A short introduction to the R statistics package

Contents

These notes are a short guide to R (not RStudio) as supplied by the Statistical Services Centre. There is some overlap with our tutorial above, some different ways of doing things, and some extra functions you might find useful. It would help for you to be familiar with these, especially if you are taking an SSC course in the future.

Introduction to R 3.0 for Windows

Or install RStudio from the Software Center on the desktop of an institute computer.

What is R?

R is a free statistical package that provides a wide range of basic and advanced data analysis capabilities. You can download it from the webpage www.r-project.org.

R is easily extended via user contributed *packages*, so if you do not find what you need in its basic installation, you are likely to find it in 3rd party contributed libraries.

R has no menu item on the standard toolbar for **Statistics** or **Graphics**, so all its features are best accessed by using commands. It can be daunting to learn the R language because it is object oriented [see the online manual: **Help→Manuals (in PDF)→R language definition**], but once you have understood the concepts, it is similar to using commands in other statistical packages.

1 Introduction

(a) It is much easier to use a computer package once you have used it! Hence this practical deliberately introduces you to the R package before you know much about it.

(b) If you finish this practical or later exercises early, you may like to explore the potential of the package a little further by trying some of the other "commands".

(c) Be adventurous. If you are not sure whether something is correct, try it out.

(d) Most R commands are typed into, and run from the R **console** window. They cannot be abbreviated and they are case sensitive. The default output from a command can be modified by the use of options. Some general commands for saving a workspace, changing directory or loading extra packages are also available in its limited pull-down menu.

2 Starting Off

Load R by **double-clicking** on the R icon. In R you view a file or work on it in a window.

The R **console** window opens automatically on starting; in this window we can type and execute commands, whose output is also sent to this window.

R does not open many windows simultaneously; there's the console and one **graphics** window, this is to minimise required memory, hence R works very well on old computers.

R also has a simple **Data editor** window in which you can type your data. Dataset are just one more type [or class] of object stored, just like a list of output or an executable function.

3 A Simple Example

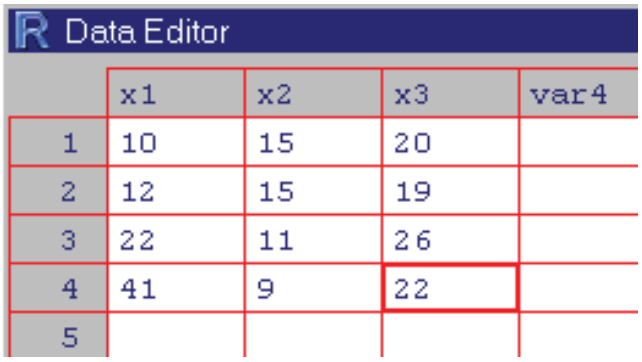
Let us enter a small data set and carry out some simple analyses. There are three columns of data to enter, with a variable name for each column.

Open the data editor window with:

```
> data<-edit(data.frame())
```

You don't need to type the > character. That's the prompt in R.

Type the data as shown in the figure below, then **double-click** the top grey row with the default variable name **var1** etc. and change them to **x1**, **x2** and **x3** in turn. Your Data editor window should look like the one below:



	x1	x2	x3	var4
1	10	15	20	
2	12	15	19	
3	22	11	26	
4	41	9	22	
5				

Now close the Data editor window and list the data object in the R console by typing:

```
> data
```

then pressing the Enter key.

If you try to work on an individual column in this data object, its individual name (x1, x2, or x3) will not be recognised by R. To have direct access to the column making up this data object, issue the command.

```
> attach(data)
```

4 Graphics

Next we will produce a scatter plot of **x2** vs. **x1**.

```
> plot(x1,x2,type='b')
```

To find out what the option `type='b'` does try the Help system with:

```
> ? plot
```

You will now see the scatter plot in its own Graph window. To discard the plot **click** on  in the top right-hand corner of the Graph window.

R has no facilities for editing **graphics** once they have been produced, so the user must get the chart right using the commands.

Graphics can be saved in a selection of formats which can be included in word processed reports: press **File→Save as** for a full selection of **graphics** file formats.

Graphics can also be saved in project files, which R calls **workspace**. See section 6 for more information on workspaces.

Here are a few statistical graphics that you can produce:

```
> plot(data, main='a matrix scatterplot')
> boxplot(data, main='three boxplots in the same frame')
```

5 Descriptive Statistics

Now we will look at some descriptive statistics using basic commands. Try:

```
> summary(data)
```

This command returns a number of descriptive statistics including the sample mean, minimum, maximum and quartiles for the three variables.

Other summary statistics can be obtained by individual commands.

```
> range(data)
> apply(data,2,range)
```

For example, the standard deviation and sample size may be calculated using:

```
> sd(data)
> length(x1)
> apply(data,2,length)
```

Instead of sd(data), you need the command sd(as.matrix(data))
to get an SD over all the data in the dataframe or apply(data,2,sd)
to get an sd for each column separately. Try it!

