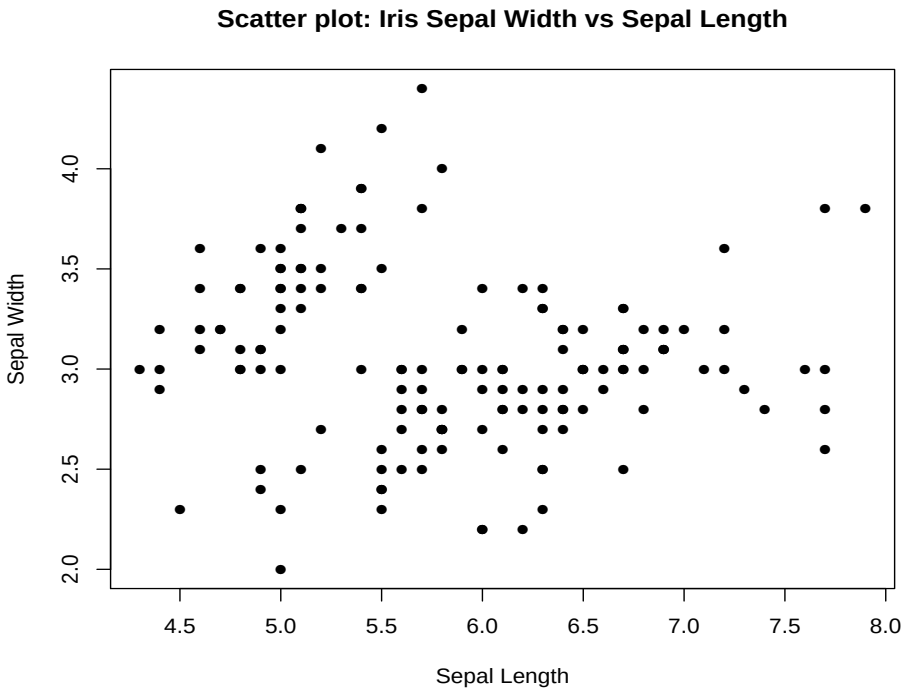
- Now consider a plot of the concentrations of nitrous oxides in engine exhaust as the richness of the air/ethanol mix is varied. There is clearly a strong association between the two, but not linear. It seems that the emissions rise about linearly until reaching a peak, but then falls about linearly. It would be foolish to use a measure of linear association here.

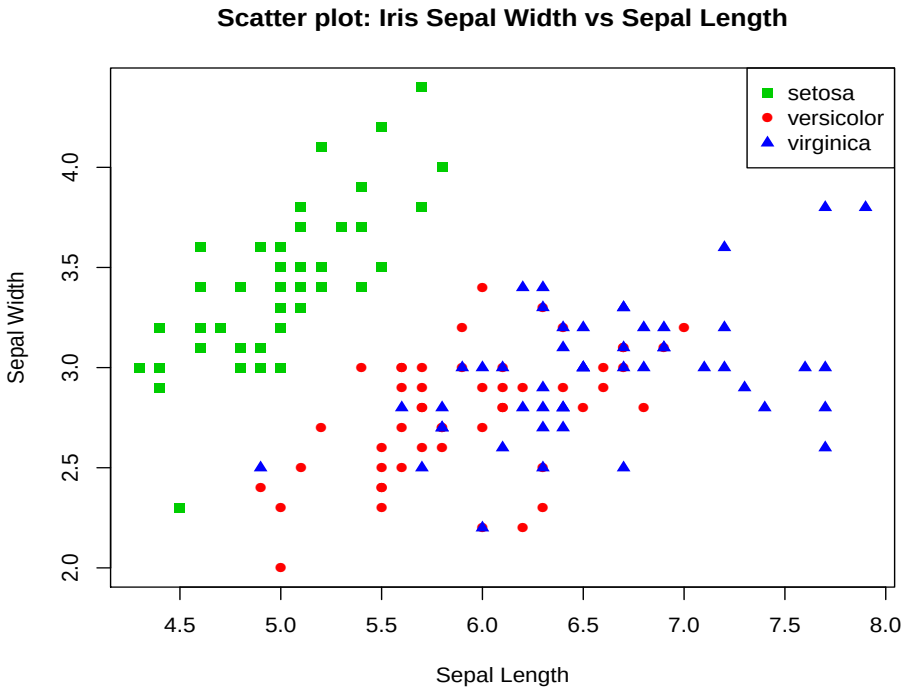


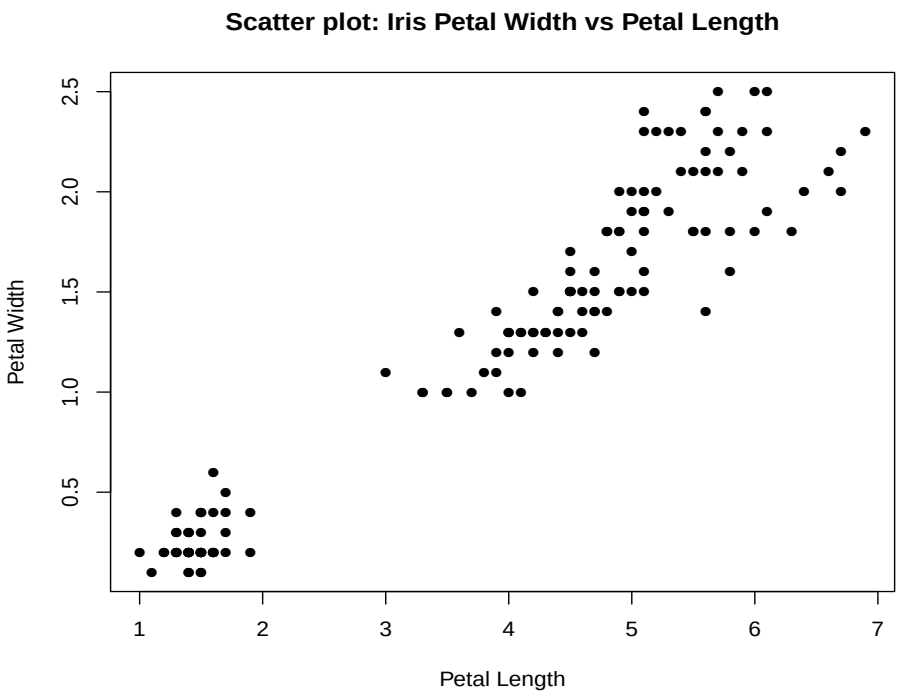
- The next plot is of sepal length against sepal width for 150 iris plants. There is no particular pattern discernible. Maybe there is a cluster to the upper left?

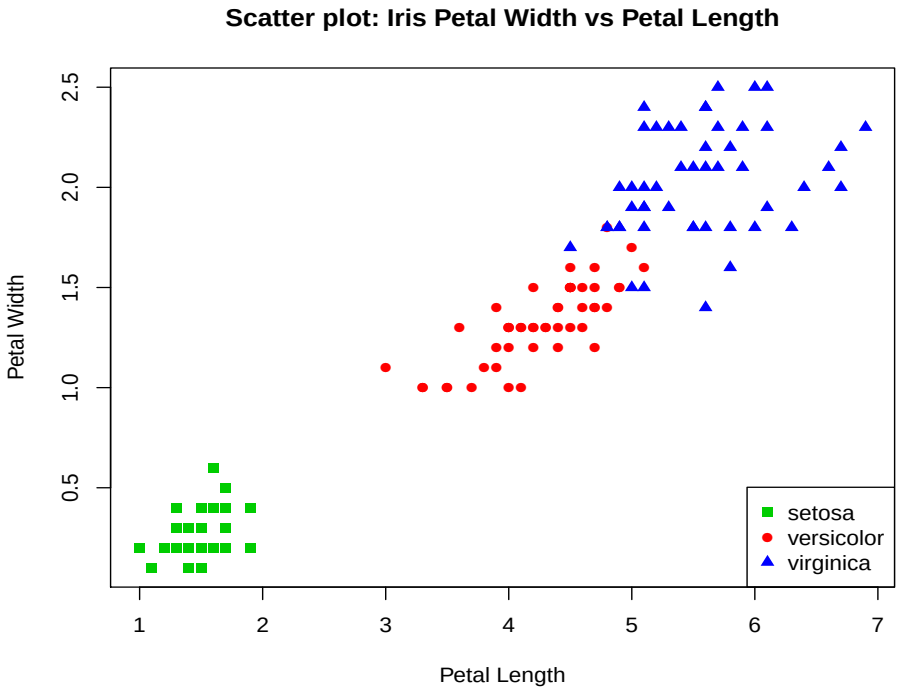
- It may be possible to use a third *categorical* measurement to label a two dimensional plot. Doing so often reveals a good deal of information, in particular we might deduce that some patterns are due to a lurking variable.

- In fact, there were 50 measurements for each of three species, *Setosa* (S), *Versicolor* (V), and *Virginica* (G). Suppose that we now label all the points by species. We obtain instead a labelled plot. This shows clearly the cluster of Setosa to the upper left. The other two species are more mixed up, but the Virginica tend to have slightly larger measurements.


Scatter plot: Iris Sepal Width vs Sepal Length
Sepal Width
Sepal Length
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0


Scatter plot: Iris Sepal Width vs Sepal Length
setosa
versicolor
virginica
Sepal Width
Sepal Length
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0





- The next pair of plots shows even more clearly the value of labelling by a third variable. For this plot, the same 150 iris plants have their petals measured. The first plot implies a linear relationship between petal length and petal width, as it is possible to draw a straight line through the data. However, the

labelled plot reveals the falsity of this argument. Once the measurements have been labelled by species, we see three clusters of points, with no particular pattern visible for any of the clusters.

- This gives an example where the relationship between the two variables plotted cannot be fully understood without knowledge of a third variable. You should bear in mind that an apparently strong relationship in a scatter plot may dissolve when we bring in other variables. Such variables are sometimes called *lurking* variables.

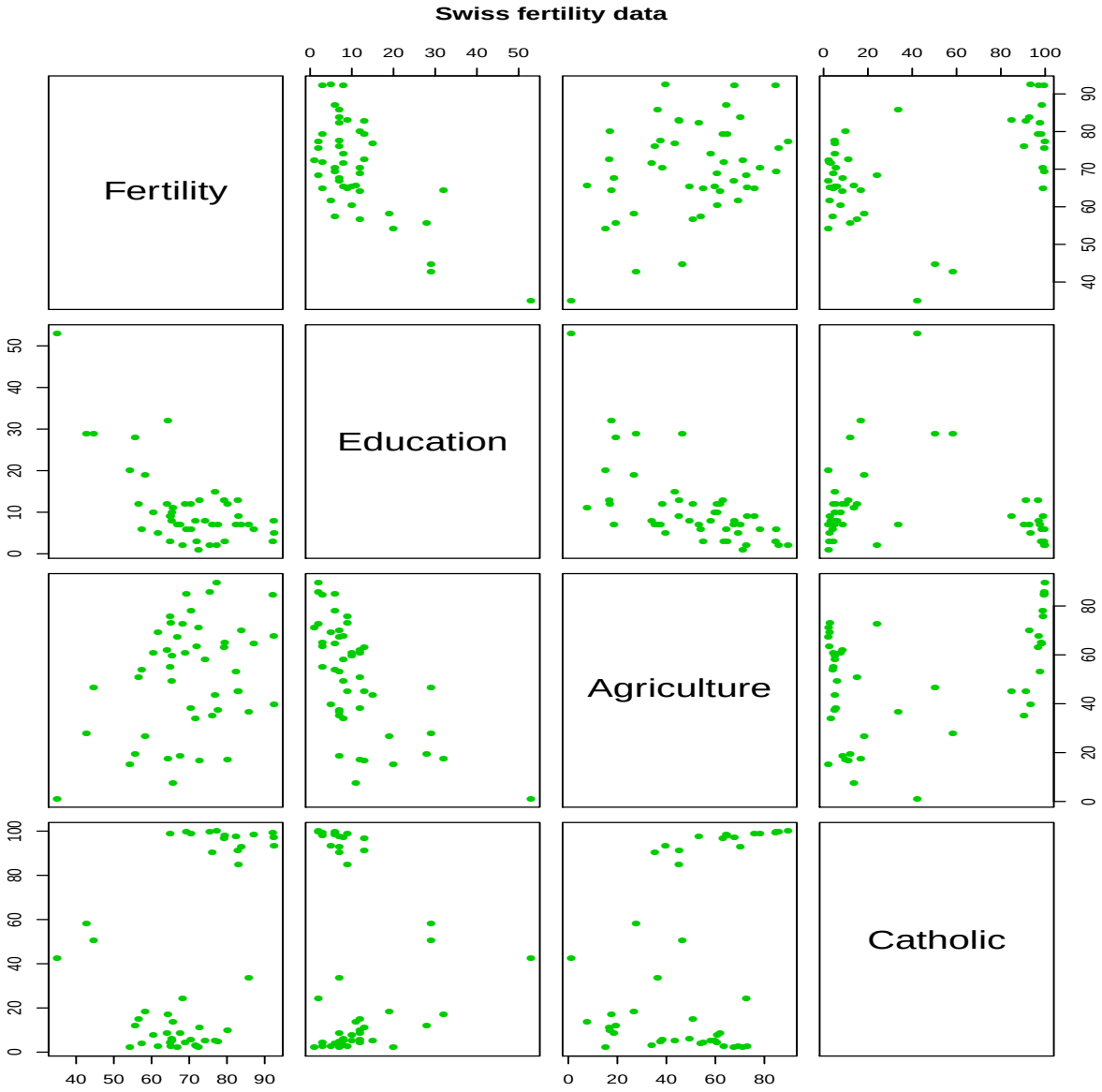
- Multivariate scatter plots can be used to display bivariate relationships between many variables simultaneously.

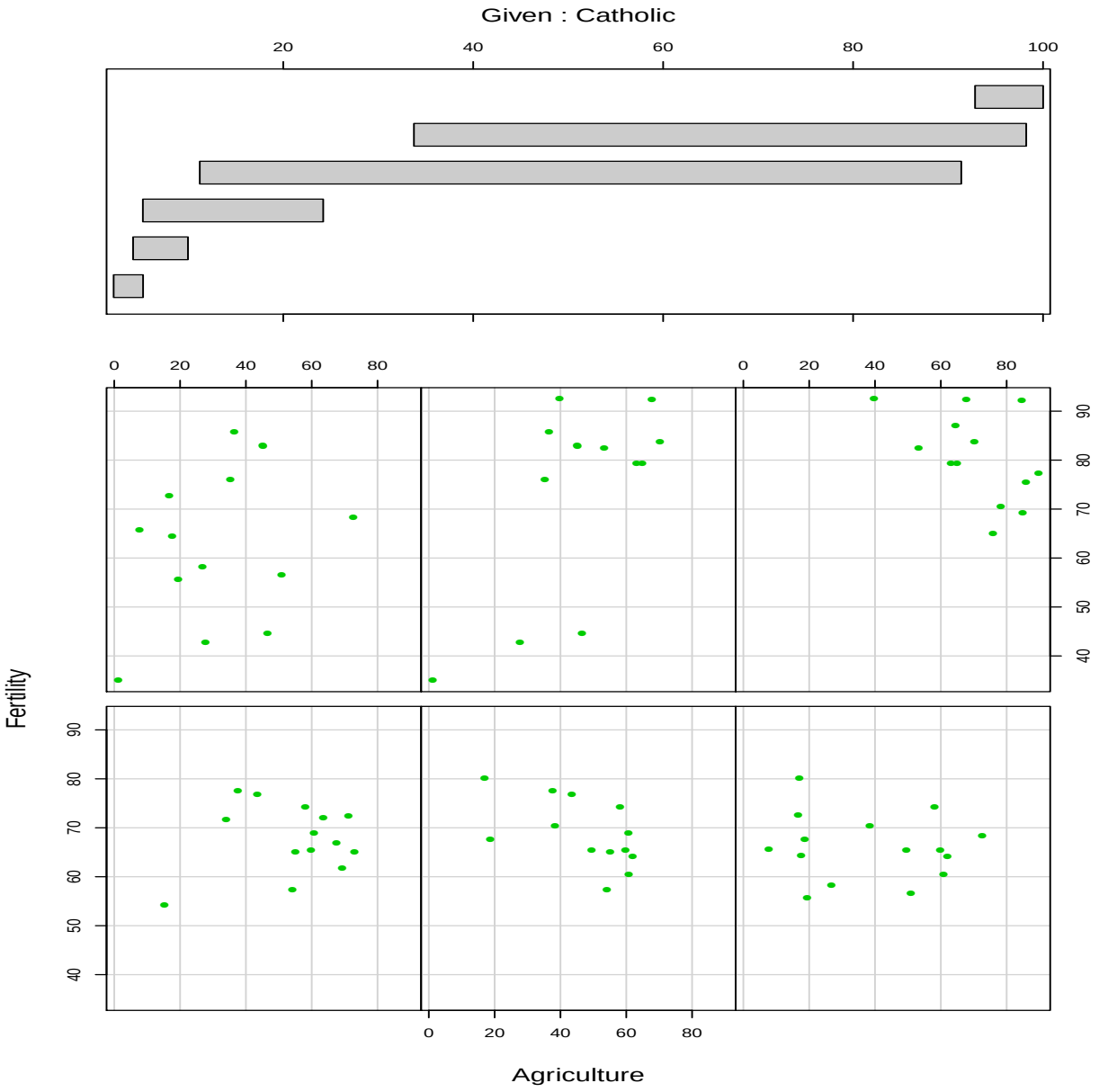
- The Swiss fertility data set is of 47 cases from French speaking provinces of Switzerland around 1888, from the era of industrialization. The variables are:

  - Fertility - standardized fertility measure for each province
  - Agriculture - % population involved in agriculture
  - Examination - draftees receiving highest mark in army exam
  - Education - % of population educated beyond primary school
  - Catholic - % of population who are catholic
  - Infant.Mort - live births living less than 1 year

  The interest is to relate Fertility (number of children born, the response) to the other variables (the predictors). The hypothesis is that the predictors are good proxies for the true causes of high and low fertility. These variables are probably not direct causes of fertility but part of a cultural mix that overall determines fertility rates.

- We see, for example, higher fertility for catholic provinces, higher fertility for agricultural provinces. But we see a higher proportion of catholics associated with a higher proportion of agricultural workers, so perhaps the relationship is more complicated.


Swiss fertility data
Fertility
Education
Agriculture
Catholic



- Coplots may be used to show a separate scatter plot for each value of a third variable.
- Suppose we plot the relationship of proportion of agricultural workers with level of fertility, but controlling for the proportion of catholics in each province. We see no particular relationship between fertility and agricultural proportion in any plot, suggesting that the level of catholicism explains the apparent relationship between fertility and agricultural proportion of workers. Clearly we could study these relationships in far more detail.

3.2 Measuring linear association: Correlation

- Our main interest will lie in measuring the *linear association* between two variables.
- This shows up on a scatter plot by the plotted points seeming mostly to lie along a line drawn in a particular direction.
- We will often use the extra point which we get when we plot the pair $(\bar{x}, \bar{y})$.
- This point, which we call the *point of averages*, locates the *centre* of the scatter plot.
- The numerical measure of linear association is known as the correlation coefficient, r.
- We only use the word *correlation* to mean this measure of linear association.
- Writing the two variables as x and y, and assuming that we have n pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$, the correlation coefficient is defined to be

$$r = \text{“average}'' \text{ of ((x in standard units) times (y in standard units))}.$$

- We could calculate it as

$$r = \frac{1}{n-1} \sum_{i=1}^{n} (\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_y})$$

- where $\bar{x}, \bar{y}, s_x, s_y$ are respectively the means and standard deviations of x and y.
- We divide by $n-1$ rather than n when taking the “average” again for technical reasons.
- The above formula is suitable for computers, etc. For hand calculations, you might prefer to use either of the following equivalent formulae:

$$\begin{aligned} r &= \frac{\frac{1}{n-1}[\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}]}{s_x s_y} \\ r &= \frac{\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}}{\sqrt{[\sum_{i=1}^{n}(x_i^2) - n\bar{x}^2][\sum_{i=1}^{n}(y_i^2) - n\bar{y}^2]}} \end{aligned}$$

- The correlation coefficient tells us two things: the strength of association (strong, weak or zero); and the direction of association (-ve or +ve).
- The strength of association can be gauged by measuring the degree of clustering of points about a line:
  - a strong association is present when the all the points lie on, or very close to, the line
  - a weak or no association is present when the points are scattered randomly over the plot.

Case no.	Manufacturer	Market share, x_i	Sample size, y_i
1	BMW	2.30	1096
2	Citroen	4.54	2177
3	Fiat	2.41	640
4	Ford	21.46	3855
5	Honda	1.74	1397
6	Lada	0.57	122
7	Mazda	0.98	793
8	Mercedes-Benz	1.19	507
9	Mitsubishi	0.60	341
10	Nissan	5.02	2684
11	Peugeot	8.02	2116
12	Proton	0.80	175
13	Renault	5.24	1768
14	Rover	13.38	3264
15	Saab	0.51	568
16	Subaru	0.24	164
17	Suzuki	0.57	97
18	Toyota	2.93	1910
19	Vauxhall	17.09	3897
20	Volkswagen-Audi	4.72	1988
21	Volvo	2.46	1197

- Summary of Car Manufacturer information given:

$$\begin{aligned} n &= 21 \\ \sum x_i &= 96.77 \\ \sum y_i &= 30756 \\ \sum x_i^2 &= 1124.6 \\ \sum y_i^2 &= 74249888 \\ \sum x_i y_i &= 269189 \end{aligned}$$

- Find the five-number summary for bivariate data.

Correlation and the SD Line

- The strength of association or correlation can be gauged by measuring the degree of clustering of points about a line.
- This line is approximately the *SD line*.
- It passes through all points (real and imaginary) that are approximately an equal number of standard deviations away from the point of averages: $(\bar{x}, \bar{y})$.
- The SD line passes through $(\bar{x}, \bar{y})$ and $(\bar{x} + s_x, \bar{y} + s_y)$ when $r > 0$, and $(\bar{x}, \bar{y})$ and $(\bar{x} + s_x, \bar{y} - s_y)$ when $r < 0$.
- For example, a point whose y value is about 1.5 standard deviation distant from the y average *and* whose x value is about 1.5 standard deviation distant from the x average, will fall on the SD line.
- But a point which is 2 standard deviation and 1 standard deviation distant from the y, x averages respectively, will fall off the line.

Interpreting Correlation

- Graphically, we can be misled by the degree of clustering: we have to take into account the size of the standard deviations as well.
- Small standard deviations will make a plot look more clustered than one with large standard deviations, but the correlation may be the same.
- The correlation coefficient r can take any value between -1 and $+1$, inclusive, with positive numbers representing positive correlation, and negative representing negative correlation.
- A correlation of zero means *uncorrelated*: no association.
- Thus the value of r can range from $r = -1$ (perfect -ve correlation) through weaker -ve correlation until $r = 0$ (no correlation) through weak +ve correlation to $r = 1$ (perfect +ve correlation).
- The correlation coefficient is *scale-free*.
- That is, if you add the same number to all the entries of one distribution, and/or multiply them by another number, the correlation coefficient is unchanged.
- This is because the correlation coefficient is determined from their standardised values.
- We can also change the order of the variables without changing the correlation coefficient.
- An association is positive if high values of one variable are associated with high values of the other, and low values associated with low.
- The association is negative when the reverse is true: high values of one variable associated with low values of the other, and low associated with high.
- For example, A-level results are positively correlated with degree results (high-with-high; low-with-low).
- Sales of hot-water bottles are negatively correlated with temperature (high with low; low with high).

Correlation: Difficulties

- It is difficult to interpret correlation numerically: a value of $r = 0.7$ can mean different strengths of association for different numbers of pairs of points.
- Generally, the more points there are, the lower the correlation has to be to indicate association.
- Strong correlations (near ± 1) are quite rare, especially if the size of the distribution is large.
- Weak correlations (near 0) are common.
- Quite small correlations can indicate a linear association if the size of distributions is large, though it is difficult to set a hard and fast rule.
- r doesn't tell us by how much a change in one variable effects change in the other - it just tells us the rough direction and degree of change.

- For two separate scatter plots, the correlation coefficient might be numerically the same, but the x, y relationship may be very different.
- r is not resistant to extreme values.
- The correlation coefficient measures *linear* association, and so can be misleading when there is *nonlinear* association, or when their are outliers present.