6 Saving an R workspace

Before moving to the next example we will save all R objects temporarily stored in the default working directory (datasets, output lists and functions) in a single file. Check what is in the directory with:

```
> ls()
```

- Next, from the **File** menu choose **Save Workspace...**
- In the next dialog box, you need to add a filename (e.g. your first name) to the default extension *.RData. The saved workspace file is of type **R images**: this is just a name for a bundle of objects saved together. Save the workspace onto **C:\USER**.

This can also be done with the command:

```
> save.image('c:/user/myname')
```

To quit R, type q() or use the menu at the top.

If now you quit R, then relaunch it, you can retrieve the saved project by **clicking** on

File→Load workspace

and select the ***.RDdata** file on c:\user.

Alternatively, you can also use the following command:

```
> load('c:/user/myname.Rdata')
```

and list the data.

```
> data
```

1. On the JIC system, it's better to use the u: drive, which you can access from any terminal on-site.
2. (This might be specific to JIC's network but) you need to use \\ instead of / between folder names.
3. In my experience, you need to add the file extension .rdata to save the file correctly.
Hence:
save.image('u:\\myname.rdata')

You should also be able to double-click the icon for a .rdata file to load it into R. If that doesn't work first time:
Right-click the icon > Open with... > R for Windows GUI & make sure the box for 'Always use this app to open .rdata files' is ticked.

7 Calculator

R can be used as a calculator. Calculations will either give a single number or range of numbers. For instance:

```
> product <- x2*x3
> product
```

This creates a new vector in the working directory. Note that unless you save your objects before leaving R, the new vector **product** will not be saved in the current workspace.

Note that – is used for subtraction, + for addition, * for multiplication, / for division, ^ for raising a number to a power. For help about arithmetic operators try:

```
> ? '+'
```

You can execute calculations interactively, by not assigning the result to a new object. Try:

```
> sqrt(x1); x1^2
```

To compute the standard error of the x1 column:

```
> sd(x1)/sqrt(length(x1))
```

8 Simulation

R is also very convenient for simulation work. For simulating 100 rows of numerical data from a normal distribution with mean 0 and standard deviation 1, use:

```
> rnorm(100)
```

To create a histogram of another simulation from the same distribution without explicitly saving the results, use

```
> hist(rnorm(100))
```

Now simulate tossing an unbiased coin 100 times and noting each single outcome:

```
> b100 <- rbinom(100,1,0.5)
> b100 <- factor(b100,labels=c('head','tail'))
```

Find a tally of how many heads and tails with:

```
> table(b100)
```

Make a bar chart of these counts with

```
> barplot(table(b100))
```

Make a pie chart of these counts with

```
> pie(table(b100))
```

9 Importing data from other applications

9.1 Import from other statistics packages

You can read into R data formats from a variety of other statistics packages. To do this, we need to load in the current workspace an extra package called 'foreign' that comes with the base installation of R. Do this with:

```
> library(foreign)
```

Now to find out exactly which other data formats this library can read from, use:

```
> library(help='foreign')
```

9.2 Import an individual worksheet from an Excel workbook

You can also read in to R from an Excel workbook, up to Excel 2007, using the xlsx package.

However, the xlsx package is not included in the base installation of R and needs to be installed first then loaded into the current working session of R from the official website.

Do this with the following commands:

```
> install.packages('xlsx',rep='http://cran.r-project.org')
```

Then fill the firewall dialog box as required (if you your computer has one).

Now load this package with:

```
> library(xlsx)
```

To access the help files for this feature type:

```
> ? xlsx
```

Note that 'xslx' is a different package to 'readxl' that we use in other set of notes. There's no strong reason to use one or the other! In R there is often more than one way to do the same thing.

Note that R is case sensitive, so Xlsx is not the same as xlsx.

Now you are ready. To import data stored in a sheet named 'Exam_Scores' from an Excel workbook called 'sms_data.xls' stored on your computer in c:/user, use the following commands:

```
> setwd('c:/user')
> exam_scores <- read.xlsx('sms_data.xls',sheetName='Exam_Scores')
> exam_scores
```

Should be .xlsx

When you type "library(xlsx)", if you get a message saying that rJava can't be loaded, it's probably because you're using a 64-bit version of R and a 32-bit version of Java.

The best way of fixing this is to install the 64-bit version of Java:
Go to www.java.com & click "Free Java Download".
Near the bottom of the page, click "See all Java downloads".
Then download and install "Windows Offline (64-bit)".

Another way is to use the version of R called "R i386", not "R x64".

In my email about this course, there's an attachment with a file called "sms_data.xlsx" (my own version, not the one from Reading). Save it to the u: directory.

Then you need the command
setwd('u:')

STATISTICAL SERVICES CENTRE

Basic Statistics & Design Principles for John Innes Centre

9-10 November 2017

For enquiries about our short course programme or commissioned courses, please contact:

Training Co-ordinator
Statistical Services Centre
Harry Pitt Building
University of Reading
Whiteknights Road
RG6 6FN

Email statistics-courses@lists.reading.ac.uk
Tel +44 (0)118 378 6408

www.reading.ac.uk/ssc

University of Reading

Practical 1 - R

Descriptive Statistics and Exploratory Data Analysis

Objectives

- The aim of this practical is for you to become comfortable using R to:
  - Read in your data
  - Understand how R treats different 'types' of data
  - Obtain summary statistics
  - Obtain graphs

1. Work through the R 3.0 tutorial (in the Appendix), if you did not do this immediately prior to the course.

2. The data set discussed in Session 1 is available in the Excel file named **Introstat.xlsx** in the worksheet **P1-TreeSpeciesData**. Import these data into R as follows:

   - Review section 11 of the introduction to R tutorial to remind yourself of how to read excel files into R. Install and load the xlsx package if you have not already done so.
   - Set the working directory to the folder in which the file is saved. For example, if the dataset is in the folder c:/user:

   ```
   > setwd("C:/User")
   ```

   - Read in the data and assign it to a named data frame within R called **TreeData**.

   ```
   > TreeData<-read.xlsx("Introstat.xlsx",
   sheetName="P1-TreeSpeciesData")
   ```

4. Use the boxplot() function to produce a boxplot for height split by species
   This requires the use of an R formula, which works in the general form of "y~x,data". In this case we want to use height as a variable on the y axis, split by species on the x axis with the data coming from the TreeData data frame. So the command needed is:

   ```
   > boxplot(height~species.name,TreeData)
   ```

   Explain what this boxplot can show us about the relationship between height and species.

5. To produce a scatter plot of height against diameter, with different colours for each species:

   ```
   > plot(height~dgl,data=TreeData,col=species.name,pch=16)
   ```

   You can add a legend to this plot to identify which colour is which using the legend() function:

   ```
   > legend("bottomright", pch=16, col=1:6,
   legend=levels(TreeData$species.name))
   ```

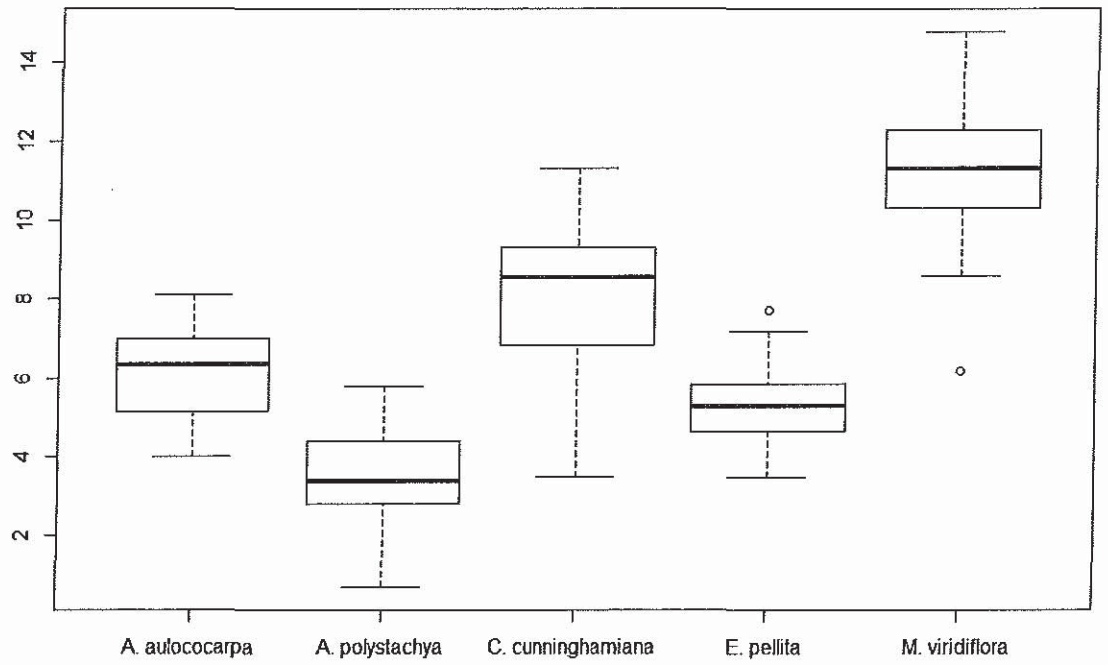
6. You can produce trellis scatter plots like those demonstrated in the lecture using xyplot from the lattice library:

   ```
   > library(lattice)
   > xyplot(height~dgl|location.name,data=TreeData,
   group=species.name, auto.key=list(corner=c(0.9,0.05)))

   > xyplot(height~dgl|species.name,data=TreeData,
   group=location.name, auto.key=list(corner=c(0.9,0.75)))
   ```

7. And trellis box plots using bwplot in a similar manner:

   ```
   > bwplot(height~species.name|location.name,data=TreeData)

   > bwplot(height~location.name|species.name,data=TreeData)
   ```

Solutions



This boxplot has been created using the variable species_name rather than the variable species, in order to do this the variable species_name in needs to be of type *factor* so convert if necessary.