Association does not imply cause and effect

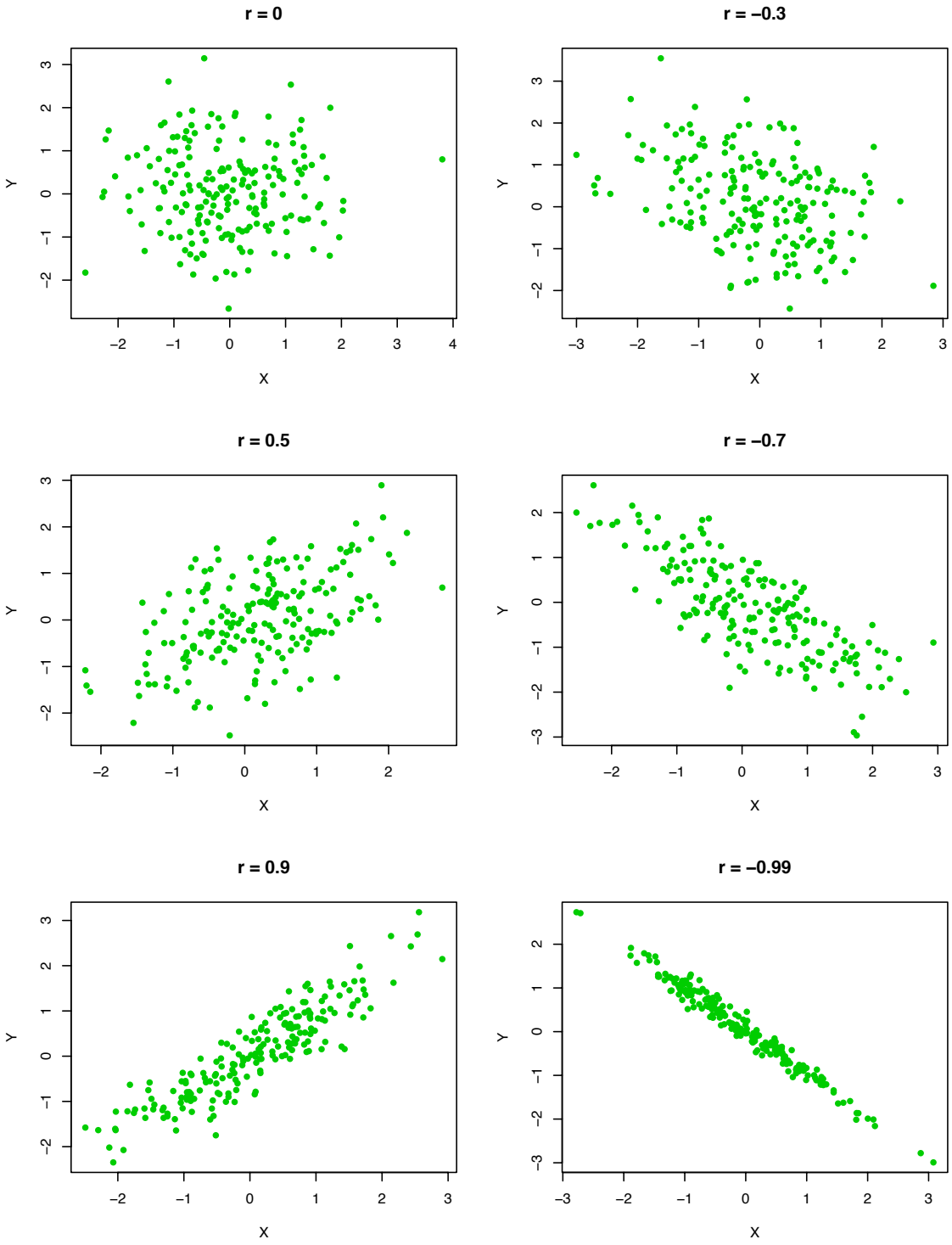
- A strong correlation between two variables suggests that knowing one helps in predicting the other.
- In particular, we can often label one variable as being *independent* and the other as *dependent*.
- The value of the dependent variable is thought to be dependent to some extent upon the value of the independent variable. Here are some examples:

Independent	Dependent
Engine size	mileage per gallon
Volume of baby's cry	sleep per parent
Boringness of lecture	sleep per person

- However we must be careful, because evidence of association is not necessarily evidence of causation.
- If we measure a high correlation, it does not mean that we establish and measure cause.
- In particular, there may be other, confounding, variables at work.
- This leads to the well know mantra:

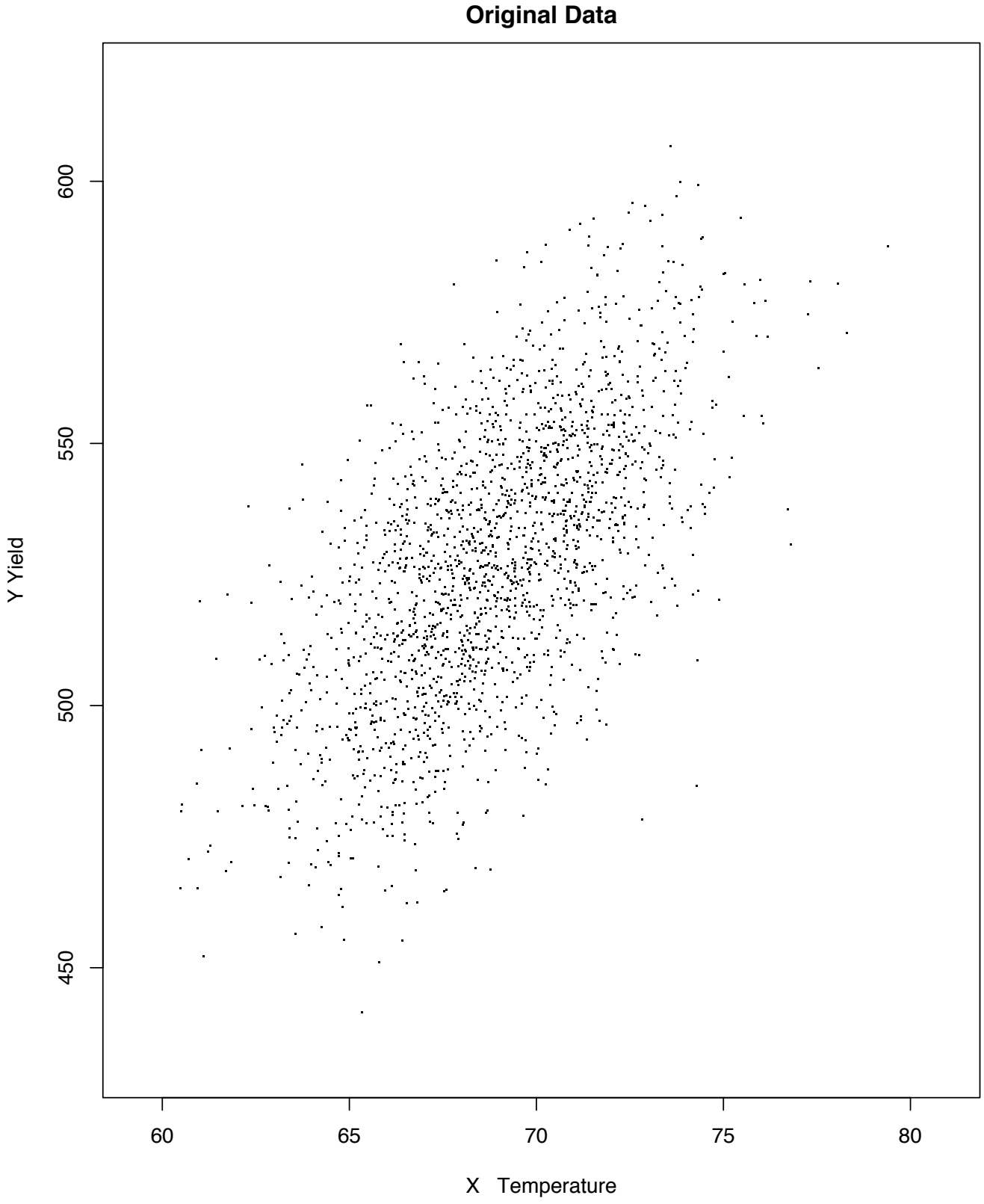
CORRELATION DOES NOT IMPLY CAUSATION!

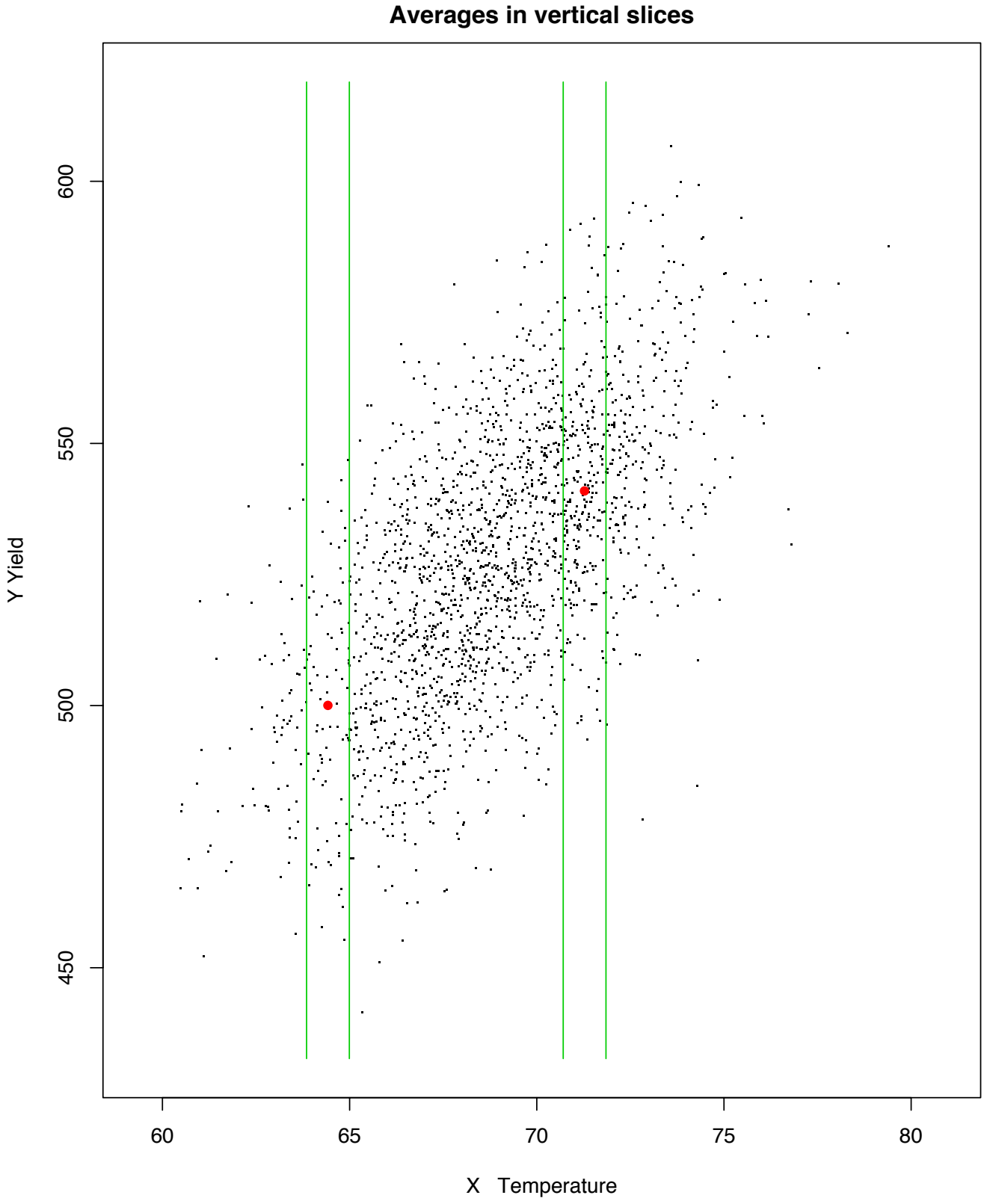
- Correlations based upon averages often mislead: these are sometimes called *ecological* correlations.
- Assuming that what holds true for the group also holds true for the individual is often called the *ecological fallacy*.
- The problem is that the underlying averages have a lot less spread than the original measurements, and so tend to overstate the amount of clustering, and thereby the degree of correlation.
- As an example, consider the swiss fertility data. The correlations which you can see are based on averages for provinces, and so will overstate the level of correlation.
- The level of individual variation is understated.

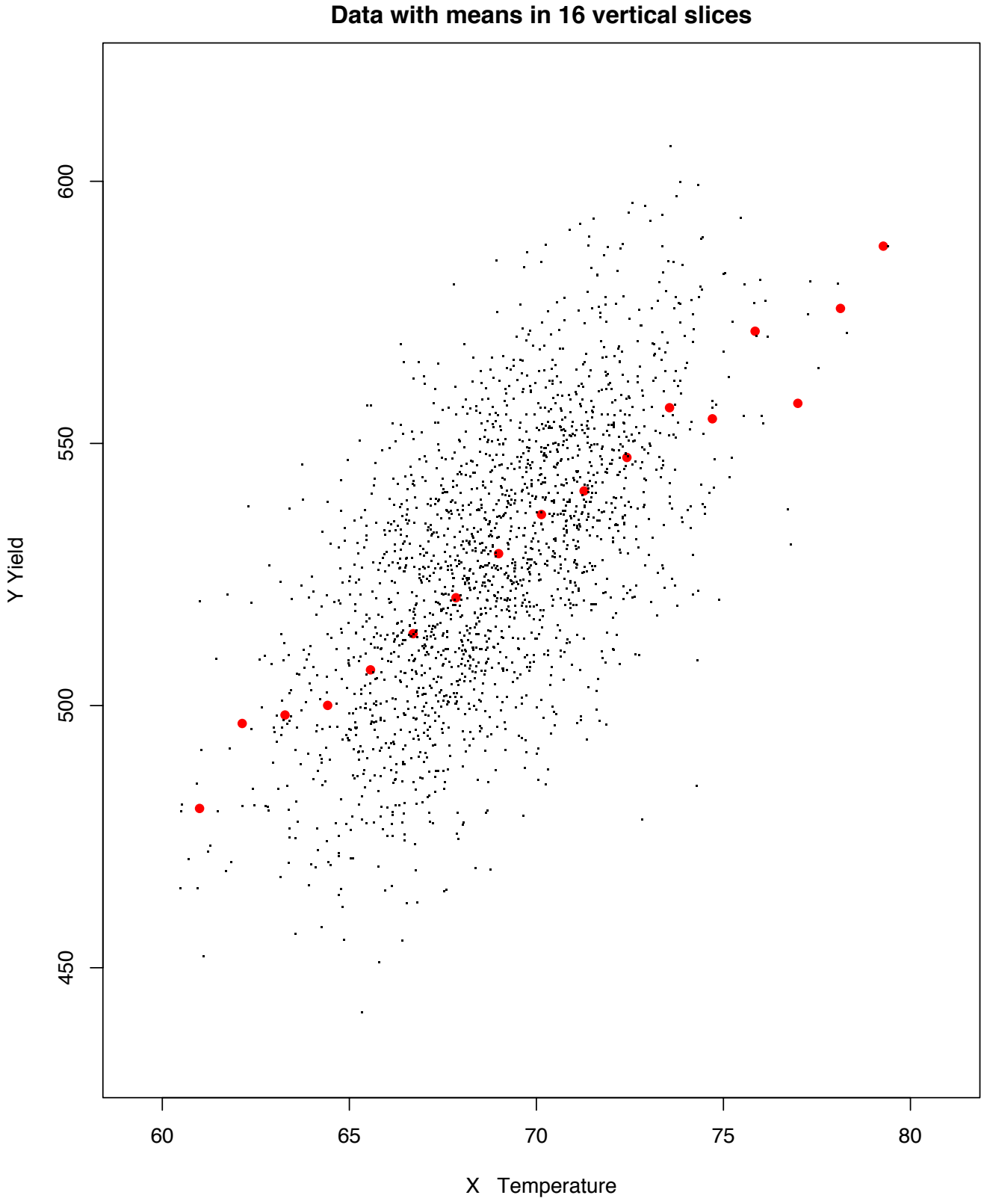

r = 0
r = −0.3
r = 0.5
r = −0.7
r = 0.9
r = −0.99
X
Y

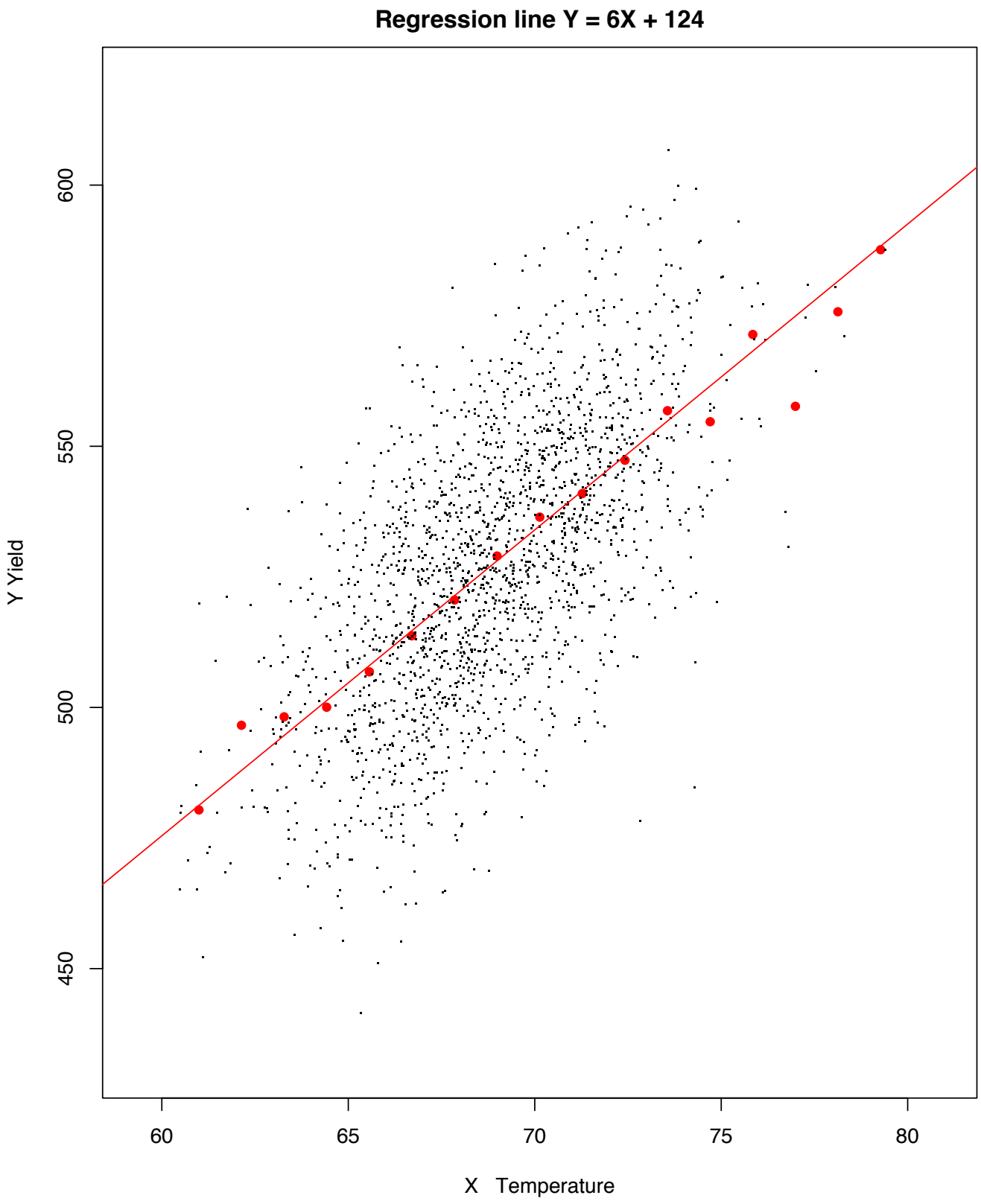
3.3 Prediction from linear association: simple linear regression

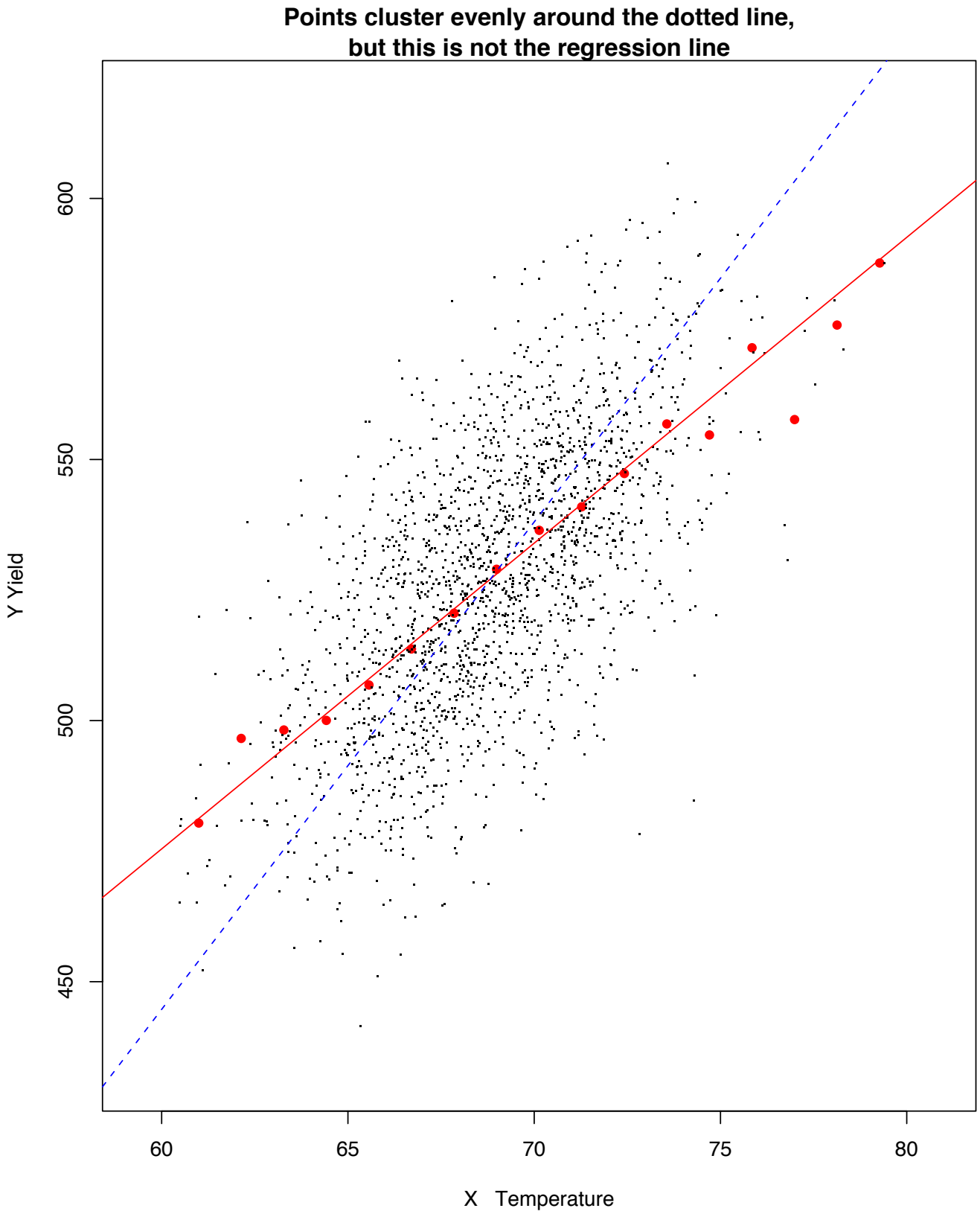
- Suppose that we believe that two variables are linked causatively.
- Generally we let x be the *independent* or *explanatory* variable, and let y be the *dependent* or *response* variable, so that changes in x "cause" changes in y.
- We will be concerned with predicting y from a given value of x.
- If we don't know x, our best guess for y is probably $\bar{y}$. Otherwise we can do a little better.
- Where should we draw the prediction line on the following graph?
- Many people make the error of assuming that the line about which the points cluster (the *SD line*) is the regression line.
- Suppose instead we think about predicting within a narrow vertical slice, for a fixed value of x.
- It would seem reasonable to take all the y values in that slice, calculate their average (mean) and use that as our prediction.
- If there are several y values associated with an x-value, we can draw the *graph of averages*.
- For each x value we calculate the average of the associated y values, and then connect the points $(x_i, \bar{y}_i)$.
- Otherwise, we could divide up the x axis into a number of strips, and take the average in each strip.
- This gives instead an average y for similar x values.
- Connecting the averages in some way gives us a better method of prediction.
- The closer the graph of averages to a straight line, the better the *linear regression fit*.
- The further the graph of averages from a straight line, the less appropriate the fit.
- Sometimes the graph of averages can be used for prediction when the association is not linear.
- Similarly, it is also possible to calculate the medians in each vertical strip. This is often very useful and informative, but we shan't pursue this matter.
- There is also a relationship with what are called generalized additive models and smoothers such as loess.

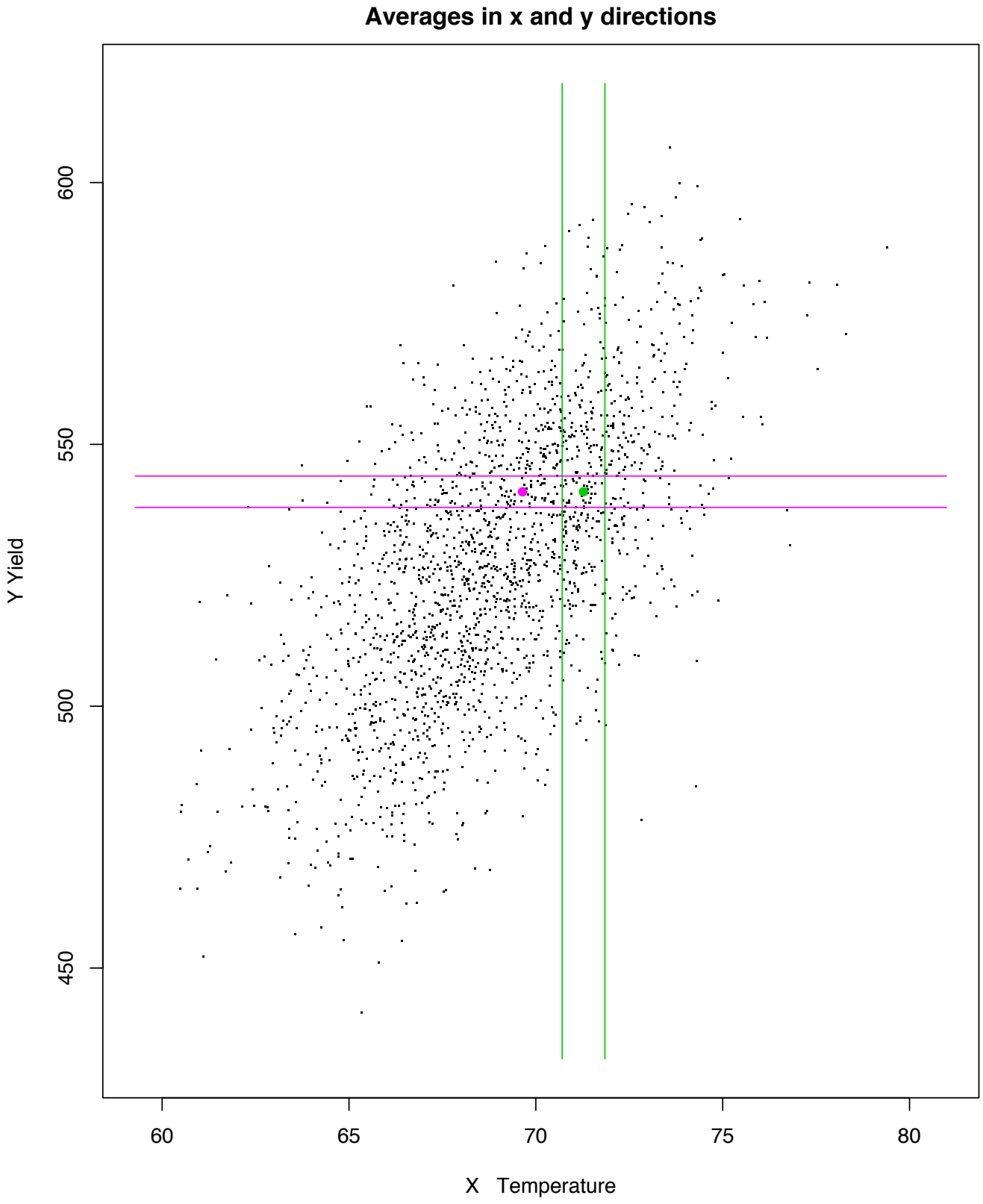

Original Data
Y Yield
600
550
500
450
60
65
70
75
80
X Temperature


Averages in vertical slices
Y Yield
600
550
500
450
60
65
70
75
80
X Temperature


Data with means in 16 vertical slices
Y Yield
600
550
500
450
60
65
70
75
80
X Temperature


Regression line Y = 6X + 124
Y Yield
X Temperature
450
500
550
600
60
65
70
75
80


Points cluster evenly around the dotted line,
but this is not the regression line
Y Yield
600
550
500
450
60
65
70
75
80
X Temperature


Averages in x and y directions
Y Yield
600
550
500
450
60
65
70
75
80
X Temperature

3.4 Method of least squares

- The key concept here is that our concern is with vertical slices and prediction within them.
- The errors we make in vertical slices are the vertical distances from our observed points to their predictions.
- We model the cause-and-effect relationship as $y = a + bx$, a linear relationship.
- We usually never can know truly what a and b are; they must be estimated, and there are many ways of doing this.
- The most common is the method of *least squares*, where we fit a straight line through the points so as to minimise the average squared distance from the points to the straight line.
- Suppose that the straight line will actually pass through the point $(x_i, \hat{y}_i)$.
- Then the vertical distance from the ith point to the straight line is $y_i - \hat{y}_i$.
- We choose the straight line which minimises the sum of these squared distances:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- We can show that the line achieving this minimum is given by $y = \hat{a} + \hat{b}x$, where

$$\begin{aligned} \hat{b} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}}{\sum_{i=1}^{n}(x_i^2) - n\bar{x}^2} \\ \hat{a} &= \bar{y} - \hat{b}\bar{x} \end{aligned}$$

- Outline Proof. Begin by minimising:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

- We then obtain simultaneous equations in a, b, which are solved to give the result.
- The technical term for this particular line is the regression line of y (depending) on x.
- The slope of the regression line, b, and the correlation coefficient r are related as

$$b = r\,\frac{s_y}{s_x}$$

- To predict y from x, simply insert the value of x into the regression equation and calculate the y prediction.
- Beware that prediction outside the range of x-points used to calculate the regression (such prediction is called *extrapolation* as opposed to *interpolation*) may well be misleading.

3.5 SD Line and Regression Line

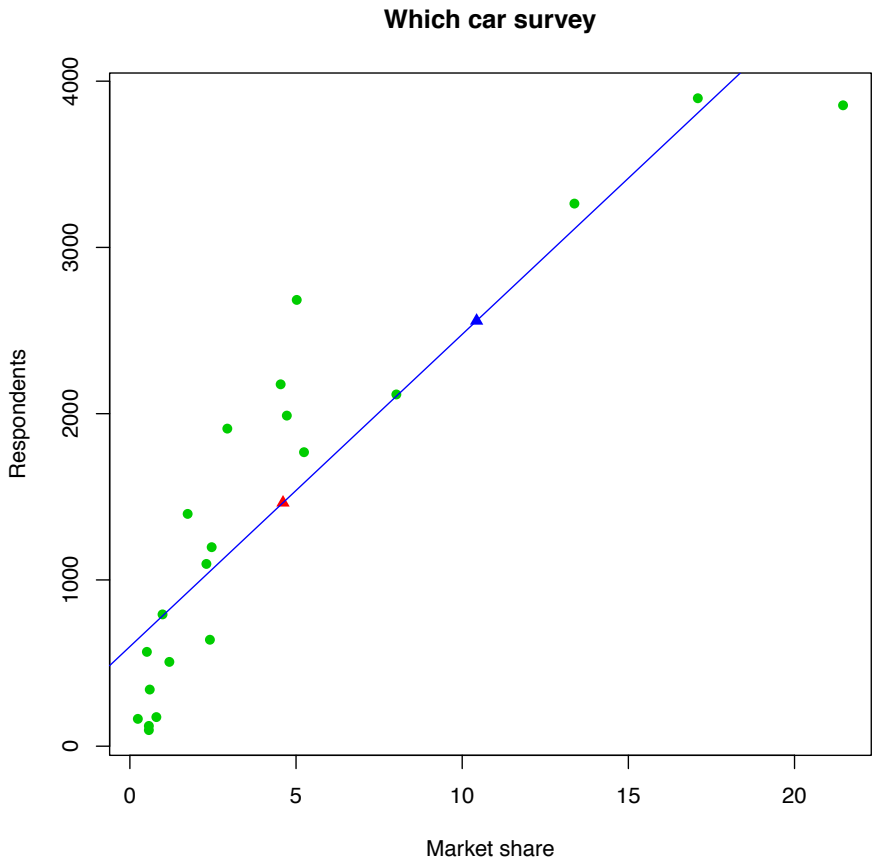
- Recall that the *SD line* passes through the point of averages:

$$(\bar{x}, \bar{y}) \quad \text{and} \quad (\bar{x} + s_x, \bar{y} + \text{Sign}(r) s_y),$$

- where Sign(r) is the sign (plus or minus 1) of r, or zero when r is zero.
- The slope of this line is Sign(r) s_y/s_x.
- In contrast, the regression line passes through the point of averages:

$$(\bar{x}, \bar{y}) \quad \text{and} \quad (\bar{x} + s_x, \bar{y} + r s_y),$$

- The slope of this line is $r\ s_y/s_x$.
- Notice that the slope of the SD line is thus steeper than the slope of the regression line.



- The main difference between the SD lines and the regression line lies in the fact that for the SD line, the distance from each point is measured approximately diagonally to the line.
- For the regression line, the distance is measured vertically.

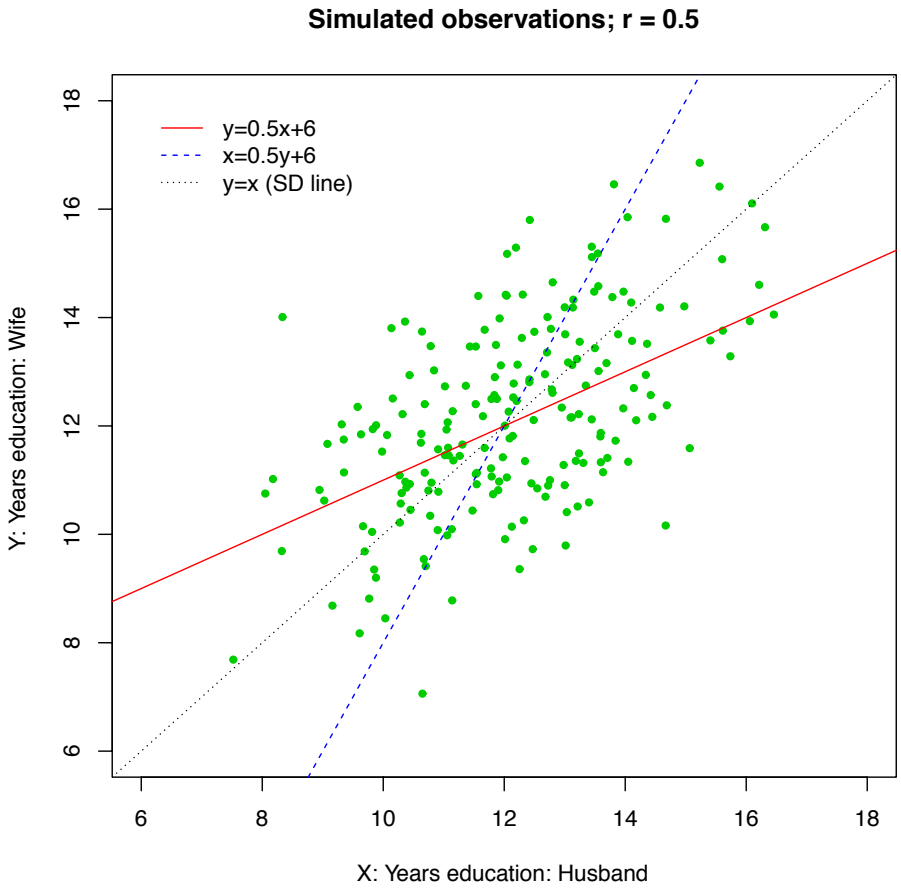
- The SD line is roughly the closest *diagonally* to the points.
- The regression line is roughly closest *vertically* to the points.
- The regression line and the SD line coincide when the correlation is perfect ($r = 1, r = -1$) and when the correlation is zero ($r = 0$).
- Rarely, we might be interested in obtaining a prediction of an x value from a new y value.
- Here we try to obtain a line such that the points are *horizontally* close to the line.
- This is called the regression of x (depending) on y.
- It is suitable only when the new point we are trying to predict has been sampled in the same way as before.
- The interpretation or meaning of the prediction of x from y is a bit more subtle.
- Mainly we will deal with the standard (vertical) kind of regression.

3.6 Regression effect and the regression fallacy

- Example. In one study, the correlation between the educational level of husbands and wives in a certain town was about $r = 0.5$.
- Both husbands and wives averaged 12 years of schooling, with a standard deviation of 3 years.

  1. Predict the educational level of a woman whose husband has completed 18 years of schooling.
  2. Predict the educational level of a man whose wife has completed 15 years of schooling.
  3. Apparently, well-educated men marry women who are less well educated than themselves, but the women marry men with even less education. How can this be explained?

- The regression of the educational level of women y (depending) on the educational level of men x, thus has $\hat{b} = (0.5)(3)/(3) = 0.5$, and $\hat{a} = (12) - (0.5)(12) = 6$.
- Therefore the regression line is $y = \frac{1}{2}x + 6$.
- When $x = 18$ the predicted value of y is 15.
- However, the regression of the educational level of men x (depending) on the educational level of women y is the same: $x = \frac{1}{2}y + 6$.
- When $y = 15$ the predicted value of x is 13.5.
- This apparent paradox is explained by the *regression effect*.
- (The SD line in this example has slope $s_y/s_x = 1$ and intercept $\bar{y} - s_y\bar{x}/s_x = 0$.)



- The *regression effect* is the term for spread about the regression line - really just spread about the average in each vertical strip.
- An example is test-retest situations, where there is a tendency for those who do poorly to show some improvement and for those who do well to fall back a little *on average*.
- The *regression fallacy* is the fallacy that the regression effect is due to something important, rather than simply spread about the regression line.
- It arises because we confuse the diagonal clustering about the SD line with the vertical clustering about the regression line.
- The regression fallacy is also frequently confusing in the case of *regression to mediocrity* (Galton's term), which asserts essentially that inherited characteristics tend to revert to average.
- For example:
  - The IQs of the children of geniuses tend to be very high (because there is an association) but not so high as their parents (because of regression to mediocrity),
  - The children of ESN parents tend to also have low IQs, but be brighter than their parents.

- Bear in mind the graph of averages whenever these issues become confusing.
- In the following example, if we want to predict y when $x = 71.3$, we find the y average in that vertical slice is about $y = 541$. If we want to predict x when $y = 541$, we find that the x average in that horizontal slice is about $x = 69.6$.

3.7 Residuals and standardized residuals

- Usually the regression fit won't be exact: there will be some sort of error involved, usually because the linear relationship is too simple to capture the full relationship.
- The imperfections are absorbed into an error term which we hope resemble chance errors:

$$\begin{array}{rcll} y_i & = & \hat{y}_i & + \ \hat{\epsilon}_i \\ \text{observed value} & = & \text{fitted value} & + \ \text{estimated chance error} \end{array}$$

- Fitted value is another phrase for value of the regression line at that point.
- The regression line passes through the points $(x_i, \hat{y}_i)$ rather than (x_i, y_i), where $\hat{y}_i = \hat{a} + \hat{b}x_i$ is our notation for the predicted value for y at x_i.
- For each point on the scatter plot we calculate the *residual* at each point as $\hat{\epsilon}_i = y_i - \hat{y}_i$.
- This is the vertical distance from the data point to the regression line.
- The least squares method has the property that the residuals ϵ_i add to zero (so that the average residual $\bar{\epsilon}$ is zero).
- This is a useful check.
- It is useful (but tedious by hand) to standardize the residuals.
- To do this, we simply divide them by their standard deviation, which we denote s_ϵ, where

$$s_\epsilon = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} \hat{\epsilon}_i^2}$$

  (Remember, the average residual is zero.)
- We can show that we may derive this to be:

$$s_\epsilon = s_y\sqrt{1-r^2}$$

- Proof of $s_\epsilon = s_y\sqrt{1-r^2}$. Write

$$\sum \hat{\epsilon}_i^2 = \sum (y_i - \hat{a} - \hat{b}x_i)^2$$

  Put $\hat{a} = \bar{y} - \hat{b}\bar{x}$ to give

$$\sum ([y_i - \bar{y}] - \hat{b}[x_i - \bar{x}])^2$$

  and expand to give

$$\sum [y_i - \bar{y}]^2 + \hat{b}^2 \sum [x_i - \bar{x}]^2 - 2\hat{b} \sum [y_i - \bar{y}][x_i - \bar{x}]$$

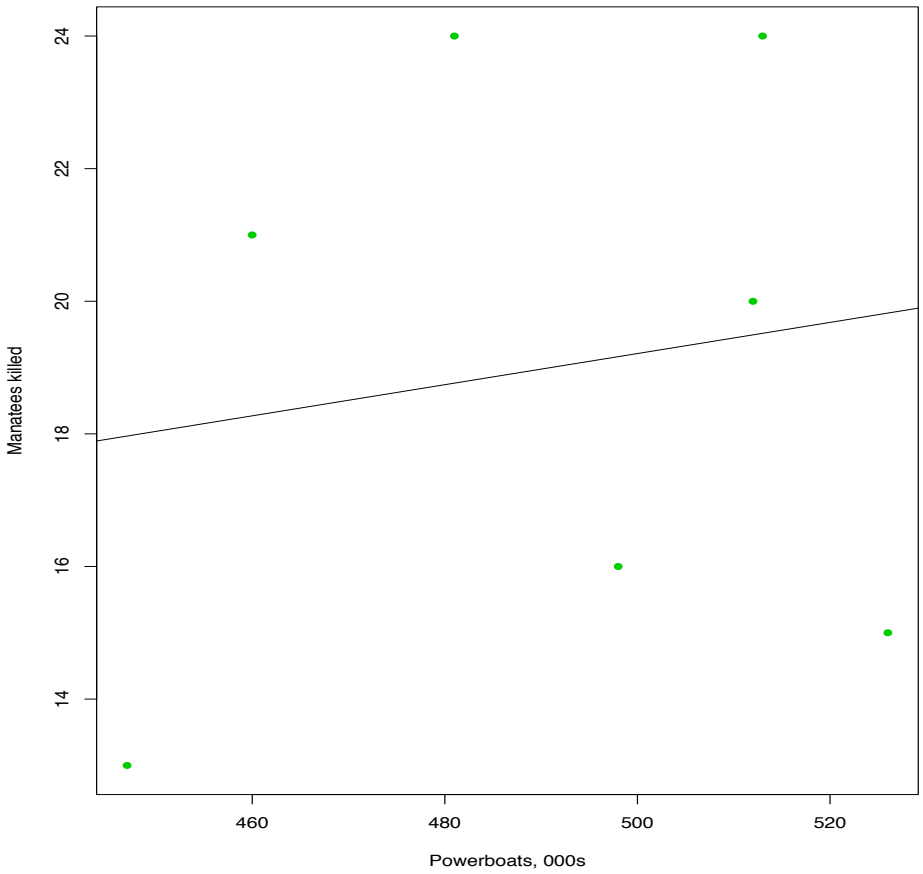
  Now remember that $\hat{b} = rs_y/s_x$ and that

$$\sum [y_i - \bar{y}][x_i - \bar{x}] = (n-1)rs_x s_y,$$

  and the result follows.

- s_ϵ is also known under the special name **the root mean square error for the regression**, abbreviated rmsr.
- (Technical note. For inferential purposes we prefer to divide by $(n-2)$ rather than $(n-1)$ in the definition of s_ϵ. For large sample sizes, there won't be much difference and for this part of the course we continue to use $(n-1)$.)
- Notice that because it is a measure of "typical" error in the vertical, y direction, it is measured in the same units as y.
- If we have approximately Normal residuals, the standardized residuals should follow a standard Normal distribution.
- For example, we would expect few of the values (about 5%) to be larger than ± 2.
- Let's return to our powerboat/manatees example:

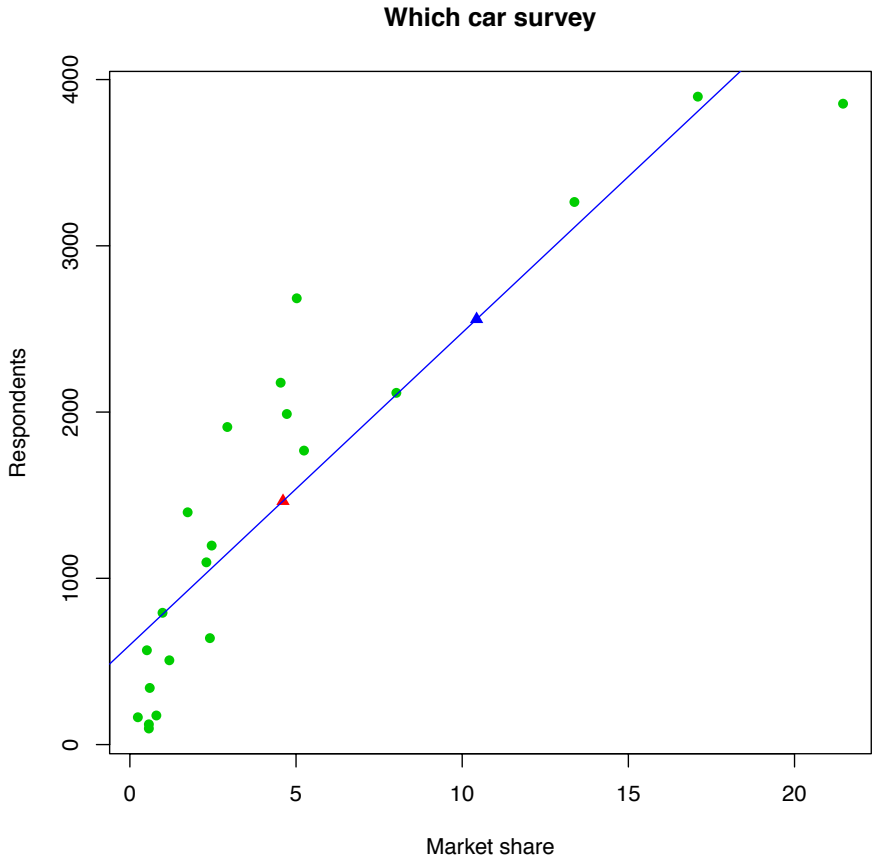
	boats x_i	manatees y_i	prediction fit $\hat{y}_i$	residuals $\hat{\epsilon}_i = y_i - \hat{y}_i$	standardised residuals $(\hat{\epsilon}_i - \bar{\epsilon})/s_e = \hat{\epsilon}_i/s_e$
1	447	13	18.0	-5.0	-1.1
2	460	21	18.3	2.7	0.6
3	481	24	18.8	5.2	1.2
4	498	16	19.2	-3.2	-0.7
5	513	24	19.5	4.5	1.0
6	512	20	19.5	0.5	0.1
7	526	15	19.8	-4.8	-1.1

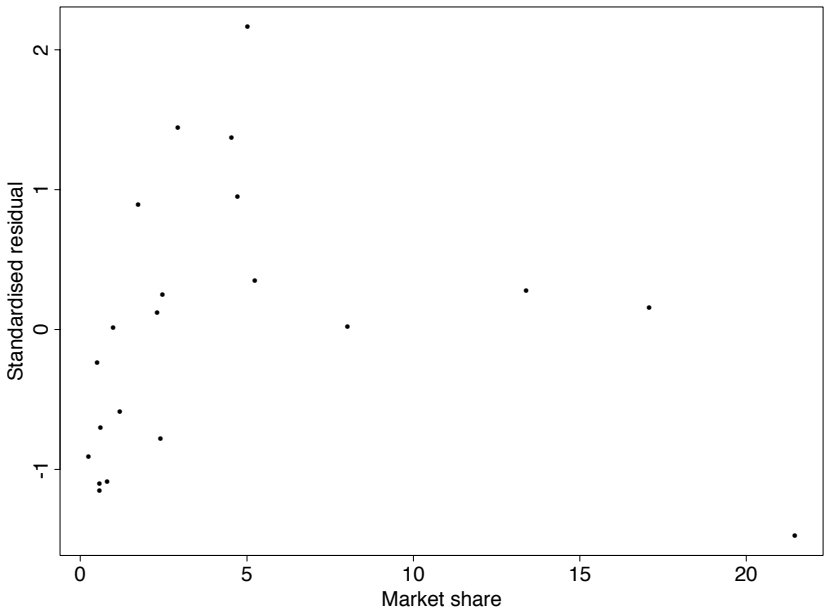


Case	Manufacturer	Market share x_i	Sample y_i	Regression prediction $\hat{y}_i = 187.8x_i + 598.8$	Residual $y_i - \hat{y}_i$	Standardised residual $\hat{\epsilon}_i/s_e$
1	BMW	2.30	1096	1031.1	64.9	0.12
2	Citroen	4.54	2177	1451.8	725.2	1.37
3	Fiat	2.41	640	1051.7	-411.7	-0.78
4	Ford	21.46	3855	4629.6	-774.6	-1.47
5	Honda	1.74	1397	925.9	471.1	0.89
6	Lada	0.57	122	706.2	-584.2	-1.10
7	Mazda	0.98	793	783.2	9.8	0.01
8	Merc.-Benz	1.19	507	822.6	-315.6	-0.59
9	Mitsubishi	0.60	341	711.8	-370.8	-0.70
10	Nissan	5.02	2684	1541.9	1142.1	2.16
11	Peugeot	8.02	2116	2105.4	10.6	0.02
12	Proton	0.80	175	749.4	-574.4	-1.09
13	Renault	5.24	1768	1583.3	184.7	0.35
14	Rover	13.38	3264	3112.0	152.0	0.28
15	Saab	0.51	568	694.9	-126.9	-0.24
16	Subaru	0.24	164	644.2	-480.2	-0.91
17	Suzuki	0.57	97	706.2	-609.2	-1.15
18	Toyota	2.93	1910	1149.4	760.6	1.44
19	Vauxhall	17.09	3897	3808.8	88.2	0.16
20	VW-Audi	4.72	1988	1485.6	502.4	0.95
21	Volvo	2.46	1197	1061.1	135.9	0.25

3.8 Residual plots

- For model checking, we can draw plots of the raw residuals versus x.
- Sometimes we add a line with slope zero and intercept zero to aid interpretation.
- Using raw residuals or standardized residuals will show the same patterns, but the standardized residuals have the advantage of having a readily interpretable y-axis.
- It is sometimes also useful to plot residuals versus other variables: this is an important mechanism for detecting *lurking variables*.
- The "ideal" residual plot shows no pattern at all, and the points should seem scattered randomly about the plot.
- The regression technique is such that if we now determined the correlation for the x values versus the residuals, we will find a correlation of zero (if not our calculations are WRONG!).
- A pattern (particularly a curve) amongst the residuals indicates non-linear association; and very large residuals tend to indicate outliers.