A. aulococarpa has relatively small diameters with little variation. *A. polystachya* has the lowest recorded diameters. There is one outlier in *C. cunninghamiana*, the diameter value is particularly small given the tree belongs to that species, however there are other trees in the sample with similar diameters so it is a plausible value (we should probably check it hasn't been recorded against the wrong species). The largest diameter values were recorded for M. viridiflora, this species also has a large variation in its diameter values compared to other species.

Practical 2 - R

Probability Ideas, The Normal Distribution, Estimation and Confidence Intervals

Objectives

By the end of this practical participants should:

- Understand the difference in interpretation between standard deviation and standard error of mean, and recognise their relationship
- Become familiar with the calculation and interpretation of a confidence interval
- Understand the effect that the sample of size has on the width of a confidence interval

1. A study was conducted to investigate the effect of applying cow dung on *Albizia zygia* seedlings. The heights in cm of seedlings, all treated equally for a certain period of growth are recorded below:

   39.5 60.1 44.8 47.5 49.8 52.0 43.7 56.1

   Use these data to estimate the expected seedling height of *Albizia zygia* seedlings when treated with cow dung under similar conditions (μ) and its standard error.

   Import the data in the worksheet **P2-Seedling** from the Excel workbook **Introstat.xlsx** into a data frame. If you have started a new R session since Practical 1 you will need to set the working directory and load the xlsx package using library in the same way as before. Otherwise, use

```
> seedling<-read.xlsx('Introstat.xlsx',
sheetName='P2-Seedling')
> seedling; attach(seedling)
> summary(seedling)
```

2. For these sample data, obtain the summary statistics mean, standard deviation, sample size and standard error. As there is no function in base R to compute all the above efficiently, we use a contributed package instead.

```
> install.packages('psych',,'http://cran.r-project.org')
> library(psych)
```

Then we use a following dedicated summary function:

```
> describe(height8)
```

Or more succinctly:

```
> describe(height8)[,c(2:4,13)]
```

Record sample size, mean, standard deviation and standard error (SE Mean) below.

What is the algebraic relationship between the standard deviation and the standard error of the mean? How do you interpret the standard error of the mean?

3. Derive a 95% confidence interval for the mean using

```
> t.test(height8)
```

What is the 95% confidence interval for the true mean? Ignore the other components of the output, these will be addressed in the next session.

What does this confidence interval tell you?

If you had found a 99% confidence interval instead, would you expect its width to be larger or smaller than that for a 95% confidence interval?

Check your answer by adding the conf.level argument to the t-test command:

```
> t.test(height8,conf.level=0.99)
```

4. Another investigation into the effect of cow dung on *Albizia zygia* seedlings, which used similar management practices, recorded the heights for 40 seedlings (also measured in cm). These are stored in the column **height40**.

   Produce the same summary stats as above, and obtain a 95% confidence interval for expected height of a seedling under these conditions.

```
> describe(height40)[,c(2:4,13)]
> t.test(height40)
```

What effect does an increased sample size have on the estimate and the confidence interval, and why?

How do your estimates of the mean and standard deviation for this dataset compare with the smaller one?

5. If you have some information in advance about the variability of data, you can use it to help decide how large a sample you might need in a future investigation.

For example, assuming the standard deviation of the seedling heights is about 8cm, determine how large your sample size should be for your estimate $\bar{x}$ to lie within ± 2.8cm of the true mean (expected height) μ with 95% confidence. Use the idea that the 95% confidence interval is approximately $\bar{x} \pm 2$s.e.($\bar{x}$). Ask a resource person if you are unsure what to do here.

Solutions

2. Summary statistics output:

```
> summary(seedling)
    height8          height40
 Min.   :39.19   Min.   :34.86
 1st Qu.:45.04   1st Qu.:44.54
 Median :49.02   Median :49.17
 Mean   :49.20   Mean   :49.20
 3rd Qu.:53.27   3rd Qu.:52.50
 Max.   :59.57   Max.   :65.33
 NA's   :32
> describe(height8)[,c(2:4,13)]
  n mean   sd   se
1 8 49.2 6.75 2.39
> sd(height8,na.rm=TRUE)/sqrt(length(na.omit(height8)))
[1] 2.386485
```

The standard error of the mean is calculated as: $s.e.(mean) = \frac{s}{\sqrt{n}} = \frac{6.75}{\sqrt{8}} = 2.386$

The standard error of the mean is the standard deviation from the theoretical distribution of our sample means (imagine taking many samples from the same population all of the same size). We use it to calculate confidence intervals, as well as conduct statistical tests about our the mean of our sample, our estimate for the population mean.