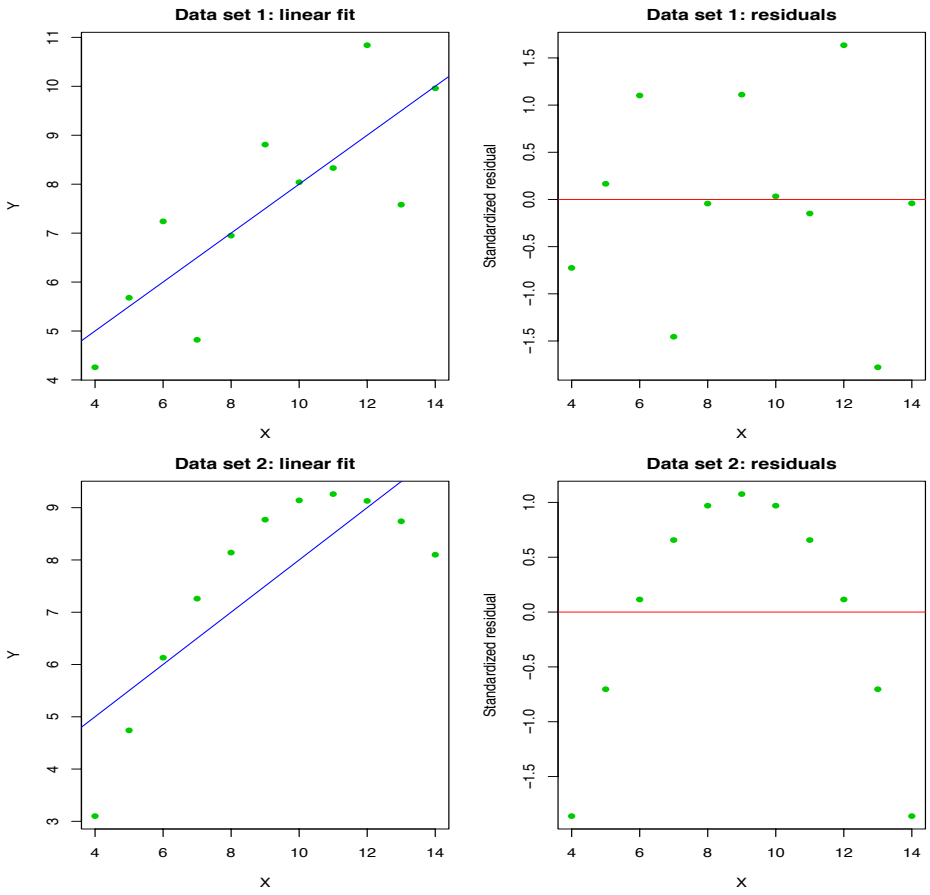
- For the Which? car manufacturer data, the residual plot shows some curved association in the residuals.
- This implies that the linear relationship is inadequate to capture the relationship between market share and number of respondents to the questionnaire.

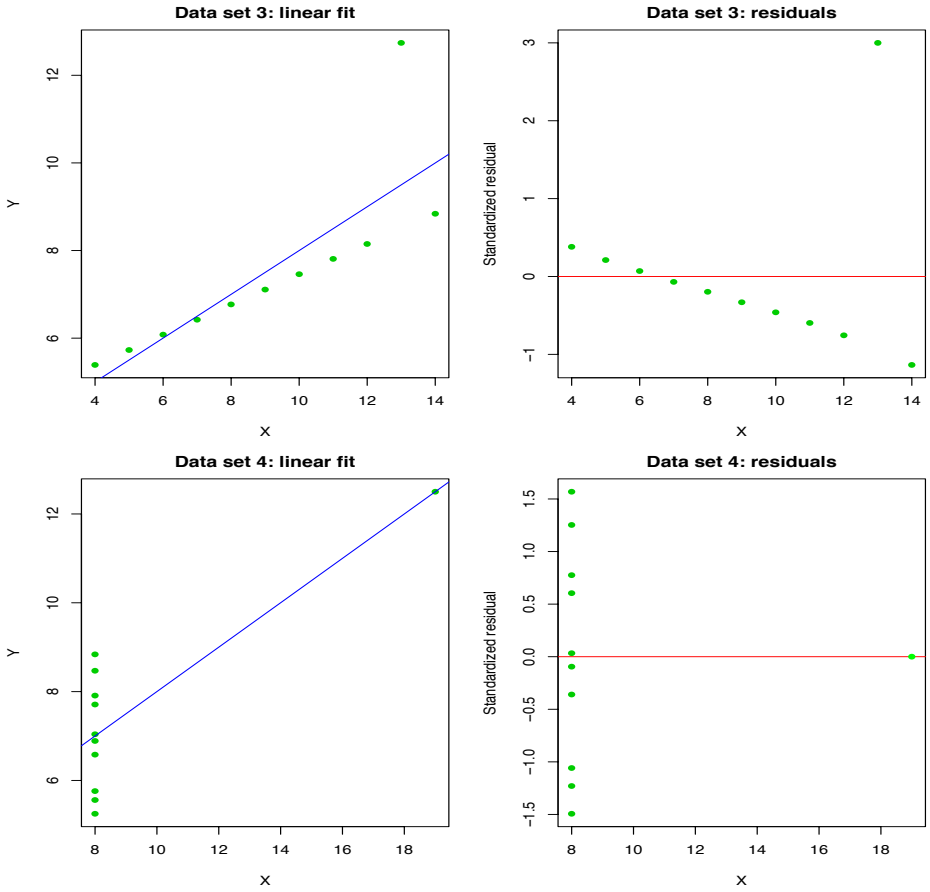
Residual plots: Identifying problems.

- The Anscombe data sets are four sets of data constructed by F.W. Anscombe to have the same means, standard deviations, and correlations.
- As such, the regression equation is the same for all 4 data sets.
- They each have $\bar{y} = 7.5$, $\bar{x} = 9$, $s_y = 2.03$, $s_x = 3.32$, $r = 0.8164$.
- Due to this they all have the same regression line: $y = 3 + 0.5x$.
- The correlation coefficient $r = 0.8164$ implies quite strong positive linear association.
- What do the residual plots tell us?

	Data set 1		Data set 2		Data set 3		Data set 4	
i	y_1	x_1	y_2	x_2	y_3	x_3	y_4	x_4
1	4.26	4	3.10	4	5.39	4	7.04	8
2	5.68	5	4.74	5	5.73	5	6.89	8
3	7.24	6	6.13	6	6.08	6	5.25	8
4	4.82	7	7.26	7	6.42	7	7.91	8
5	6.95	8	8.14	8	6.77	8	5.76	8
6	8.81	9	8.77	9	7.11	9	8.84	8
7	8.04	10	9.14	10	7.46	10	6.58	8
8	8.33	11	9.26	11	7.81	11	8.47	8
9	10.84	12	9.13	12	8.15	12	5.56	8
10	7.58	13	8.74	13	12.74	13	7.71	8
11	9.96	14	8.10	14	8.84	14	12.50	19

Table 7: Anscombe data sets. These four data sets have approximately the same summary statistics. For each, $\bar{y} = 7.5$, $\bar{x} = 9$, $s_y = 2.03$, $s_x = 3.32$, $r = 0.8164$. As such, they have the same regression line: $y = 3 + 0.5x$.





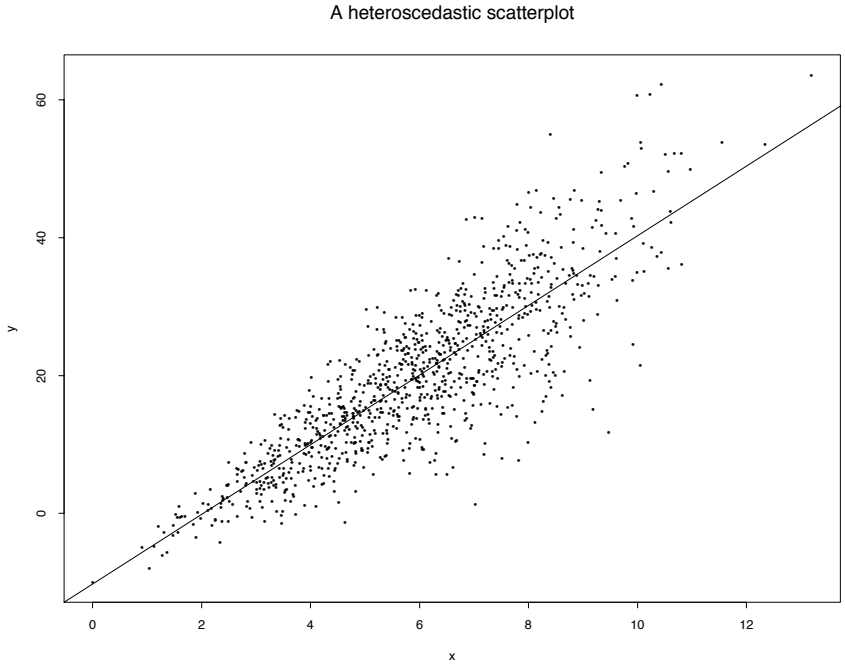
- The residual plots for these data sets shown that only data set 1 is free of problems:
- Data set 1 shows the ideal situation: a random scatter of residual points.
- Data set 2 shows a quadratic relationship which is not captured by linear regression.
- Data set 3 shows a probable outlier.
- Data set 4 shows no relationship between x and y except as driven by the outlying value which essentially determines the regression line.
- The key idea is that the summary statistics $\bar{y}, \bar{x}, s_y, s_x, r$ don't reveal these difficulties. Plots do.

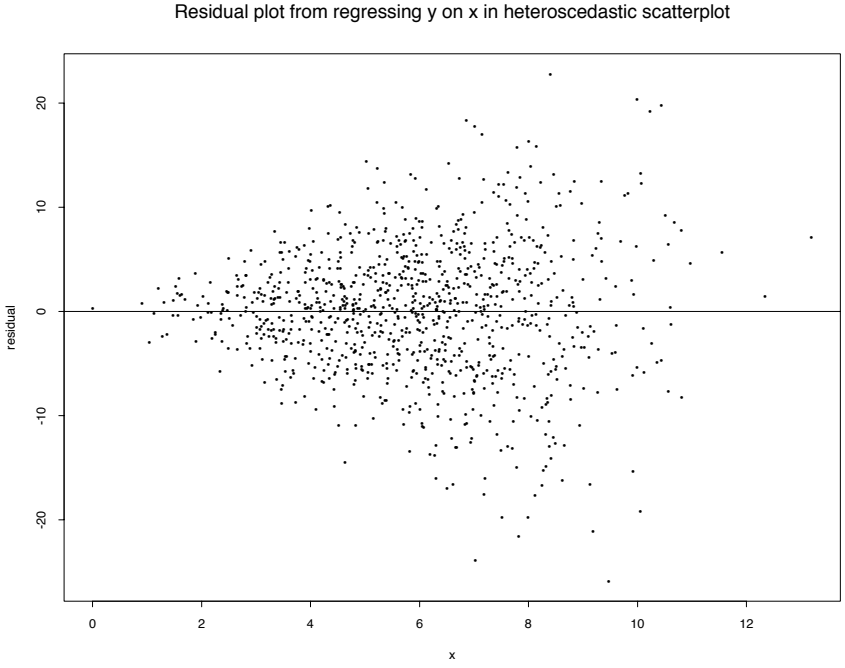
3.9 Heteroscedasticity

Residual plots: Heteroscedasticity

- If the residuals tend to increase in size with x (i.e. from left to right) this indicates that the data are *heteroscedastic* (spread differently) rather than *homoscedastic* (spread similarly).
- If we consider the points in each vertical strip as constituting a separate distribution, the spreads will be similar for each distribution if the data is homoscedastic, and not otherwise.

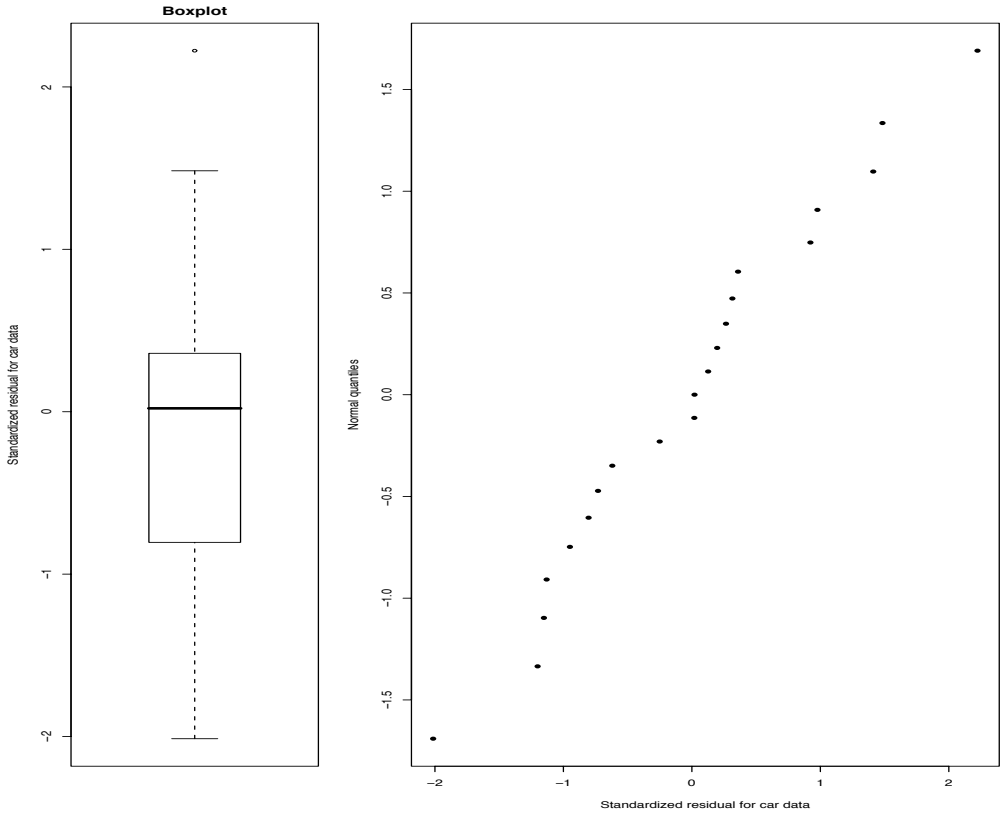
- An oval scatter plot is "well-behaved" in the sense that the data will be homoscedastic and each distribution may be approximately normal.
- However this well behaved oval shape depends on the distribution or histogram of the x values.
- It is sometimes possible to remove heteroscedasticity by a suitable transformation.
- In the following plots we see heteroscedasticity in the original plot and then in the residual plot,





3.10 Prediction and precision

- For the values within a vertical slice, we can imagine both an average and a standard deviation.
- We have seen that our average, or prediction, for y for a vertical slice corresponding to a given x is given by $\hat{y}$ calculated from the regression equation.
- For the standard deviation within a vertical slice, we will assume that it is reasonable to use s_e, the standard deviation of the residuals.
- This means that we use the same value s_e for every vertical slice.
- This is only reasonable if our data is homoscedastic, meaning that as we move up and down the regression line, the spread of points in the vertical direction is about the same.
- This is hard to check, but simplest is just to examine the scatter plot.
- If we can make this assumption of homoscedasticity, we are able to say something about the *precision* of our prediction.
- We can say something like: our predicted value for a given x is $\hat{y}$, and the likely error is s_e.
- If we want to obtain a range of likely values of y for a fixed x, one possibility is to use the mean for that vertical slice, $\hat{y}$, together with the standard deviation s_e for that vertical slice (assuming homoscedasticity).
- We don't know the *shape* of the distribution of values in the slice, but it is common to assume that the shape is Normal.
- Therefore, we assume that for a fixed point x, the distribution of y values is $\sim N(\hat{y}, s_e^2)$ approximately.
- For example, about 68% of all the points are within $\pm s_e$ of the regression line, and about 95% of all points are within $\pm(1.96)s_e$.
- This assumption of Normality may be checked by calculating all the residuals, and assessing them for Normality using a Normal quantile plot.



3.11 Correlation and regression: Tasks and Assumptions

Task	Method	New Assumption
1. Measuring association	Calculate correlation coefficient r	Association is linear
2. Estimating a linear relationship	Calculate regression line $y = a + bx$	Linearity, (causality?)
3. Predicting value at x	$\hat{y} = a + bx$	Sample is representative
4. Estimating accuracy of prediction	Standard deviation of residuals, s_e	Scatterplot is homoscedastic: same standard deviation in each vertical slice
5. Predicting proportions in ranges	Normal distribution with mean $\hat{y}$ and standard deviation s_e	Residuals have Normal distribution

1. Check: Scatter plot.
2. Check: Consider plausibility, possibility of lurking variables, plots of residuals versus other variables.
3. Check: Consider circumstances of data collection.
4. Check: Plot of residuals versus x.
5. Check: Normal quantile plot of residuals.

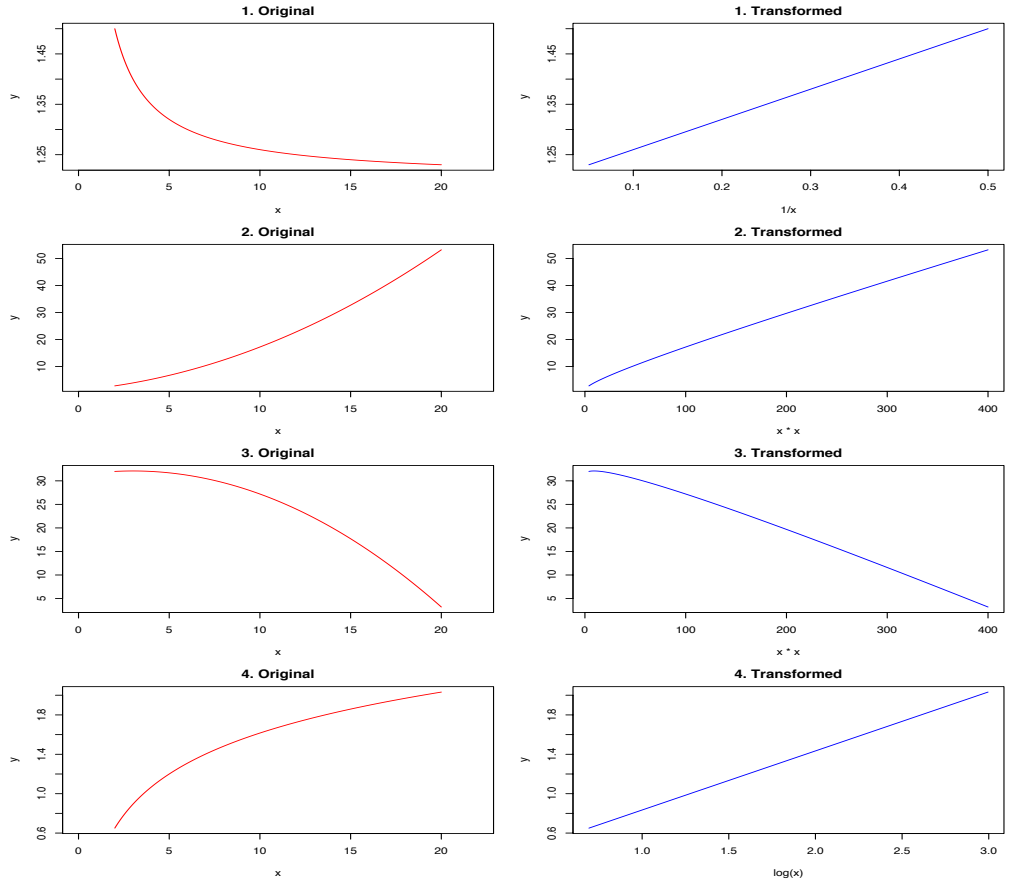
3.12 Transformations for bivariate data

- The main goal when we transform bivariate data is to linearize the relationship between x and y.

- There is rarely a perfect transformation to solve all problems.
- Sometimes the physical context might suggest an appropriate transformation.
- For example, bacteria double in number during a specified time period, so that the total number of bacteria grow exponentially in time.
- Therefore a transformation from x to e^x, or of y to $log(y)$ should straighten the relationship.
- The yield of a crop from a given area of land might be expected to increase quadratically with area, so a plot of y versus x^2 might give a straight relationship.
- Typically we may attempt to transform the y variable or the x variable, but not both together.
- We usually only attempt a transformation when the relationship is monotonic.
- It is usually best to choose a transformation by eye, with the aid of a computer. There are more sophisticated methods available.
- See the graphs below for some suggested transformation types.
- A simple family of transformations is $y = \frac{x^\lambda - 1}{\lambda}$. These are known as Box-Cox transformations.
- λ can be estimated, but is usually chosen from powers e.g. -1,-1/2, 0, 1/2, 2. Note that

$$\lim_{\lambda \to 0}(\frac{x^\lambda - 1}{\lambda}) = \log(x)$$

- It is sometimes possible to remove heteroscedasticity by a suitable transformation, often by replacing y by $log(y)$.
- Unfortunately, while this might fix the heteroscedasticity problem, it tends to de-linearize the relationship if the relationship was linear to start with.



3.13 Power law models

- Consider predicting the calorific content y of pizza from its area. What transformation should spring to mind?
- $y \propto A = \pi r^2$, so we should have in mind that $\sqrt{y} \propto r$.
- Thus, taking square roots of y should produce a linear relationship with pizza radius.
- A log transform also linearizes the relationship: $\ln y \propto \ln \pi + 2 \ln r$.
- A power law model is one of the form $y = a \times x^p$ for some power p.
- These turn out to be fairly common, or at least useful as an approximation in many disciplines.
- For example, in Biology suppose that y is the rate at which animals use energy and x is their body weight.
- If q_0 is the animal's metabolic rate, and M the animal's mass, then Kleiber's law states that $q_0 \propto M^{\frac{3}{4}}$. (For plants the exponent is close to one.)
- This is an approximation, but seems to work for animals from bacteria to whales.

- Alternatively, $\ln q_0 \propto \frac{3}{4} \ln M$ linearizes the relationship, so take $y = \ln q_0$ and $x = \ln M$, and we expect $\hat{b} \simeq \frac{3}{4}$.
- The point to notice is that the slope b in the log-linear relationship is the power p in the power model.
- Therefore, for such models we tend (a) to transform to logs; (b) carry out regression analysis on the log transforms; (c) make predictions on the log scale and back-transform to the original scale.
- The estimated slope in the regression b is an estimate of the power, p.

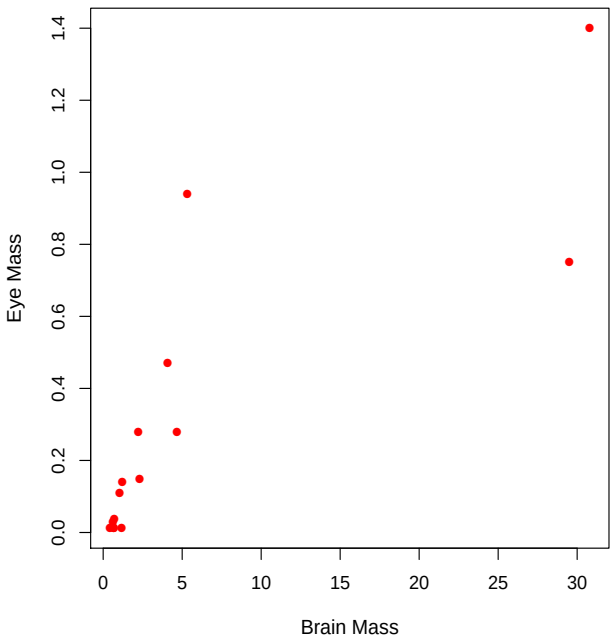
Power Law Models: Exam Question

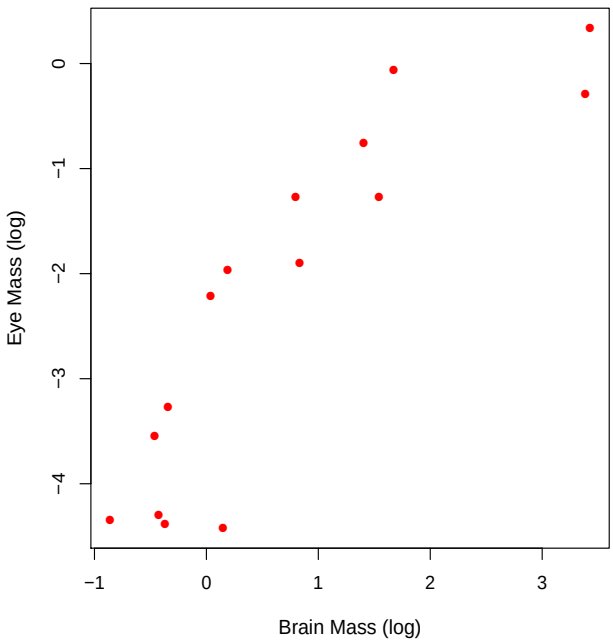
- As an extended example condider this problem, from the 2012 exam.
- The following data concern the masses (in grams) of eyes y and brains x in rodents (R.F. Burton, 2006).
- The principal hypothesis is that eye masses are proportional to brain masses, but possibly affected also by head shapes and visual requirements.
- The relationship is to be modelled using the power law: $y \propto x^p$.

i	Rodent species	Brain mass, x	Eye mass, y
1	Tamias striatus	2.22	0.28
2	Sciurus hudsonicus	4.67	0.28
3	Castor canadensis	29.50	0.75
4	Zapus hudsonius	0.65	0.01
5	Lemmus trimucronatus	1.16	0.01
6	Microtus pennsylvanicus	0.69	0.01
7	Ondatra zibethicus	5.33	0.94
8	Cricetus cricetus	1.04	0.11
9	Meriones unguiculatus	1.21	0.14
10	Mastomys coucha	0.71	0.04
11	Rattus norvegicus	2.30	0.15
12	Peromyscus sp.	0.42	0.01
13	Claviglis saturatus	0.63	0.03
14	Erethizon dorsatum	30.80	1.40
15	Cavia cutleri	4.07	0.47

Printed below is R output containing some summaries, and on the next page are scatter plots of the data on original and log scales.

```
print(mean(rodents$eye))
0.3092067
print(mean(rodents$brain))
5.693733
print(sd(rodents$eye))
0.4146261
print(sd(rodents$brain))
10.05515
print(cor(rodents$brain,rodents$eye))
0.829227
print(mean(log(rodents$eye)))
-2.242703
print(mean(log(rodents$brain)))
0.7317723
print(sd(log(rodents$eye)))
1.695043
print(sd(log(rodents$brain)))
1.338845
print(cor(log(rodents$brain),log(rodents$eye)))
0.8784087
```





Questions:

1. Use the scatter plots to assess the relationship between eye mass and brain mass.
2. What value of p corresponds to the principal hypothesis? Estimate p and comment on your answer.
3. Calculate a prediction of eye mass for a species with a brain mass of $15g$, and calculate a range of values which might be expected to include eye mass for 95% of species whose brain mass is $15g$. Comment on your answer.
4. Specify *in the context of this data* what assumptions you are making in giving your answers, and how you would check them.

Answer to question 1:

- The shape of the relationship depends on how we view the three rodents with eye mass exceeding $0.6g$.
- We might argue for some curvature on the original scale, but the species *Castor canadensis* perhaps then outlies.
- The cluster of points with small values makes any relationship hard to determine.
- Increasing variability in eye mass (heteroscedasticity) with brain mass.
- On a log scale we might argue that the relationship is approximately linear, although some curvature perhaps remains.
- The data appear more homoscedastic on this scale (hence better for regression).

Answer to question 4:

- Computing the linear regression line assumes linearity in the relationship. There is a hidden assumption of causality. (You may not know that brain mass is a proxy for body mass, which is at least pretended to be causal in influencing eye mass.)
- The prediction assumes that the sample is representative of the population of rodents for which we wish to make predictions. Hard to check unless one has expertise.
- Estimating accuracy of prediction depends on an assumption of homoscedasticity. Looking at the scatter plot, this looks a reasonably safe assumption.
- Predicting proportions in ranges requires the assumption that the residuals have Normal distribution. This could be checked by computing the residuals and drawing a Normal quantile plot.

Other Remarks

- For data that results from controlled experiments, prediction may be quite reliable. However, where the data result from an observational study, we must beware confounding, and suspect the causality in our regression.
- *Multiple regression* techniques can be used to predict a y-value from two or more independent values, for example $y = a + bx + cz$ uses two explanatory variables to help predict y.
- *Polynomial regression* can be used to counter the effects of non-linearity, for example $y = a + bx + cx^2$ introduces a quadratic term in x.
- We will deal with *testing* the degree of association, and so forth next term.

4 Multiple Regression

Multiple Regression

- What can we do if there is more than one possible predictor?
- The answer is Multiple Regression using least squares.
- The general set up is that we have a response variable y and p explanatory variables $x_1, \ldots, x_p$.
- We also assume observations on these variables for each of n individuals.
- We will suppose that the observations on the ith individual are

$$(x_{1i}, x_{2i}, \ldots, x_{pi}, y_i)$$

- We have n of these $(p+1)$-tuples,

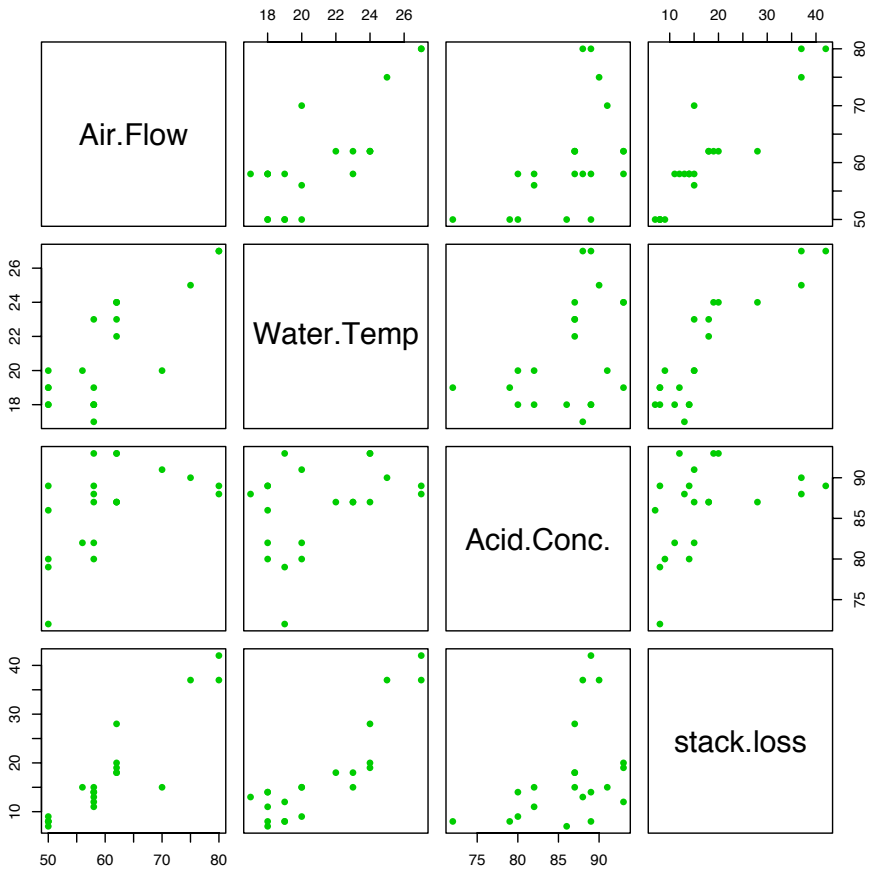
$$(x_{11}, \ldots, x_{p1}, y_1), \ldots, (x_{1n}, \ldots, x_{pn}, y_n)$$

- So for each x the first subscript indicates which variable it is, and the second subscript indicates which individual it is.
- The idea is that we wish to predict y using all the explanatory variables, and the simplest way is to do this linearly using the linear model

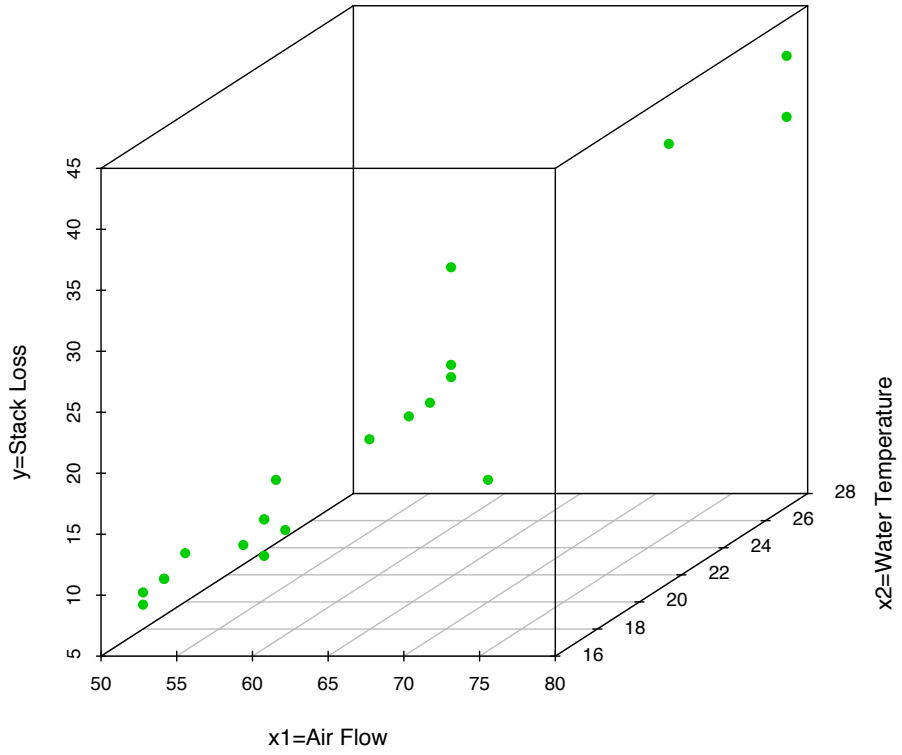
$$y = a + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

- As an example, the following famous data set is from operation of a plant for the oxidation of ammonia to nitric acid, measured on 21 consecutive days.
- Taken from: K. A. Brownlee, *Statistical Theory and Methodology in Science and Engineering*, Wiley (1965), p. 454.
- The response variable is NH_3 loss, known as stack loss y
- There are three predictor variables:
  - Air Flow x_1
  - Water Temperature x_2
  - Acid Concentration x_3

	x_1 Air Flow	x_2 H_2O Temp.	x_3 Acid Conc.	y NH_3 Loss
	80	27	89	42
	80	27	88	37
	75	25	90	37
	62	24	87	28
	62	22	87	18
	62	23	87	18
	62	24	93	19
	62	24	93	20
	58	23	87	15
	58	18	80	14
	58	18	89	14
	58	17	88	13
	58	18	82	11
	58	19	93	12
	50	18	89	8
	50	18	86	7
	50	19	72	8
	50	19	79	8
	50	20	80	9
	56	20	82	15
	70	20	91	15
Mean	60.4	21.1	86.3	17.5
Standard deviation	9.2	3.2	5.4	10.2
Correlation with x_1 Air Flow	1	0.78	0.50	0.92
Correlation with x_2 H_2O Temp	0.78	1	0.39	0.88
Correlation with x_3 Acid Conc	0.50	0.39	1	0.40
Correlation with y NH_3 loss	0.92	0.88	0.40	1



- We begin by interpreting a plot, created by the **R pairs** command, of the data.
- Our main interest is in how the x variables relate to the y variable.
- The bottom row of plots is easiest to interpret as these have the axes correctly orientated.
  - Air flow (x_1) and Stack loss (y) seem weakly positively related. There are repeated values of some Air flow measurements.
  - Water Temperature (x_2) and Stack loss (y) seem positively related, but perhaps the relationship is curved?
  - Acid Concentration (x_3) and Stack loss (y) have a more complicated relationship. Possibly weakly positively correlated, maybe showing some heteroscedasticity. There is a cluster of four observations which seem distant from the other observations.
- We also calculate summary statistics and the *correlation matrix*.
- The correlations are perhaps a bit stronger than I would have expected, looking at the plot.
- It is much harder to visualise relationships in higher dimensions.
- The following plot shows that y tends to rise about linearly with x_1 and x_2 together.



- To predict y using p predictors, we will use the rule

$$\hat{y}_i = a + b_1 x_{1i} + \cdots + b_p x_{pi}$$

so that our problem is to choose $a, b_1, \ldots, b_p$.

- We do this by minimising the sum of squared errors between the observations and their fitted values:

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (a + b_1 x_{1i} + \ldots b_p x_{pi}))^2$$

- We minimise by differentiating with respect to each unknown and equating to zero.

- This gives $p + 1$ simultaneous equations:

$$\frac{dQ}{da} = 0, \frac{dQ}{db_1} = 0, \ldots, \frac{dQ}{db_p} = 0$$

which must be solved. (These are called the *normal* equations, but they don't have anything to do with the Normal distribution!)

- Solving these equations and finding values for $a, b_1, \ldots, b_p$ is outside the scope of this course - the calculations are very elegant if you know some matrix algebra.
- Instead, we will assume that we can use a computer package such as R to solve these equations.
- We can look at the equations that we get when $p = 2$. The simultaneous equations that we need to solve turn out to be

$$\begin{aligned} b_1 s_1 + b_2 s_2 r_{12} &= r_{1y} s_y \\ b_1 s_1 r_{12} + b_2 s_2 &= r_{2y} s_y \\ a &= \bar{y} - b_1 \bar{x_1} - b_2 \bar{x_2} \end{aligned}$$

- Our notation here is that s_1 is the standard deviation for x_1; r_{12} is the correlation between x_1 and x_2, and r_{1y} is the correlation between x_1 and y, and with other notation following similarly.
- Notice that these equations only involve averages, standard deviations, and pairwise correlations: $\bar{x}_1, \bar{x}_2, \bar{y}$; s_1, s_2, s_y; and r_{12}, r_{1y}, r_{2y}.
- Thus, to construct the regression equation we are using basic summaries of the data and excluding the rest.
- This remains the case as we move beyond $p = 2$.
- Return to the Nitric Acid example: First, $p = 3$.
- Secondly the simultaneous equations result in

$$\begin{aligned} \hat{a} &= -39.9 \\ \text{Air flow } \hat{b}_1 &= 0.72 \\ \text{Water temp } \hat{b}_2 &= 1.30 \\ \text{Acid concentration } \hat{b}_3 &= -0.15 \end{aligned}$$

- The implication is that stack loss increases as Air Flow increases and as the water gets hotter, but falls slightly for high levels of acidity.
- This is a bit of a surprise, because Acid concentration is positively associated with stack loss, as we saw in the scatter plot and correlation matrix.
- This effect is due to the three predictor variables also being correlated: multiple regression provides an analysis to correctly assess their effects.

4.1 Residuals and assumptions

- In general, the residuals are

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - (\hat{a} + \hat{b}_1 x_{1i} + \ldots \hat{b}_p x_{pi})$$

  and s_e denotes their standard deviation.

- There is no easy short-cut formula for calculating s_e, so we use computer output.

- A basic principle of (multiple) regression is that, subject to certain assumptions, the distribution of y for particular values of $x_1, \ldots, x_p$ is Normal with mean $\hat{y} = \hat{a} + \hat{b}_1 x_1 + \ldots \hat{b}_p x_p$ and standard deviation s_e.

- That is given $x_1, \ldots, x_p$, we have $y \sim N(\hat{y}, s_e^2)$.

- This is the same assumption as for simple linear regression, but our concept of a vertical slice is now replaced by the idea of a slice through higher dimensions.

- Assumptions:

  1. Underlying relationship between y and $x_1, \ldots, x_p$ is linear
  2. Sample is representative of future cases
  3. The amount of variation in y for particular $x_1, \ldots, x_p$ does not depend on the values of $x_1, \ldots, x_p$.
  4. The shape of the distribution of y corresponding to particular $x_1, \ldots, x_p$ is Normal.

4.2 R^2 and the calculation of s_e

- To calculate s_e using R output, one method is to find the value printed as *R-squared, the multiple correlation coefficient*, also known as *multiple R-squared*.

- R^2 is the "proportion of variation in y explained by $x_1, \ldots, x_p$", or equivalently one minus the amount of residual variation, expressed as a proportion of initial variation.

- That is,

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} = 1 - \frac{\frac{1}{n-1}\sum_{i=1}^n \hat{\epsilon}_i^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

- Note that $R^2 = r^2$ for simple linear regression.

- R^2 is most often expressed as a percentage.

- We may thus compute:

$$s_e = s_y \sqrt{1 - R^2}$$

4.3 Obtaining information from R output

- The following two tables show R output.
- Our main need is the values of the coefficients $\hat{b}_i$. We also need the value of R^2.
- The values of means and standard deviations are also needed, but are not shown in this output.
- The first table shows R output for the stackloss regression, using all three predictor variables.
- To obtain coefficients from a multiple regression, see the *Coefficients:* part.
- The other main calculation which we require is the value of R^2, which appears towards the foot of the output described as *Multiple R-squared*.
- The second table shows R output for the stackloss regression, using only the first two predictor variables.

```
### ---------------------------------------------------------------------------
### The output for model1:
### print(model1)
Call:
lm(formula = Stackloss ~ Air.Flow + Water.Temp + Acid.Conc, data = stackloss)

Coefficients:
(Intercept)     Air.Flow   Water.Temp    Acid.Conc
   -39.9197       0.7156       1.2953      -0.1521

### print(summary(model1))
Call:
lm(formula = Stackloss ~ Air.Flow + Water.Temp + Acid.Conc, data = stackloss)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2377 -1.7117 -0.4551  2.3614  5.6978

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.9197    11.8960  -3.356  0.00375 **
Air.Flow      0.7156     0.1349   5.307  5.8e-05 ***
Water.Temp    1.2953     0.3680   3.520  0.00263 **
Acid.Conc    -0.1521     0.1563  -0.973  0.34405
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-squared: 0.9136,     Adjusted R-squared: 0.8983
F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09
```

Table 8: R output for the stackloss regression, using all three predictor variables. To obtain coefficients from a multiple regression, see the *Coefficients:* part. The other main calculation which we require is the value of R^2, which appears towards the foot of the output described as *Multiple R-squared*.

```
### ---------------------------------------------------------------------------
### The output for model2:
Call:
lm(formula = Stackloss ~ Air.Flow + Water.Temp, data = stackloss)

Coefficients:
(Intercept)     Air.Flow   Water.Temp
   -50.3588       0.6712       1.2954

> print(summary(model2))

Call:
lm(formula = Stackloss ~ Air.Flow + Water.Temp, data = stackloss)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5290 -1.7505  0.1894  2.1156  5.6588

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -50.3588     5.1383  -9.801 1.22e-08 ***
Air.Flow      0.6712     0.1267   5.298 4.90e-05 ***
Water.Temp    1.2954     0.3675   3.525  0.00242 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.239 on 18 degrees of freedom
Multiple R-squared: 0.9088,     Adjusted R-squared: 0.8986
F-statistic: 89.64 on 2 and 18 DF,  p-value: 4.382e-10
```

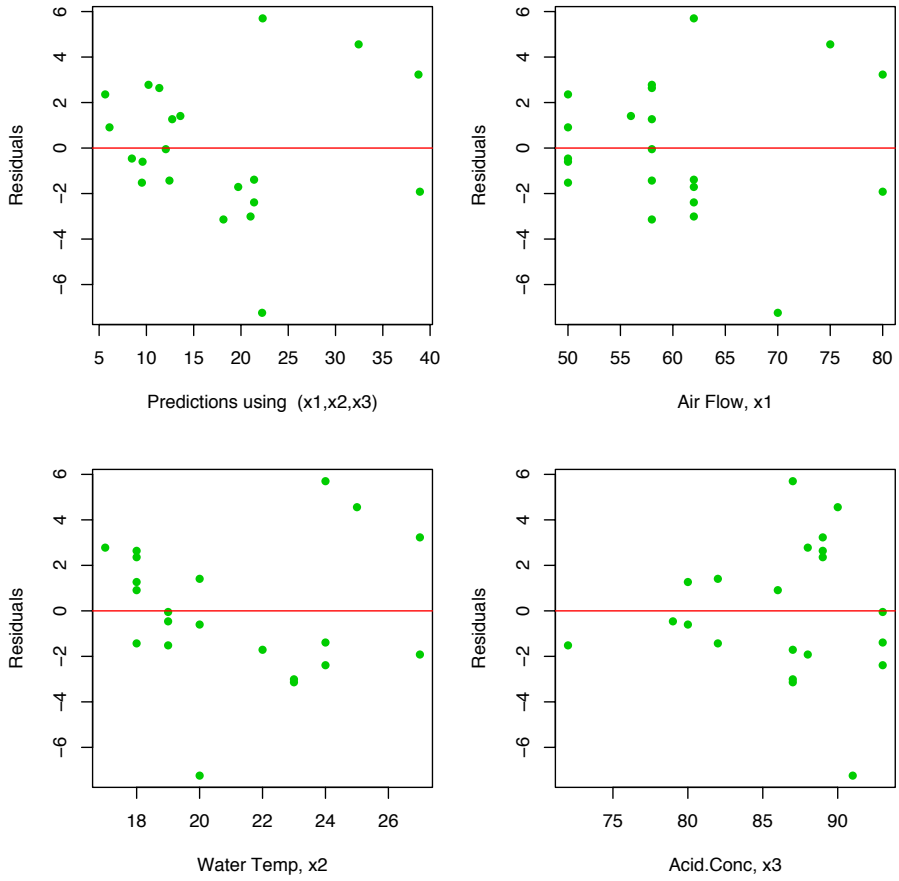
Table 9: R output for the stackloss regression, using the first two predictor variables. To obtain coefficients from a multiple regression, see the *Coefficients:* part. The other main calculation which we require is the value of R^2, which appears towards the foot of the output described as *Multiple R-squared*.

4.4 Diagnostics for predictions & variable selections

- For Multiple Regression we first have to consider which assumptions we need to satisfy:

Assumption	**Validation method**
Sample is representative for future cases	Thought/insight
Underlying relationship between y and $x_1, \ldots, x_p$ is linear	Check no curvature in residual plots
Amount of variation in y does not depend on values of $x_1, \ldots, x_p$	Check no heteroscedasticity in residual plots
Distribution of y for fixed $x_1, \ldots, x_p$ is Normal	Quantile plot of residuals

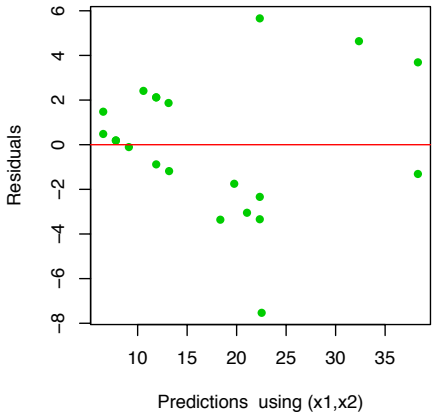
- The residual plots to draw are:
  - the residuals versus the fitted values, $\hat{\epsilon}_i$ versus $\hat{y}_i$;
  - the residuals versus each predictor variable, $\hat{\epsilon}_i$ versus x_{ji} for each x_j separately.
- We look for the same behaviour as for regression with a single predictor.
- Look at the plots for stack loss data.
- Some problems with Acid Concentration?