3.

```
> t.test(height8)

        One Sample t-test

data:  height8
t = 20.616, df = 7, p-value = 1.586e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 43.55686 54.84314
sample estimates:
mean of x
     49.2
```

The 95% confidence interval is 43.55cm to 54.84cm. We are 95% confident that such interval contains the true unknown population mean height.

A 99% confidence interval would be wider as we want to increase the chance that the interval contains the true unknown population mean, so the only way to do this would be to make the interval cover a larger range of values.

```
> t.test(height8,conf.level=0.99)

        One Sample t-test

data:  height8
t = 20.616, df = 7, p-value = 1.586e-07
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 40.84853 57.55146
sample estimates:
mean of x
     49.2
```

We can see that the width of our confidence interval has increased from 11.28cm to 16.70cm.

4.

```
> describe(height40)[,c(2:4,13)]
   n mean   sd   se
1 40 49.2 6.75 1.07
> sd(height40)/sqrt(length(height40))
[1] 1.067269
> t.test(height40)

        One Sample t-test

data:  height40
t = 46.099, df = 39, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 47.04124 51.35875
sample estimates:
mean of x
     49.2
```

The mean is 49.2cm with a 95% confidence interval of 47.04cm to 51.36cm.

The increased sample size improves the precision of our estimate, i.e. it reduces the width of our confidence interval. It does this as the sample size is in the denominator in the expression for the standard error, so although the sample standard deviation (s) may be similar, its value divided by ($\sqrt{n}$) results in a smaller standard error.

Both mean and standard deviation of the two samples are identical; not so the respective sample size, hence different standard errors.

5. Calculating the required sample size for a given level of precision:

Use the expression for calculating confidence intervals:

$$\bar{x} \pm 2s.e.(\bar{x}) = \bar{x} \pm 2\frac{s}{\sqrt{n}}$$

We want our confidence interval to be 2.8cm either side of the mean, so we substitute this into the last part of the above expression, along with our estimate of σ:

$$2.8 = 2\frac{8}{\sqrt{n}}$$

Rearrange to find n:

$$1.4 = \frac{8}{\sqrt{n}}$$

$$\sqrt{n} = \frac{8}{1.4}$$

$$n = 32.65 \approx 33$$

You can check your answer by putting 33 back into the expression.

$$2\frac{8}{\sqrt{33}} = 2.79 \approx 2.8$$

Practical 3 - R

Hypothesis Testing and Comparing One Group to a Target

Objectives

- Become familiar with the calculation and interpretation of a simple hypothesis test.
- Understand the relationship between a hypothesis test and a confidence interval.

1. Assume the usual method of growing *Albizia zygia* seedlings yields plants whose mean height, after the same duration that our researcher in Practical 2 was using, is 45cm. Our researcher is therefore interested to know whether the cow dung makes a difference to the height of the seedlings.

   If you have started a new R session since Practical 2 then reload the data saved in the worksheets P2-Seedling.

2. Use the data stored in column **height8** to test the hypothesis H_0: $\mu = 45$ versus H_1: $\mu \neq 45$. Use the t-test function as in practical 2 but add the argument mu=45.

   ```
   > t.test(height8,mu=45)
   ```

   Record your results below: ensure you understand all the output produced here. Ask if you do not. For instance, how was the T-statistic calculated?

Hypothesis test statistic (the T-statistic):
p-value:
Calculations and Interpretation:

Are your results for the hypothesis test consistent with your interpretation of the confidence interval recorded earlier?

3. Now test the same hypothesis as in question 2 above, but using the data stored in **height40**. Note down your results and conclusions below. Are your conclusions consistent with the results of the confidence interval you found using the second sample of data in Practical 2?

```
> t.test(height40,mu=45)
```

4. You would have found that the two datasets give different conclusions. Can you explain why? Note down your answer below.

Solutions

2. One-Sample t-Test with n = 8

```
> t.test(Seedling1$Height,mu=45)

 One Sample t-test

data:  Seedling1$Height
t = 1.7485, df = 7, p-value = 0.1239
alternative hypothesis: true mean is not equal to 45
95 percent confidence interval:
 43.5243 54.8507
sample estimates:
mean of x
  49.1875
```

The test statistic is calculated as: $T = \frac{49.19-45}{2.395} = 1.7485$

This test statistic is compared to a t-distribution on 7 degrees of freedom to get a p-value.

As p-value > 0.05 (testing at 5% level of significance), we conclude that there is insufficient evidence to reject the null hypothesis. There is insufficient evidence to reject H_0.

The conclusion is consistent with the confidence intervals calculated in Practical 2 as the 95% confidence interval contains our null hypothesis value of 45cm. for us to reject the null hypothesis our confidence interval should not contain the null hypothesis test value.

3. One-Sample t-Test with n = 40

```
> t.test(Seedling2$Height,mu=45)

 One Sample t-test

data:  Seedling2$Height
t = 3.3566, df = 39, p-value = 0.00177
alternative hypothesis: true mean is not equal to 45
95 percent confidence interval:
 46.84197 52.42803
sample estimates:
mean of x
   49.635
```

As the p-value from this second test is less than 0.05, we conclude that there is strong statistically significant evidence that the mean height does not equal 45cm. This conclusion is consistent with our confidence intervals calculated in Practical 2.

4.

We draw different conclusions from the two height samples, which have very similar means, due to the sample sizes. The second sample has 5 times more data than the first sample, and as such has a much smaller standard error of the mean (better precision), resulting in a larger test statistic and stronger evidence against the null hypothesis.

Practical 4 - R

Hypothesis Testing: Comparing Two Samples

Objective

- To be able to recognise when to do an independent samples t-test and when to do a paired t-test.
- Understand how R output from a two-sample (unpaired) t-test may be interpreted.
- Understand how R output from a two-sample paired t-test may be interpreted.

There are four problems that we will work with in this practical; they are stored in a worksheet named ***P4-Problems*** of the Excel workbook **Introstat.xlsx**. Read in this data into a data frame called **Problems** and attach it in R.

We will start by looking at problem 1:

Problem 1

In an investigation of the effects of dust on tuberculosis, a set of 16 rats were divided at random into two groups, A and B, the animals in group A being kept in an atmosphere containing a known percentage of dust, while those in group B were kept in a dust-free atmosphere. After three months the animals were killed and their lung weights measured. The results are given below.

Lung Weights

Rat (in group)	1	2	3	4	5	6	7	8
Group A	5.79	5.57	6.52	4.78	5.91	7.02	6.06	6.38
Group B	4.20	4.06	5.81	3.63	2.80	5.10	3.64	4.53

1. Analyse the data from problem 1 by conducting a t-test using, interpret the results and write a brief summary of your conclusions.

```
> p1<-t.test(groupA, groupB, var.equal=T); p1
```

Null hypothesis:

Hypothesis test statistic (the T-statistic):

p-value:

Difference in means (and s.e.d.*):

Calculations and interpretation:

***Note:** The R output does not print the standard error of the difference, but we can derive it working backward from the given test statistic using:

```
> names(p1)
> unname(abs(diff(p1$estimate)/p1$statistic))
```

The analysis conducted above assumes that the data in the two groups in problem 1 are independent; this assumption is valid as we know the data came from 16 different rats. Consider the data in problem 2, outlined below, are the two groups (enriched and standard) independent?

Problem 2

Two genetically identical seeds, each pair from one of 10 different sources of plant material, were randomly assigned to be raised in either a nutritionally enriched environment or under standard conditions. After a predetermined time, all plants were harvested, dried and weighed. The results, in grams are shown below.

Source	enriched	standard
1	4.81	4.17
2	4.17	3.05
3	4.41	5.18
4	3.59	4.01
5	5.87	6.11
6	3.83	4.10
7	6.03	5.17
8	4.98	3.57
9	4.90	5.33
10	5.75	5.59

The aim of the study was to investigate whether the enriched environment had an effect on plant dry weight.

The samples in problem 2 are not independent. The problem states that two seeds were used from each source, so we can break down the variation into the variation between sources and the variation within sources. We can use this partitioning of the variation to remove the between source of variation. This can give us a more precise estimate of the difference between treatments, if the variability between sources is relatively large. This is called a paired t-test as we focus on the difference between treatments at the source level, rather than the overall difference between treatments. **A paired t-test is equivalent to a one-sample t-test applied to the 10 pairwise differences.** Please ask if unclear.

2. Analyse the data from Problem 2:

```
> p2<-t.test(enriched, standard, paired=T); p2
```

Null hypothesis:

Hypothesis test statistic (the T-statistic):

p-value:

Difference in means (and s.e.d.*):

Calculations and interpretation:

* Find this quantity using

```
> unname(abs(p2$estimate/p2$statistic))
```

3. Next consider problems 3 and 4. Assess whether they are independent, in which case a Two-Sample test is appropriate, or whether the two samples are not independent in which case a Paired t-test should be used.

Your analysis should consist of the following steps:

(i) Using the appropriate two-sample structure (paired or independent), analyse the data by investigating, using hypothesis testing, whether or not there is a difference between group means.

(ii) Estimate the difference in group means and give a confidence interval for the true difference.

Record your results below, and interpret the results. Make sure you are happy with any assumptions underlying the analysis.

Problem 3

A group of 10 students scored the following marks in a test.