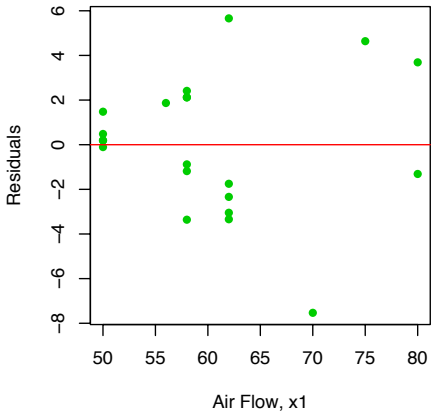


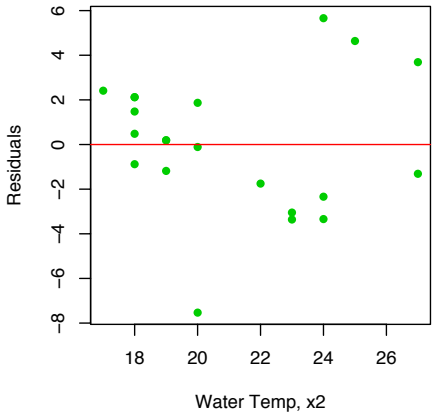
- Provided all residual plots are homoscedastic and residuals are approximately Normal, prediction calculations can be performed.
- These use the regression equation prediction as the mean for y and the standard deviation s_e of the residuals as the standard deviation of y.
- That is, given a value for x, we have that $y \sim N(\hat{y}, s_e^2)$.
- We can't make reliable predictions for the original stack loss regression, but we can do if we drop Acid Concentration from the regression.
- The new regression equation is

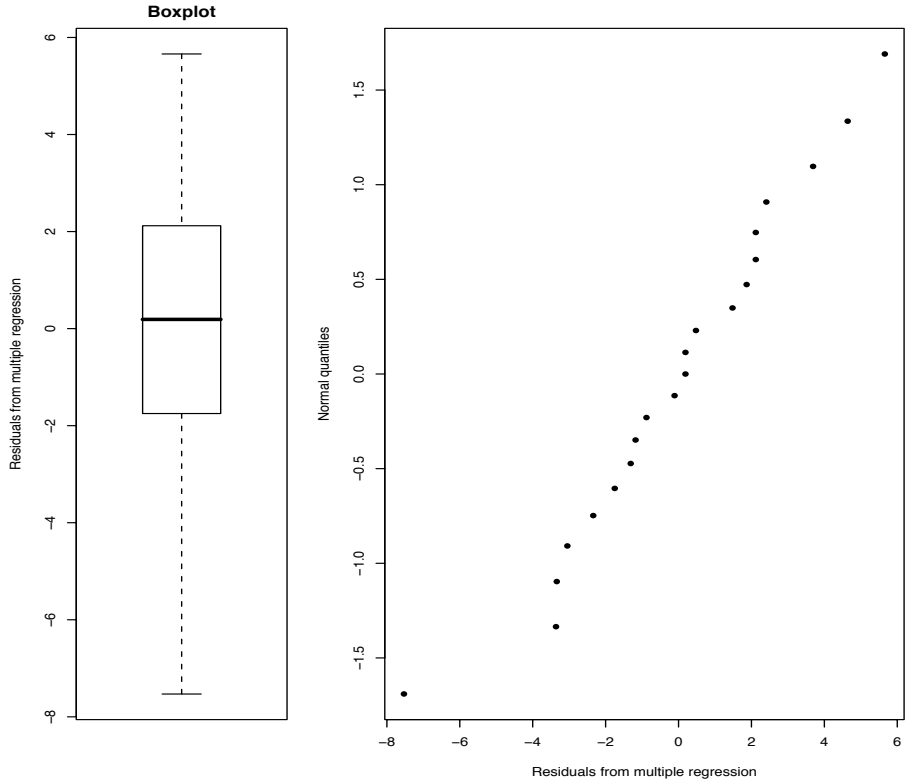
$$\hat{y} = -50.4 + 0.67 \times \text{Air Flow} + 1.30 \times \text{Water Temp}.$$

- We interpret this much as we did before.

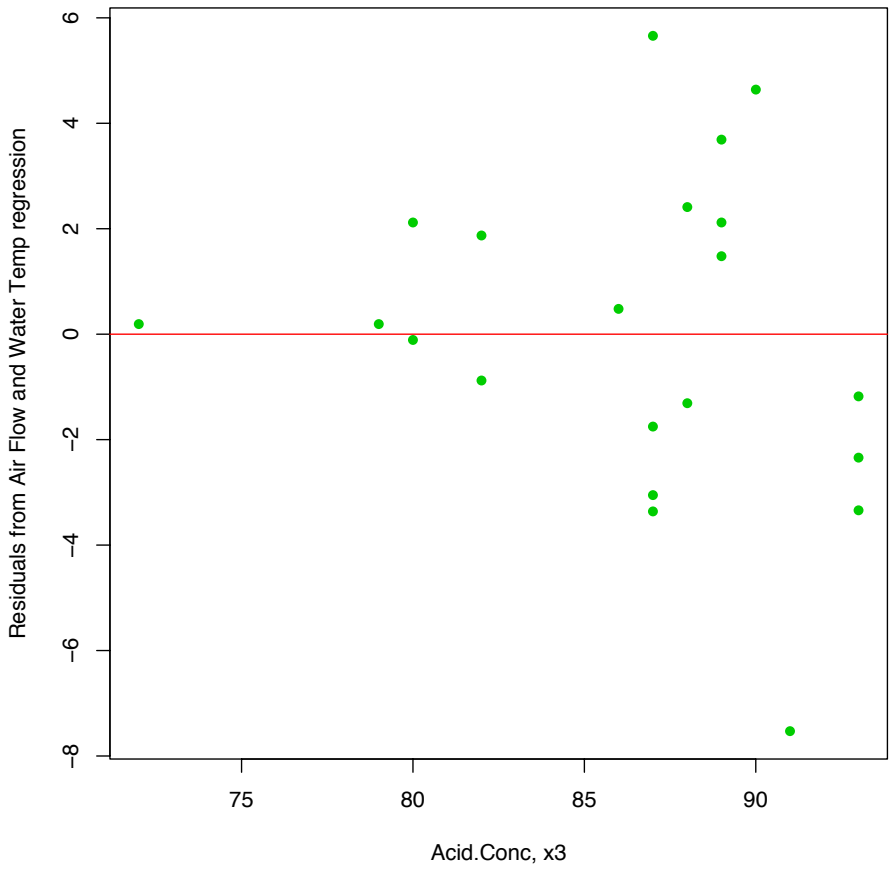

Residuals
6
4
2
0
−2
−4
−6
−8
10
15
20
25
30
35
Predictions using (x1,x2)


Residuals
6
4
2
0
−2
−4
−6
−8
50
55
60
65
70
75
80
Air Flow, x1


Residuals
6
4
2
0
−2
−4
−6
−8
18
20
22
24
26
Water Temp, x2



- The residual plots now look OK, so we now check Normality of residuals by checking a quantile plot of the residuals.
- There is one possible outlier, otherwise the assumption of Normality looks quite safe.
- We will suppose that we can go ahead and make predictions.
- Note that extrapolation outside range of known predictors is dangerous AND this data came from an uncontrolled observational study; so DON'T trust this prediction much!
- Have we lost anything by excluding x_3, acid concentration?
- A plot of residuals versus a variable not in the regression will show a linear pattern if that variable needs to be included. For this example, it seems not.



4.5 The Value of Variables

The Standardised Regression Coefficient

- Clearly some variables might be more relevant than others. How do we judge this?
- One thing is to look at the size of the (suitably scaled) coefficient for x_i in the regression equation.
- So we might consider $|\hat{b}_i|s_i$ (the "standardised regression coefficient") as a measure of how much this contribution to $\hat{y}$ varies.
- So therefore:

$$\text{Value of variable } x_i \quad \propto \quad |\hat{b}_i|s_i$$

  where s_i is the standard deviation for variable x_i.

- This arises because if we transform x_i to standard units by defining

$$x_i^* = \frac{x_i - \bar{x}_i}{s_i}$$

then the term $b_i x_i$ in the regression equation becomes

$$b_i s_i x_i^* + b_i \bar{x}_i$$

- The latter portion doesn't involve x_i^* (it gets absorbed into the intercept term) and so the actual coefficient of the standardized value x_i^* is $b_i s_i$.
- For the stack loss data, and the first regression we fitted, we have $s_1 = 9.2$ and $\hat{b}_1 = 0.72$, so that $|\hat{b}_i| s_i = 6.6$ for Air Flow.
- We have also 4.1 for Water Temp and -0.82 for Acid Conc.
- This suggests that Air flow is the most important variable in the equation, and Acid concentration the least important.

Multiple R^2

- R^2 can be used as a measure of the overall "quality" of the multiple regression model.
- Interpreting R^2 takes a little care: an R^2 value of 91% is not much better than one of 90%.
- However, an R^2 of 99% is much better than 98%.
- Why? We have that $s_e = \sqrt{1 - R^2} s_y$, and therefore:

R^2	s_e
90%	$0.31 s_y$
91%	$0.30 s_y$
98%	$0.14 s_y$
99%	$0.10 s_y$

- So the relative change in size of prediction errors is large for the second situation and small for the first.

The Change in R^2

- Another way of judging the value of a single variable x_i is to look at how R^2 changes as individual predictors are left out.
- By leaving some out, we get a different set of simultaneous equations in a smaller number of unknowns and so the coefficients of other predictors generally change.
- R^2 always decreases if we leave out a variables.
- If it R^2 doesn't decrease much, we might consider that the variable wasn't very important for prediction.
- (There are highly sophisticated methods for such variable selection.)
- For the stack loss data, we can see in the following table how model value changes with the variables selected.

Predictors	R^2	Prediction formula
AF, WT, AC	91.4%	$\hat{y} = -39.9 + 0.72\text{AF} + 1.30\text{WT} - 0.15\text{AC}$
AF, WT	90.9%	$\hat{y} = -50.4 + 0.67\text{AF} + 1.30\text{WT}$
AF	84.6%	$\hat{y} = -44.1 + 1.02\text{AF}$
WT	76.7%	$\hat{y} = -41.9 + 2.82\text{WT}$
None (just $\hat{a}$)	0.0%	$\hat{y} = \bar{y} = 17.5$

- One fundamental difficulty is that of *co-linearity* of predictor variables.
- The same information may be carried by more than one predictor.
- This will tend to be the case when the predictor variables are highly correlated.
- For example, if $x_3 = x_1 + x_2$ then we don't need x_3.
- Or, suppose we are trying to predict risk of heart disease from height, weight, and waist measurement - we would expect these to be highly correlated.
- In a very extreme situation, we may have a regression where the following happens:

Predictors	R^2
All variables	92%
All except x_1	90%
All except x_2	90%
All except x_1 and x_2	1%

- The problem here is that x_1 and x_2 contain essentially the same information about y and are very strongly correlated with each other.
- In such cases, nothing in the data alone can tell us which prediction rule we should use.
- Sometimes insight into the the particular context may suggest that one set of predictors is better than another.
- The problem stems from the correlations between the predictors.
- If the correlations between them were zero, then each predictor would explain a certain percentage of the variation in y and it wouldn't overlap with the variation explained by another variable.
- Were we to be asked to **design** the experiment in advance, and hence choose the values of x_{ij}, then this provides a very efficient strategy.

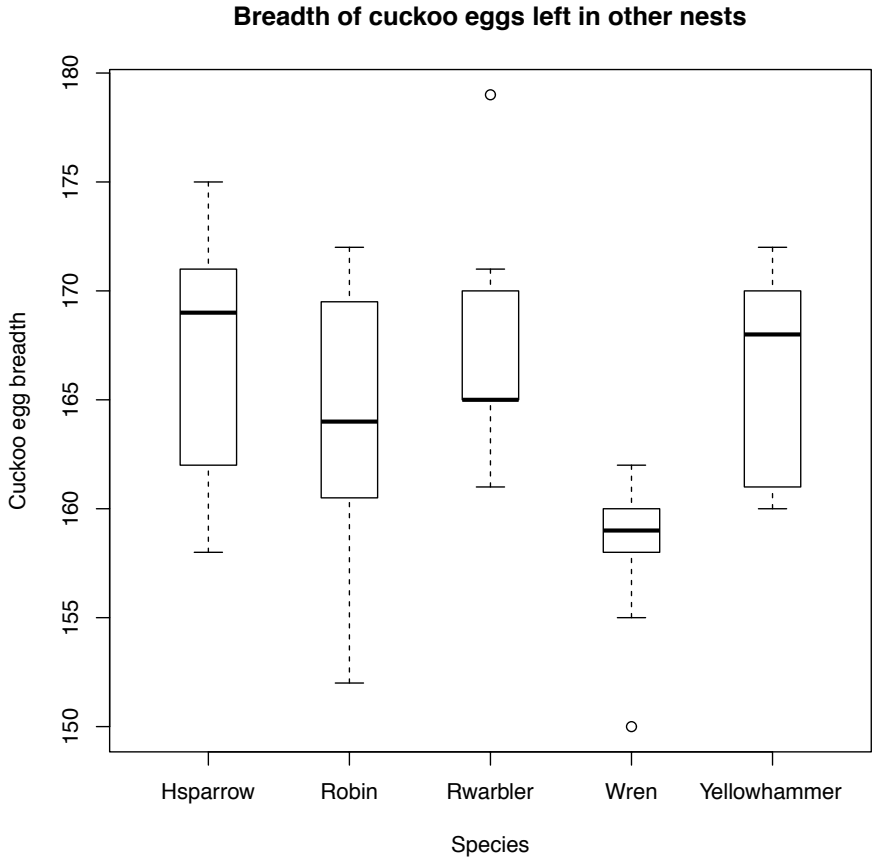
5 Exploring differences between many groups

5.1 The one-way layout

- The term one-way layout is used to describe the situation where we study the dependence of one continuous variable (called the response variable) on another discrete variable (called a factor).
- The different values ("levels") of the latter divide the values of the response into groups.
- The next two panels show data from a one-way layout.
- The data are the breadths of cuckoos' eggs found in nests of a number of other species.
- In this example, the response variable is the breadth of the egg and the factor is the host species in whose nest each egg was found.
- The goal of the analysis is find a simple representation/description of the variation found in the response variable.
- In particular, we wish to distinguish variation associated with the factor and unexplained variation.
- Example: One-way layout. These are breadths of cuckoos eggs found in other birds' nests:

Host species	Breadths of eggs
Hedge sparrow	170, 169, 158, 173, 175, 175, 162, 165, 162, 171, 161, 169, 167, 170.
Reed warbler	169, 171, 170, 161, 165, 165, 161, 165, 179, 165.
Robin	160, 159, 171, 166, 169, 161, 172, 162, 169, 152, 163, 170, 160, 164, 164, 170.
Wren	150, 160, 162, 159, 162, 155, 160, 159, 155, 159, 160, 157, 159, 160, 160.
Yellowhammer	160, 160, 161, 172, 165, 170, 170, 170, 168.

- The boxplot is a natural way to plot the groups of data. We should spend some effort on interpreting the boxplots for features such as:
  - whether the groups are centred at about the same location
  - whether the groups have about the same spread (homogeneity);
  - symmetry of groups and whether Normality looks plausible



- For example, cuckoo eggs left in wrens' nests seem clearly smaller than cuckoo eggs left in wrens' nests.
- This implies something remarkable: do cuckoos know how large their eggs will be when laid?
- Can cuckoos distinguish nests for different birds?
- Are some cuckoos evolved only to lay within wrens' nests?
- So how can we represent this situation mathematically?

5.2 Notation

- p is the number of levels of the factor (number of groups into which the factor divides the observations).
- n_i is the number of observations in group i, where $i = 1, 2, \ldots, p$.
- The values of the response variable are y_{ij}
  - The subscript i denotes the group in which the observation lies and ranges from 1 to p.
  - The subscript j denotes the observation within the group. For each value of i, the subscript j ranges from 1 to n_i.

- For the cuckoo example we have 5 different host species. So $p = 5$.
- There 14 eggs from hedge sparrows' nests, 10 from reed warblers, etc. So $n_1 = 14$, $n_2 = 10$, $n_3 = 16$, $n_4 = 15$ and $n_5 = 9$.
- From the table of the data we can now see that:
  - $y_{1,1} = 170$, $y_{1,2} = 169$, $y_{1,3} = 158$, ..., $y_{1,14} = 170$.
  - $y_{2,1} = 169$, $y_{2,2} = 171$, ..., $y_{2,10} = 165$.
  - $y_{3,1} = 160$, ..., $y_{3,16} = 170$.
  - $y_{4,1} = 150$, ..., $y_{4,15} = 160$.
  - $y_{5,1} = 160$, ..., $y_{5,9} = 168$.
- Note: for clarity, I have temporarily written a comma between i and j
- For each group of observations, we can compute the mean which we denote by $\bar{y}_i$, where $i = 1, \ldots, p$. Thus

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

- We call $\bar{y}_i$ the "group mean" for group i.
- Similarly, we use s_i to denote the standard deviation of the observations in group i. Thus

$$s_i = \sqrt{\frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}$$

5.3 Decomposition

- We denote the mean of the all the values of the response variable by $\bar{y}$ (called the "overall mean"). Thus

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n_i} y_{ij} = \frac{1}{n} \sum_{i=1}^{p} n_i \bar{y}_i \qquad \text{where } n = n_1 + n_2 + \cdots + n_p$$

- We define $g_i = \bar{y}_i - \bar{y}$ and we call it the "group effect" for group i.
- We define $e_{ij} = y_{ij} - \bar{y}_i$ and we call it the "residual" for observation y_{ij}
- Using the notation above, we have

$$y_{ij} = \bar{y} + g_i + e_{ij}$$

- Using the notation above, we have

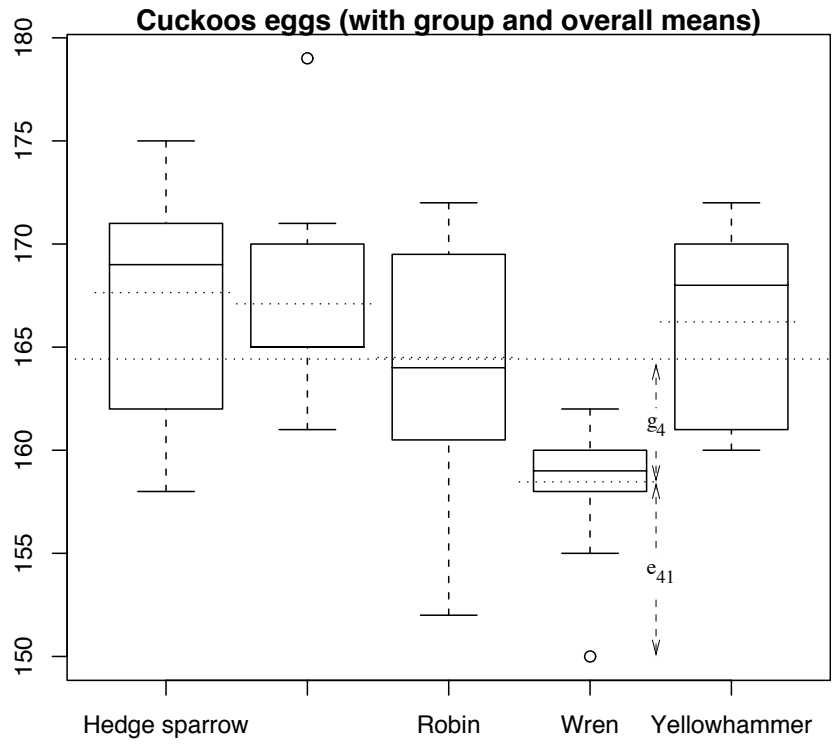
$$y_{ij} = \bar{y} + g_i + e_{ij}$$

- We have decomposed (taken apart) each observation into three pieces:
  - an overall mean: typical value of response variable.
  - a group effect: the difference between mean of group and overall mean.
  - residual (left over): variation due to other factors.
- In each group, the residuals have the same shape and scale as the original observations but have mean zero.
- The group effects will average to zero if the data are *balanced*, meaning that the sample size in each group is the same.
- Otherwise the data are said to be unbalanced and the group effects don't necessarily sum to zero.
- Table of group summaries and effects:

Host species	i	n_i	$\bar{y}_i$	s_i	g_i
Hedge sparrow	1	14	167.64	5.34	3.22
Reed warbler	2	10	167.10	5.38	2.68
Robin	3	16	164.50	5.48	0.08
Wren	4	15	158.47	3.11	-5.95
Yellowhammer	5	9	166.22	4.82	1.80
All species	–	64	164.42	5.89	–

- Table of residuals:

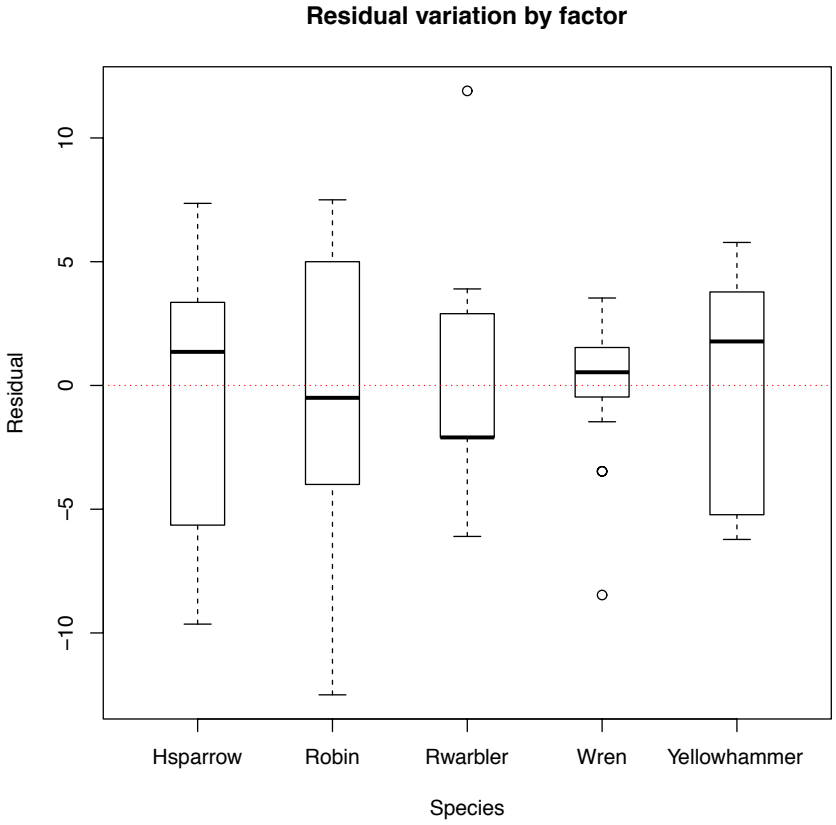
Host species	Residuals
Hedge sparrow	2.36, 1.36, −9.64, 5.36, 7.36, 7.36, −5.64, −2.64, −5.64, 3.36, −6.64, 1.36, −0.64, 2.36
Reed warbler	1.9, 3.9, 2.9, −6.1, −2.1, −2.1, −6.1, −2.1, 11.9, −2.1
Robin	−4.5, −5.5, 6.5, 1.5, 4.5, −3.5, 7.5, −2.5, 4.5, −12.5, −1.5, 5.5, −4.5, −0.5, −0.5, 5.5
Wren	−8.47, 1.53, 3.53, 0.53, 3.53, −3.47, 1.53, 0.53, −3.47, 0.53, 1.53, −1.47, 0.53, 1.53, 1.53
Yellowhammer	−6.22, −6.22, −5.22, 5.78, −1.22, 3.78, 3.78, 3.78, 1.78



- Decomposition of data: short dashed lines across boxes are group means, long dashed line is overall mean.