Marks in test (before coaching)

Candidate	A	B	C	D	E	F	G	H	I	J
Score	53	60	61	47	40	56	75	46	82	61

After some extra special coaching, their marks were as follows:

Marks in test (after coaching)

Candidate	A	B	C	D	E	F	G	H	I	J
Score	60	58	67	51	60	72	71	48	94	76

Objective: To determine if the coaching improved the students' marks.

Problem 3:

Independent Samples or Paired?

Test used:

Hypothesis:

Hypothesis test statistic (the T-statistic):

p-value:

Difference in means (and s.e.d.):

Calculations and interpretation:

Problem 4

A motoring organisation decided to investigate the relative merits of two brands of petrol A and B. Using cars of the same make, nine cars were run on brand A and another nine were run on brand B. The distances (km) obtained on a supply of 10 litres of petrol are shown below.

Distance (km)

Car (in group)	1	2	3	4	5	6	7	8	9
Brand A	131.4	131.1	132.4	129.6	133.3	129.2	130.5	130.5	131.2
Brand B	124.6	130.2	128.2	127.5	128.9	130.0	127.2	128.5	129.2

Problem 4:

Independent Samples or Paired?

Test used:

Hypothesis:

Hypothesis test statistic (the T-statistic):

p-value:

Difference in means (and s.e.d.):

Calculations and interpretation:

Solutions

1. The Rat Data

```
> t.test(Problems$groupA,Problems$groupB,var.equal=T)

 Two Sample t-test

data:  Problems$groupA and Problems$groupB
t = 4.3718, df = 14, p-value = 0.0006386
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.908005 2.656995
sample estimates:
mean of x mean of y
  6.00375   4.22125
```

H_0 : The mean lung weights of the rats in each population are equal: $\mu_A = \mu_B$, or that there is no difference between the mean lung weights of the populations: $\mu_A - \mu_B = 0$.

H_1 : mean lung weights of the rats in each population are not equal: $\mu_A \neq \mu_B$, or

there is a difference between mean lung weight of the two populations: $\mu_A - \mu_B \neq 0$.

p-value is <0.001, very strong statistical evidence against the null hypothesis. Conclude that there is a difference in the mean lung weights between the two atmospheres.

Difference (A vs B) estimated at 1.782g (6.004 – 4.222) with s.e.d. = 0.4077, giving 95% CI of 0.908g to 2.657g.

2. The Seed Data

```
> t.test(Problems$enriched,Problems$standard,paired=T)

 Paired t-test

data:  Problems$enriched and Problems$standard
t = 0.8673, df = 9, p-value = 0.4083
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3313193  0.7433193
sample estimates:
mean of the differences
                  0.206
```

As this is a matched, or paired, t-test, the differences are calculated for each pair of observations, and then the average of the differences is compared to 0 using a one-sample t-test. Mean difference is 0.206 (this is positive, so enriched seeds have a larger weight than standard seeds).

H_0 : no true difference in the resulting plant weight between the enriched and standard seeds on average, $\mu_{Enriched} - \mu_{Standard} = 0$.

H_1 : true difference in the resulting plant weight between the two types of seeds on average, $\mu_{Enriched} - \mu_{Standard} \neq 0$.

p-value is 0.408, so there is insufficient evidence to reject the null hypothesis, and we conclude that there is not enough evidence to suggest there is a difference in the weights from the plants resulting from the two different types of seeds.

This conclusion could have been reached using the 95% CI for the difference in plant weights from the two types of seed. Mean difference in weight estimated at 0.2060 g, with s.e.d. = 0.2375, giving 95% CI of -0.3313 to 0.7433. This interval contains zero, so we have insufficient evidence to reject H_0.

3. The Student Data

```
> t.test(Problems$before,Problems$after,paired=T)

 Paired t-test

data:  Problems$before and Problems$after
t = -3.0136, df = 9, p-value = 0.01463
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -13.304943  -1.895057
sample estimates:
mean of the differences
                  -7.6
```

This set of data contains paired or matched observations, as the test has been conducted on the same student before and after coaching, so we can remove a level of variation (difference between students) by using a paired or matched test.

H_0 : Coaching makes no difference to the students' true mean scores, $\mu_{After} - \mu_{Before} = 0$.

H_1 : Students' true mean scores change after coaching, $\mu_{After} - \mu_{Before} \neq 0$.

p-value is 0.015, so statistically significant evidence that the students' true means scores after coaching were different to those prior to coaching. Estimated difference in scores is 7.6 marks; scores after coaching are on average 7.6 marks higher, with s.e.d. = 2.252, giving 95% CI of 1.9 to 13.3 marks.

This test could be performed on either the after-before or before-after differences, the conclusion will be the same. The order in which you enter the columns in the R command will set up which way around the difference is calculated.