5.4 Homogeneity

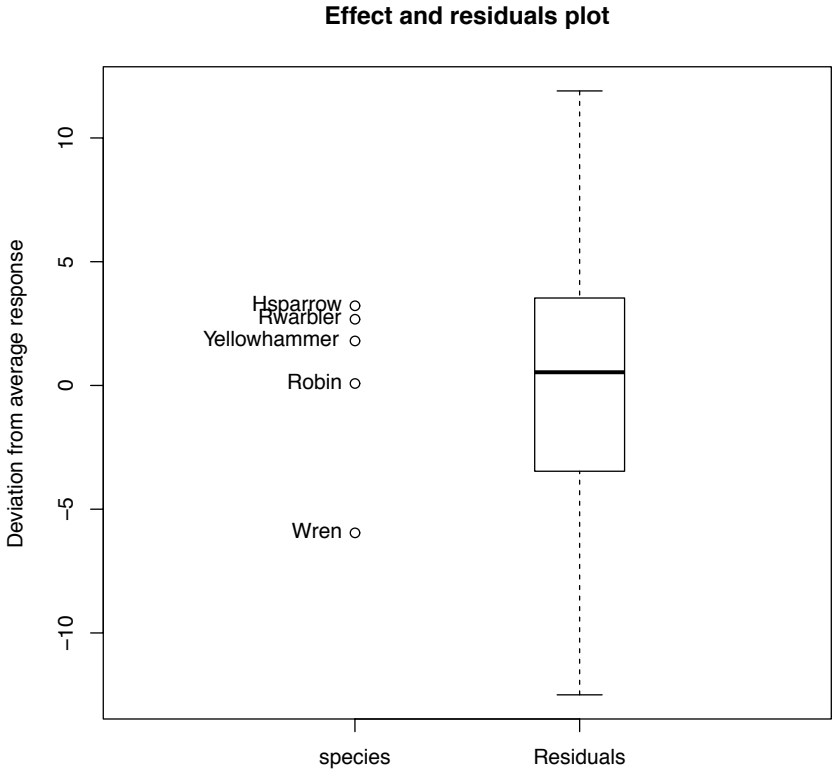
- Corresponding to each group of observations is a population distribution.
- Each group's residuals inform us about the scale and shape of the corresponding population distribution.
- If the p population distributions all have the same scale and shape, the situation is called "homogeneous".
- If homogeneous, we can combine all the residuals together to learn about the shared scale and shape.
- Homogeneity helps us to reduce the amount of information we need to present about a set of data.
- We can assess homogeneity by drawing a boxplot of the residuals for separate groups.



- For the cuckoo example, the centres of the residual for each group are the same, zero, as the mean is zero for each group of residuals.
- The medians might be slightly off centre.
- The spreads are reasonably similar, depending on how we view the outliers for the wren group.

Effects and residuals plot

- If we can assume homogeneity, differences between groups, together with relative amounts of variation, may be assessed graphically using an *effects and residuals* plot.
  - The first column on the plot shows the individual group effects.
  - The vertical position of each group is given by the value of g_i computed earlier.
  - The second column shows the residuals (all groups combined) as a boxplot.
- The graphic shows the main difference between the groups and highlights the relative sizes of variation due to differences between groups ("group effects") and other ("residual") variation.



- We see two features for the example data.
  - There is a large difference between the Wren and other host species.
  - The scale of the differences between groups is somewhat smaller than the scale of the residuals.

5.5 Analysis of variance: ANOVA

- ANOVA equation (not very hard to prove)

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i} g_i^2 + \sum_{i=1}^{p}\sum_{j=1}^{n_i} e_{ij}^2$$

  or

$$\text{“Total SS”} = \text{“Group SS”} + \text{“Residual SS”}$$

  where “SS” is an abbreviation for “sum of squares”.

- The left-side of the equation (the “Total sum of squares”) measures the total amount of variation in the response variable.

- In the ANOVA equation, the first sum on the right-side (the “Group sum of squares”) measures variation explained by the factor which divides the data into groups.

- The second sum on the right-side (the "Residual sum of squares") measures other (residual) variation.
- The relative sizes of these two sums of squares tell us about the relative sizes of variation explained by the grouping factor and of other variation.

The ANOVA Table

- The ANOVA table is easily computed from a table of the group sizes, summaries and effects since

$$\text{"Group SS"} = \sum_{i=1}^{p}\sum_{j=1}^{n_i} g_i^2 = \sum_{i=1}^{p} n_i g_i^2$$

and

$$\text{"Residual SS"} = \sum_{i=1}^{p}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^{p} (n_i - 1)s_i^2$$

- An even easier method for the residual SS is to start from the total sum of squares which is $(n-1)s^2$, where

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2}$$

is the overall standard deviation of all the observations, and subtract the group sum of squares.

- Then we have

$$\text{"Total SS"} = \text{"Group SS"} + \text{"Residual SS"}$$
$$(n-1)s^2 = \sum_{i=1}^{p} n_i g_i^2 \qquad + \sum_{i=1}^{p} (n_i - 1)s_i^2$$

- It is common to present the results of ANOVA in tabular form, such as follows:

	Sum of squares	Calculate as:	Variation component
Host species	778.28	$\sum_{i=1}^{p} n_i g_i^2$	35.7%
Residuals	1403.40	$\sum_{i=1}^{p} (n_i - 1)s_i^2$	64.3%
Total	2181.61	$(n-1)s^2$	100.0%

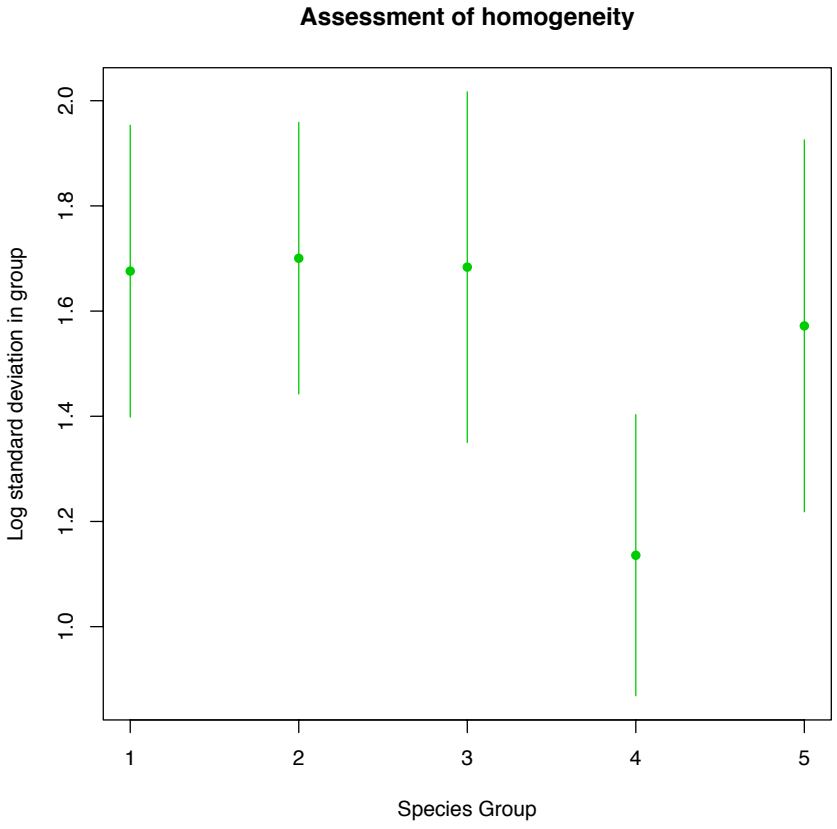
- We can express these SS as percentages of the original variation.
- One gives the percentage of variation attributable to the factor, and the other gives the percentage of variation which can't be explained by the factor.
- The "Host species" sum of squares is about 35% of the total.
- We say that 35% of the variation in the breadth of cuckoos eggs is explained by differences between host species and about 65% is left over (unexplained).

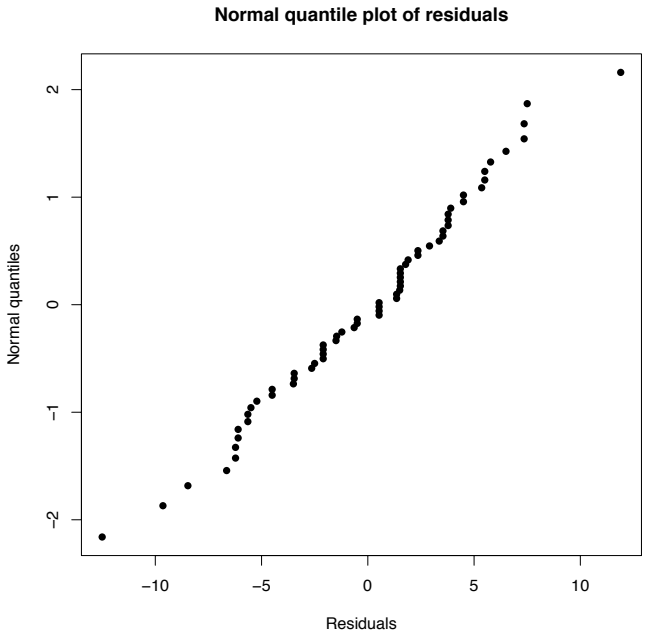
5.6 Assessing homogeneity

- When homogeneous, each of the p populations has the same standard deviation.
- The sample group standard deviations $s_1, \ldots, s_p$ will vary but should not vary too much unless the group sizes are very small.
- For a standard deviation obtained from a simple random sample of size n, the variation of the natural logarithm of s between samples is of order $\pm\frac{1}{\sqrt{n-1}}$
- A simple check on the assumption of homogeneity is to plot each $\log(s_i)$ with $\pm\frac{1}{\sqrt{n_i-1}}$ error bars and informally assess whether the underlying population standard deviations might be the same for each group.

Assessing homogeneity: Cuckoo Example

- The natural logarithms of the standard deviations of the groups are $1.68, 1.68, 1.70, 1.14$ and 1.57
- The corresponding accuracies $1/\sqrt{n_i - 1}$ are $0.28, 0.33, 0.26, 0.27, 0.35$.
- From the corresponding plot, all the bars overlap at about 1.4.
- This might suggest that the group population sds could all be near $\exp(1.4) = 4$.
- This might be taken to imply homogeneity.





- Assuming that we have homogeneous groups, we may need to assess the Normality of the residuals, and this is done using a Normal quantile plot for the combined residuals.
- For these data, the residual distribution looks pretty Normal.

5.7 Transformations

Transformations: The Location Scale Plot

- When the response is always positive, homogeneity often fails because there is a strong relationship between group mean and group standard deviation.
- One way of handling the situation is to apply a suitable transformation to the response variable so that the transformed variable has homogeneous distributions.
- To detect if a transformation is likely to succeed, we can plot the group standard deviations s_i versus the group means $\bar{y}_i$ on logarithmic scale.
- This is called a *location-scale plot*.
- Plot $log(s_i)$ on the y-axis and $log(\bar{y}_i)$ on the x-axis. We examine the slope for the relationship.
  - If the slope is one of 1/2, 1 or 2 then it is likely that an appropriate transformation will help matters.
  - A slope of 0 suggests leave well alone.
  - Slope 1/2 suggests taking the square-root of each y-value.
  - Slope 1 suggests taking the logarithm.
  - Slope 2 suggests taking the reciprocal.

- The reason for this is as follows.
- Suppose group i has population mean μ_i and sd σ_i. If we see observations that generally have larger residuals for larger group means, this is equivalent to saying that

$$\sigma_i \approx \theta \mu_i^{\alpha}$$

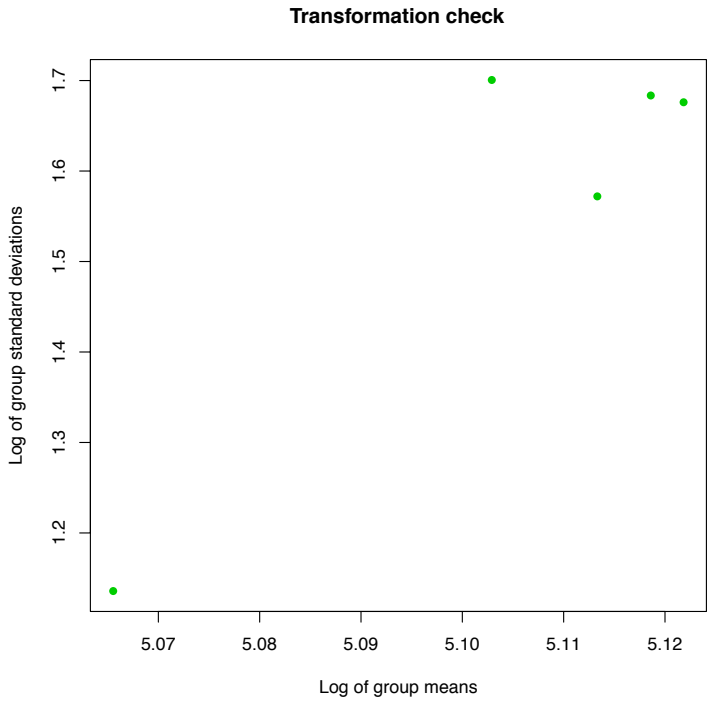
  for some constants θ (of no interest) and α.
- Our interest is in cases $\alpha > 0$ as this will imply heteroscedasticity. Taking logs, we then get

$$\log \sigma_i \approx \log \theta + \alpha \log \mu_i$$

- We can estimate σ_i by s_i and μ_i by $\bar{y}_i$ to give

$$\log s_i \approx \log \theta + \alpha \log \bar{y}_i$$

- This is a regression equation, and α is the slope of the relationship.
- Clearly $\alpha \approx 0$ implies no relationship of variation with size.



- For the cuckoo data, logarithm of group standard deviation versus logarithm of group mean. No obvious slope, and one outlier (wrens).
- For a second example, consider the data in the following table.
- This shows tensile strength of cement, y, and curing time, x.

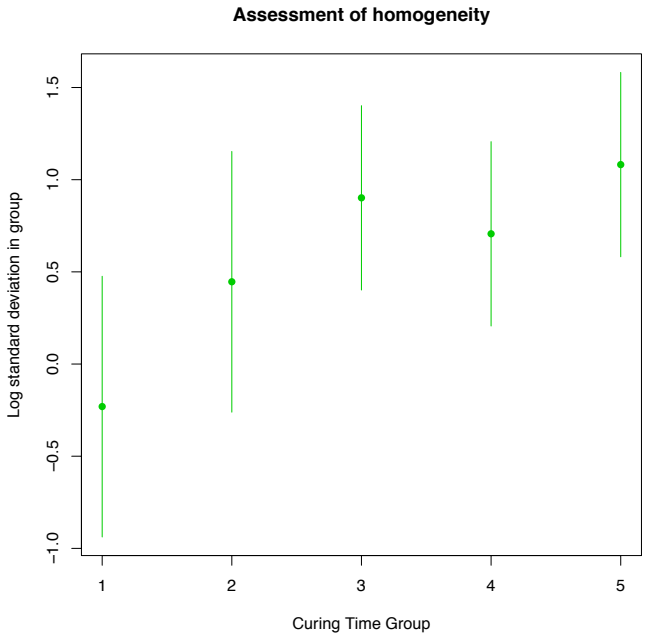


Figure 12: Cement Example: Assessment of homogeneity, not so homogeneous.

- There is some suggestion that variability increases as the mean increases.

	Curing.Time	Tensile.Strength	logstrength
	y	x	$\log y$
1	1	13.00	2.56
2	1	13.30	2.59
3	1	11.80	2.47
4	2	21.90	3.09
5	2	24.50	3.20
6	2	24.70	3.21
7	3	29.80	3.39
8	3	28.00	3.33
9	3	24.10	3.18
10	3	24.20	3.19
11	3	26.20	3.27
12	7	32.40	3.48
13	7	30.40	3.41
14	7	34.50	3.54
15	7	33.10	3.50
16	7	35.70	3.58
17	28	41.80	3.73
18	28	42.60	3.75
19	28	40.30	3.70
20	28	35.70	3.58
21	28	37.30	3.62

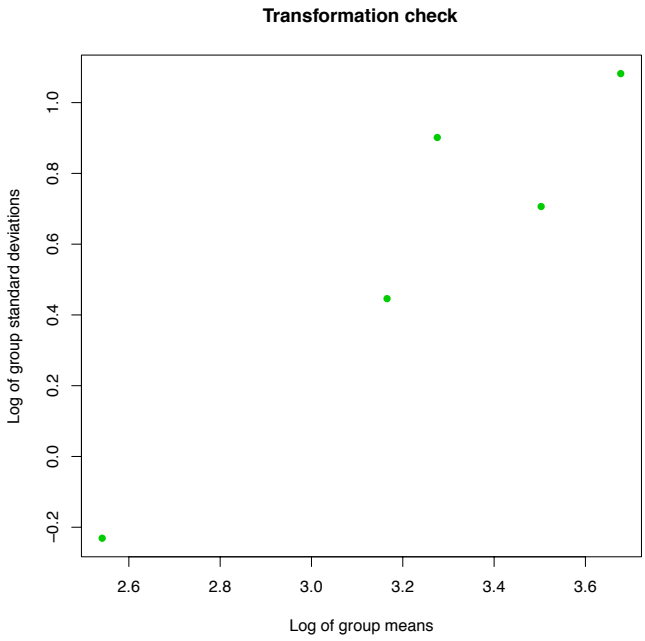


Figure 13: Location-scale plot for cement strength data: Slope roughly 1, which suggests taking logarithm of the response variable (strength).

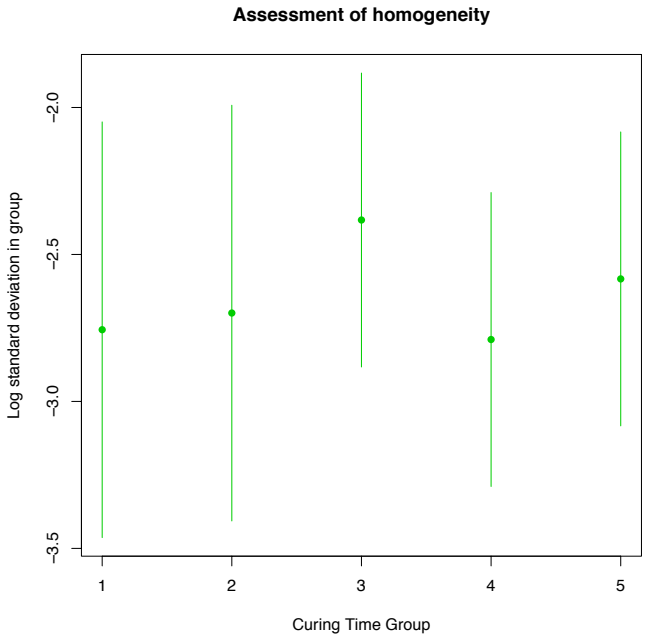


Figure 14: Logarithm of cement strength versus curing time: More homogeneous?

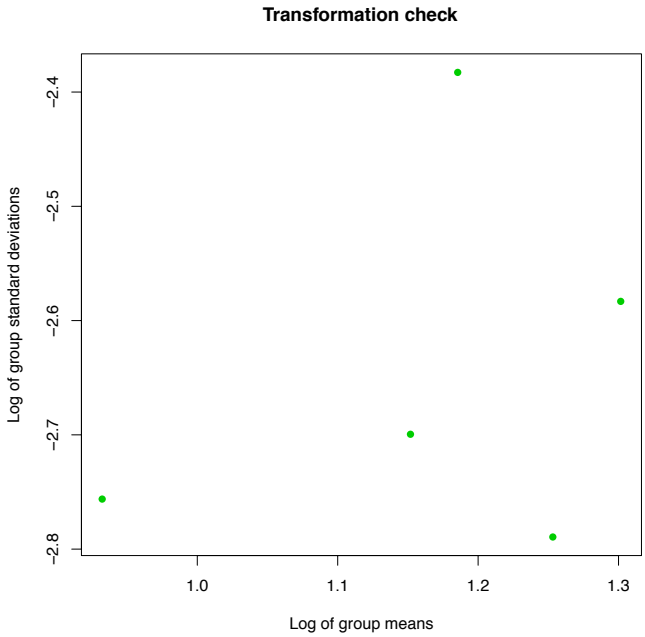


Figure 15: Location scale plot of log of cement strength versus curing time.

6 Exploring differences between two factors

6.1 Two-way layout: Introduction

- Now we have two factors instead of just one.
- As before, the different values of the factors divide the values of the response into groups.
- Let p denote the number of levels of the first factor and q the number of levels of the second factor.
- Then there may be as many as pq groups of the response.
- We shall only consider a special case, where all pq groups occur and where there is the same number of observations m in each group.
- We call this a *complete factorial design with balanced replication.*
- The example is based on data collected by Nurcombe studying maternal adaptation (how well a mother's behaviour is adapted to the needs of her child).
- Nurcombe studied a number of mother-baby pairs to try to understand the factors related to and influencing adaptation.
- We will examine a subset of the data in which two factors vary in a complete factorial design with balanced replication.
- The response is maternal adaptation (lower numbers indicate better adaptation).

- The factors:
  - Factor one ("treatment") divides mother/baby pairs into three groups: (i) low birth-weight and given an experimental treatment intended to enhance adaptation; (ii) low birth-weight given no treatment; (iii) normal birth-weight.
  - Factor two ("education") divides mother-baby pairs into two groups: (i) high-school education only; (ii) post high-school education.
- The data are tabulated and plotted below.
- Less revealing than with just one factor but still fairly clear pattern.
- Adaptation seems better for more educated mothers, and scores seem higher for LBW-C, than for NBW then for LBW-E.

	Education	
	High-school	Post high-school
Low Birth Weight — Experimental	14, 20, 22, 13 13, 18, 13, 14	11, 11, 16, 12 12, 13, 17, 13
Low Birth Weight — Control	25, 19, 21, 20 20, 14, 25, 18	18, 16, 13, 21 17, 10, 16, 21
Normal Birth Weight	18, 14, 18, 20 12, 14, 17, 17	16, 20, 12, 14 18, 20, 12, 13

6.2 Notation

- We have 3 different "treatments". So $p = 3$.
- We have 2 different amounts of "education". So $q = 2$.
- For each treatment/education combination, we have 8 observations. So $m = 8$.
- The values of the response variable are y_{ijk}
  - The subscript i denotes the level of the first factor and ranges from 1 to p.
  - The subscript j denotes the level of the second factor and ranges from 1 to q.
  - The subscript k denotes the observation within the group determined by the levels of the factors and ranges from 1 to m.

6.3 Decomposition into mean and residual

- Group means: For each group of observations defined by a combination of levels of the two factors, we have a mean

$$\bar{y}_{ij} = \frac{1}{m}\sum_{k=1}^{m} y_{ijk}$$

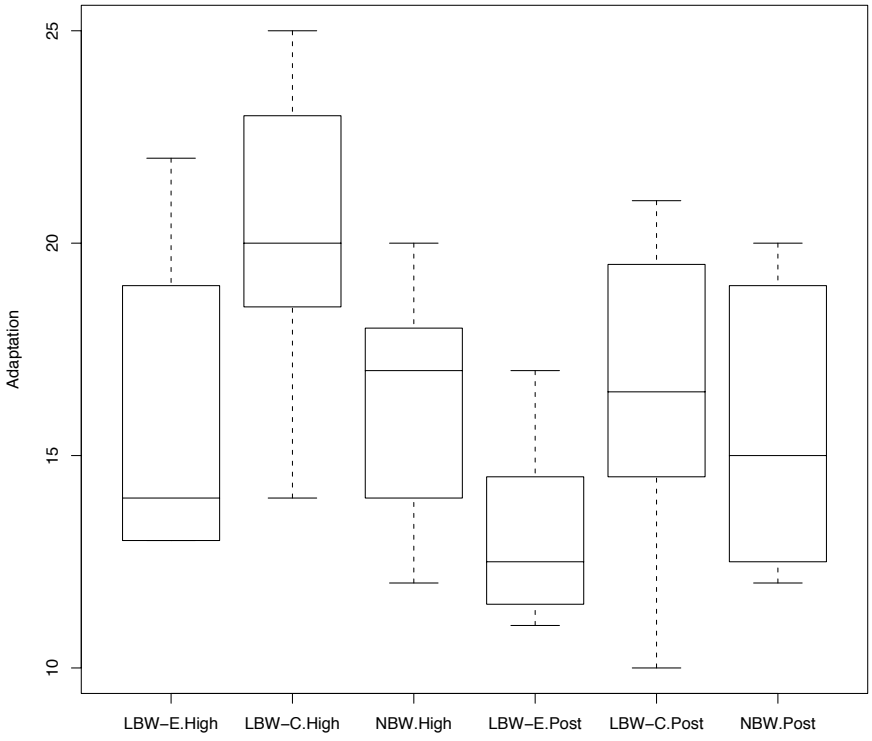


Figure 16: Boxplots of all groups in Nurcombe data set.

- Residuals: For each observation y_{ijk}, we have a residual

$$e_{ijk} = y_{ijk} - \bar{y}_{ij}$$

- Decomposition: Each observation is made up of two pieces: group mean and residual.

$$y_{ijk} = \bar{y}_{ij} + e_{ijk}$$

The group means and the residuals for the Nurcombe data are:

	High	Post high
LBW-Exp	15.875	13.125
LBW-Control	20.250	16.500
NBW	16.250	15.625

Table of residuals (note these sum to zero in each block).

	High	Post high
LBW	-1.875 4.125 6.125-2.875	-2.125-2.125 2.875-1.125
Exp	-2.875 2.125-2.875-1.875	-1.125-0.125 3.875-0.125
LBW	4.750-1.250 0.750-0.250	1.500-0.500-3.500 4.500
Con.	-0.250-6.250 4.750-2.250	0.500-6.500-0.500 4.500
NBW	1.750-2.250 1.750 3.750	0.375 4.375-3.625-1.625
	-4.250-2.250 0.750 0.750	2.375 4.375-3.625-2.625

6.4 Interaction plots

- One way of visualizing relationships between the factors is via an *interaction plot.*
- This shows how the mean of the response depends on the levels of the two factors.
- One factor appears on the horizontal axis and the other is used to draw lines with different patterns.
- Here we see that the differences between treatments are fairly clear and are much the same for the two levels of education.

6.5 Decomposition into row and column effects

- Row means: We write $\bar{y}_{i.}$ for the mean of all observations where the first factor is at level i.
- Column means: $\bar{y}_{.j}$ is the mean of all observations where the second factor is at level j.
- Overall mean: $\bar{y}$ denotes the mean of all the response values.
- Therefore

$$\bar{y}_{i.} = \frac{1}{q}\sum_{j=1}^{q} \bar{y}_{ij} \qquad \bar{y}_{.j} = \frac{1}{p}\sum_{i=1}^{p} \bar{y}_{ij} \qquad \bar{y} = \frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q} \bar{y}_{ij}$$

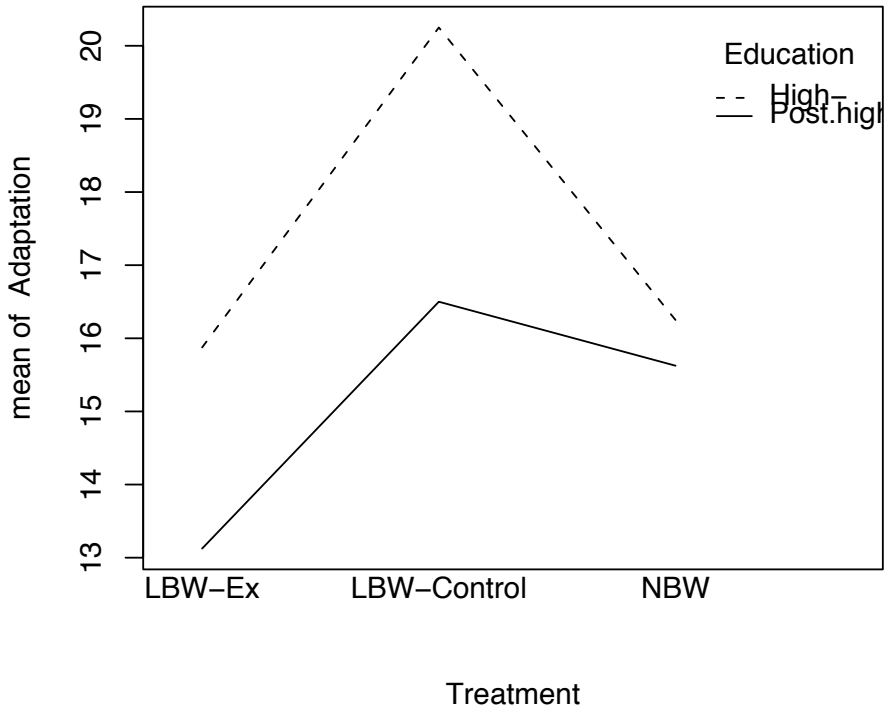


Figure 17: Interaction plot of Nurcombe data set.

- Row effects: we call $r_i = \bar{y}_{i\cdot} - \bar{y}$ the "effect" for level i of the first factor.
- Column effects: we call $c_j = \bar{y}_{\cdot j} - \bar{y}$ the "effect" for level j of the second factor.
- Interactions: We call $w_{ij} = (\bar{y}_{ij} - \bar{y}) - (r_i + c_j)$ the interaction between level i of the first factor and level j of the second factor.
- Decomposition: each group mean is made up of four pieces:

$$\bar{y}_{ij} = \bar{y} + r_i + c_j + w_{ij}$$

- Thus, each observation has the decomposition:

$$y_{ijk} = \bar{y} + r_i + c_j + w_{ij} + e_{ijk}$$

6.6 Mean polish

- We can compute the two-way layout decomposition by *mean polish*.
- Begin with the table of group means. First, polish the columns of the group means. This means we find the column average and remove it from each value in the column:
  - take each column from the original table
  - find its mean

  - write down the mean at the bottom
  - subtract the mean from each number in the column

- So we begin with:

	High	Post high
LBW-Exp	15.875	13.125
LBW-Control	20.250	16.500
NBW	16.250	15.625

  and after the column averages have been swept out we get

	High	Post high
LBW-Exp	-1.5833	-1.9583
LBW-Con	2.7916	1.4166
NBW	-1.2083	0.5416
	17.4583	15.0833

- Next we polish the rows of what we are left with. This means we find the row average of what is left, and remove it from each value in that row:

  - Take each row from the table (including the column average row)
  - find its mean
  - write down the mean at the right
  - subtract the mean from each number in the row

- This will give the final table,

	High-	Post high	
LBW-Exp	0.1875	-0.1875	-1.7708
LBW-Control	0.6875	-0.6875	2.1042
NBW	-0.875	0.875	-0.3333
	1.1875	-1.1875	16.2708

  - The six numbers in a block are the interactions between treatment and education, i.e. the values w_{ij}.
  - The single number at bottom-right is the overall mean, $\bar{y}$.
  - The three numbers at the right are the treatment effects, r_i.
  - The two numbers at the bottom are the education effects, c_j.

- Our first observation is $y_{1,1,1} = 14$. We have decomposed this as

$$\begin{aligned} 14 &= 16.2708 - 1.7708 + 1.1875 + 0.1875 - 1.8750 \\ y_{111} &= \bar{y} \quad + r_1 \quad + c_1 \quad + w_{11} \quad + e_{111} \end{aligned}$$

  So this observation is:

- in a row (experimental LBW) with lower-than-average treatment scores r_1
- in a column (high-school education) with higher than average education scores c_1
- a block where the interaction between LBW-E and education leads to a higher score w_{11}
- has a negative residual e_{111} meaning that this observation was lower than average for the other observations in its group.

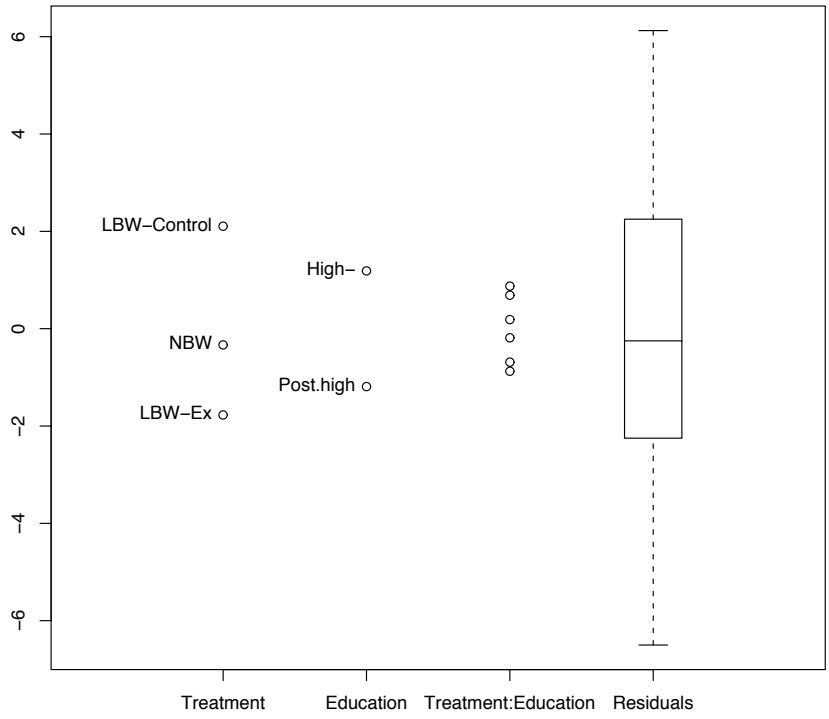
6.7 Interactions

- If the interactions w_{ij} are all zero, things are simple:
  - The only way in which the level of the first factor affects the group mean is via the row effect which is the same for all levels of the second factor. The same statement is true reversing the roles of the two factors.
  - Hence we can think (and report) separately about the dependence of the response on the two factors.
- If the interactions are small relative to the sizes of the row and column effects we get most of the same benefit as if they were actually zero.
- If the interactions are the same size or larger than the row and/or column effects, the situation is complicated.
- The "effect" of changing factor one depends on the level of factor two. We cannot think about the factors separately. They interact. In an extreme case, we could have:
  - When factor one is at level 1, the group mean *increases* greatly when we go from level 1 of the second factor to level 2.
  - When factor one is at level 2, the group mean *decreases* greatly when we go from level 1 of the second factor to level 2.

6.8 Homogeneity

- Homogeneity now means that each of the pq groups should have the same population scale and shape.
- As in the one-way case, if homogeneous, we can combine all the residuals together to learn about the shared scale and shape.
- We can examine homogeneity of scale using group standard deviation s_{ij} in much the same way as the one-way case.
- The next panel shows a two way layout Effects and Residuals Plot for our example:
  - The first column shows the row effects r_i.
  - The second column shows the column effects c_j.
  - The third column shows the interactions.
  - The fourth column shows the residuals.
- It highlights the individual factor effects and indicates relative size of four kinds of variation: differences between levels of factor one, differences between levels of factor two, interactions between the factors, and other ("residual") variation.

Effects and Residuals Plot



- We obtain an effects and residuals plot as for the one-way layout, but with separate parts of the graph showing (1) effects for the first factor, (2) effects for the second factor; (3) interaction effects; (4) residuals.

- We can clearly see here

  – important treatment and education effects
  – the interaction effects are not tiny
  – the residual variation is much larger than everything else.

6.9 Analysis of variance — ANOVA

- ANOVA equation:

$$\sum_{i,j,k}(y_{ijk} - \bar{y})^2 = \sum_{i,j,k} r_i^2 + \sum_{i,j,k} c_j^2 + \sum_{i,j,k} w_{ij}^2 + \sum_{i,j,k} e_{ijk}^2$$

or

"Total SS" = "Factor One SS" + "Factor Two SS"
+ "Interaction SS" + "Residual SS"

- The relative sizes of the four sums of squares tell us about the relative sizes of the four corresponding kinds of variation.

- Most of the ANOVA table computes from the mean polish results:

$$\text{“Factor One SS”} = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m} r_i^2 = qm\sum_{i=1}^{p} r_i^2$$

$$\text{“Factor Two SS”} = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m} c_j^2 = pm\sum_{j=1}^{q} c_j^2$$

$$\text{“Interaction SS”} = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m} w_{ij}^2 = m\sum_{i=1}^{p}\sum_{j=1}^{q} w_{ij}^2$$

- The residual sum of squares can be computed directly from the residuals by squaring and summing or from group standard deviations by:

$$\text{“Residual SS”} = (m-1)\sum_{i=1}^{p}\sum_{j=1}^{q} s_{ij}^2$$

- Here, s_{ij} is the standard deviation for the group corresponding to level i for factor 1 and level j of factor 2.

- The standard deviations for the 6 Nurcombe groups are as follows.

	High-	Post high
LBW-Exp	$s_{11} = 3.6$	$s_{12} = 2.2$
LBW-Con	$s_{21} = 3.6$	$s_{22} = 3.7$
NBW	$s_{31} = 2.6$	$s_{32} = 3.4$

- The overall sd of all $n = 48$ observations is $s = 3.734$.

- The Total SS can be calculated as

$$\text{“Total SS”} = (n-1)s^2$$

- This givea another way of computing the residual SS: calculate total SS and subtract the two factor sums of squares and the interaction sum of squares.

- For the nurcombe data, Compute sums of squares as follows.
  - Treatment SS:
$$qm\sum r_i^2 = (2)(8)[(-1.77)^2 + 2.10^2 + (-0.33)^2] = 122.79$$
  - Education SS:
$$pm\sum c_j^2 = (3)(8)[(1.19^2 + (-1.19)^2] = 67.69$$

- Interaction SS:

$$m\sum w_{ij}^2 = (8)[0.1875^2 + (-0.1875)^2 + \cdots + (-0.875)^2 + 0.875^2] = 20.38$$

- Residual SS:

$$\sum e_{ijk}^2 = (-1.875)^2 + 4.125^2 + \cdots + (-2.625)^2) = 444.62$$

- or Residual SS:

$$(m-1)\sum s_{ij}^2 = (7)[3.6^2 + 2.2^2 + \cdots + 3.4^2] = 444.62$$

- Total SS:

$$(n-1)s^2 = (47)(3.734^2) = 655.48$$

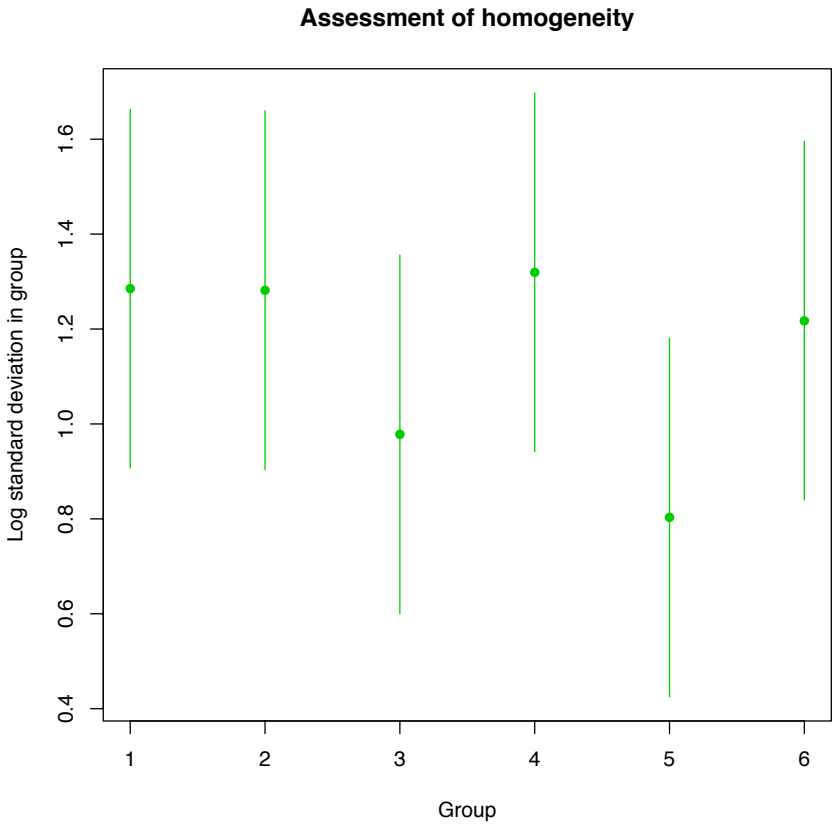
- ANOVA in tabular form:

	Sum of squares	Proportion
Treatment	122.79	18.7%
Education	67.69	10.3%
Treatment:Education (interaction)	20.38	3.1%
Residuals	444.62	67.8%
Total	655.48	100%

- The ANOVA table confirms that residual variation dominates.
- More variation associated with treatment differences than education differences
- Not very large interactions.

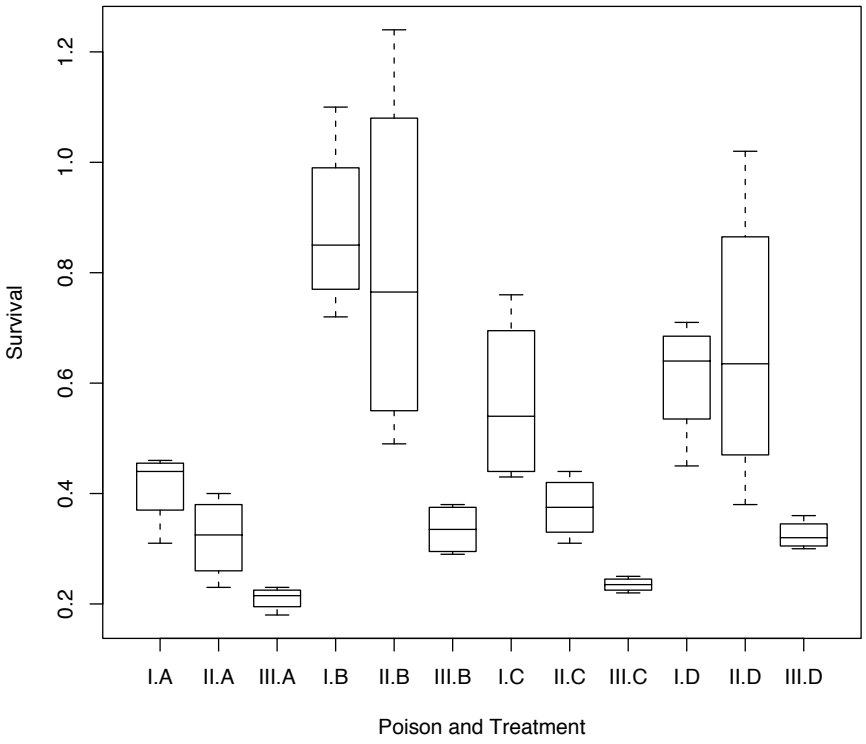
6.10 Testing homogeneity

- Testing homogeneity:
  - The original box-plots all have fairly similar scale and shape (considering only 8 observations in each one).
  - More formally, the natural logarithms of the six group standard deviations are 1.28, 0.80, 1.29, 1.32, 0.97, and 1.21.
  - Since the order of accuracy of each is $1/\sqrt{m-1} = 1/\sqrt{7} = 0.38$, we can there is no reason to reject the possibility of a common population standard deviation.
- A plot is given below.

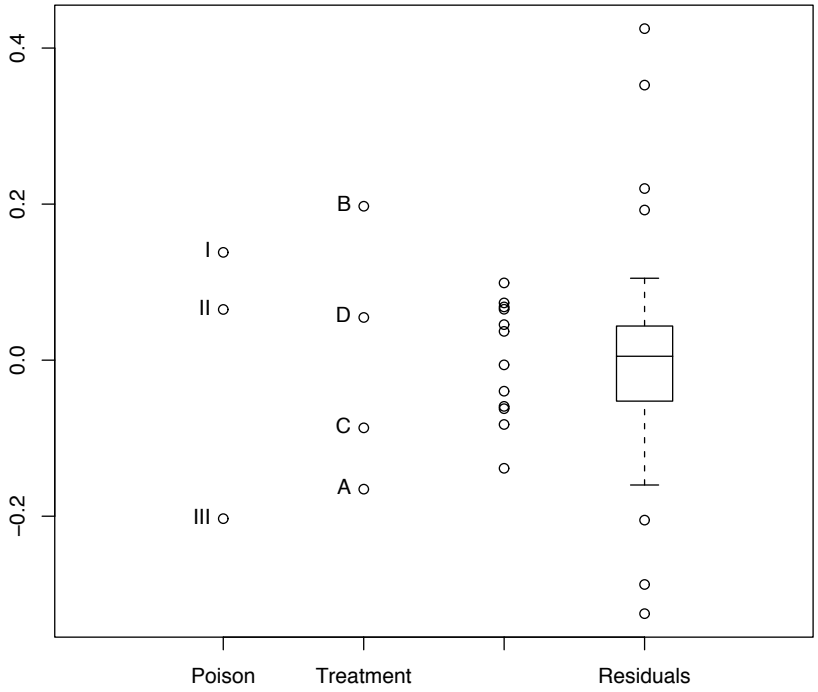


6.11 Transformations

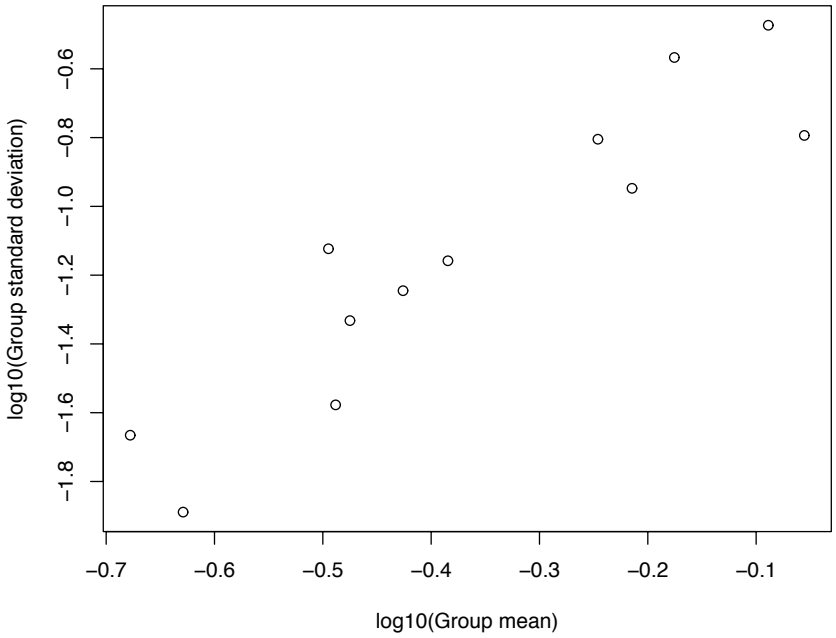
- As in the one-way case, one way of dealing with failure of homogeneity is to apply a suitable transformation to the response variable.
- We produce the same plot (group standard deviations versus group means) and interpret in the same way.
- Animal data example: gives survival times in response to a poison factor (three levels I, II and III) and a treatment factor (four levels A, B, C, D).
- Notice that estimation of effects after a transformation can change the overall conclusion.
- Without transformation, it seems as though the **Treatment** factor has a wider range (more important effects) that the **Poison** factor.
- The location-scale plot shows a slope of about 2, so a reciprocal transformation is indicated.
- After taking transformations, the **Poison** factor has a wider range.



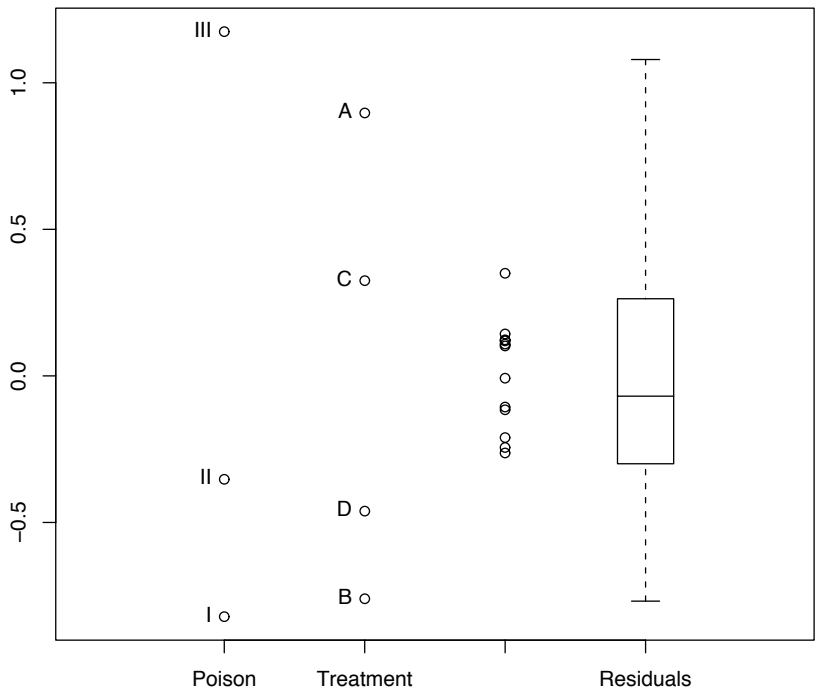
Box plots of original Animal data.



Effects and residuals plot (response is survival).



Group means and standard deviations.



Effects and residuals plot of transformed response (using inverse or reciprocal transformation)

6.12 Median polish (not examinable)

- The decomposition of data using group means and mean polish is very sensitive to outliers.
- Outliers can be unusual individual data values or unusual individual interactions.
- A robust (resistant) alternative is to decompose using group medians and median polish.
- It boils down to using the median where we would have used the mean before.
- Decompose each observation as median of group and residual. Do this by "polishing" each group using the median.
- Thus we would now have table of group medians.
- "Polish" each column using the median to obtain an intermediate table with an extra row.
- "Polish" each row of the intermediate table using the median to obtain the final decomposition.
- Potential difficulty:
  - With mean polish, the mean of each row and column of the interactions is zero as are the means of the row and column effects.
  - With median polish, the corresponding medians may not yet be zero. In this case, it is necessary to repeat the process of polishing the interactions, adding the bits swept-out to the effects.