4. The Car Data

```
> t.test(Problems$brandA,Problems$brandB,var.equal=T)

 Two Sample t-test

data:  Problems$brandA and Problems$brandB
t = 3.8912, df = 16, p-value = 0.001298
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.259408 4.273925
sample estimates:
mean of x mean of y
 131.0222  128.2556
```

As there are 18 cars of the same model and make were used in total there is no matching or paired observations in this set of data.

An F test concludes that the assumption of equal variances between our two samples is appropriate. The null hypothesis being that the true brand variances are equal, the alternative being that the variances are unequal, the p-value = 0.44, therefore there is insufficient evidence to reject the null hypothesis of equal true variances.

H_0 : no difference (in terms of true mean distance travelled) between the two brands of petrol, $\mu_{Brand\ A} - \mu_{Brand\ B} = 0$.

H_1 : difference (in terms of true mean distance travelled) between the two brands of petrol, $\mu_{Brand\ A} - \mu_{Brand\ B} \neq 0$.

p-value is 0.001, so there is strong statistically significant evidence of a difference in the mean distance travelled by 10 litres of the two brands of petrol. The estimated difference is 2.767 km in favour of brand A, with s.e.d. =0.711, giving 95% CI of 1.259 to 4.274 km.

Practical 8 - R

Introducing Simple Linear Regression

Objectives

- Learn to construct and interpret simple linear regression models.
- Be comfortable with checking the assumptions underlying analysis of variance models.

1. In this question you will go through Section 4 of the Genstat guide which details the dialogue boxes used for Regression in Genstat.

   Import the data from the Excel file ***Introstat.xlsx*** using the sheet called **P8-cmtut5**.

   First look at a scatter plot of uptake against concentration:

   ```
   > plot(cmtut5$conc,cmtut5$uptake)
   ```

   To produce the regression model use the lm function and assign the model to an object called lm1 and then use the relevant functions to obtain different outputs from the model:

   ```
   > lm1<-lm(uptake~conc,data=cmtut5)
   > anova(lm1)
   > summary(lm1)
   > confint(lm1)
   ```

   Summarise what output is produced from each of these 3 commands, and what it can show us. Write down and interpret the regression equation from this output.

   To superimpose the regression line onto your original plot then use

   ```
   > abline(lm1)
   ```

Model Checking:
Using the plot() function on an lm object produces four graphs automatically, however we shall concentrate on the first two. To plot the first two plots side by side you can change the graphical parameter settings:

```
> par(mfrow=c(1,2)); plot(lm1,1:2)
```

(i) The first plot is the standardised residuals plotted against the fitted values, this plot should ideally contain a random scatter of points between -2 and 2. Common problems observed from this plot is a funneling out of data points (or funneling in) which is an indication of violating the constant variance assumption, and points following a curved pattern which indicates another variable may be needed in the analysis. *Helpful tip:* It can be tricky to conclude whether the points are random, especially if you have a small sample size, so if you could add one extra point anywhere on the plot and you would be satisfied that the points are randomly scattered then there is no need to worry – but if you have to add several extra points to be satisfied then the model is not appropriate.

(ii) The second plot is a Normal qq plot of the standardised residuals; this should be approximately normally distributed, as the t and F tests rely on the normality assumption.

If you want future plots to be plotted as one graph per page rather than two graphs side by side you need to reset the graph parameters:

```
>par(mfrow=c(1,1))
```

2. Coakley and Lines (1981) carried out a study of the climatic variables affecting the incidence of stripe rust disease on winter wheat. Part of this involved the following data. Negative degree days are a measure of the severity of the winter. The data are listed below and are also available in the Excel file ***Introstat.xls*** in the worksheet named **P8-Rust**.

Negative Degree Days X	Disease Index y
580	4.0
500	5.5
460	5.5
710	3.0
570	3.0
630	4.0
420	7.0
480	7.0
870	1.0
520	4.0
480	4.0
460	6.0

Import the data into R, and plot disease index (Y) against negative degree days (X).

```
>plot(rust$X,rust$Y)
```

What do you conclude about the relationship between Y and X?

Fit the linear regression model $y = \alpha + \beta x$ relating disease index (y) to negative degree days (x).

```
> lm2<-lm(Y~X,data=rust)
> anova(lm2)
> summary(lm2)
> par(mfrow=c(1,2)); plot(lm2,1:2)
```

Review and interpret the output, use the prompts in the box below to fully interpret the model.

See the end of *Session Notes 8b: Regression and ANOVA* for patterns to look out for when examining model checking plots.

ANOVA:

Hypothesis:

p-value: Interpretation:

s^2, RMS Interpretation:

How much of the total variance is explained by the straight line model?
R^2_{adj} =

Regression Parameters:

What is the equation of the fitted line?

Interpret the slope parameter estimate:

What is the approximate 95% confidence interval for the true value of the regression slope?

Modeling Checking:

From the fitted model plot does the model look appropriate?

Is the normality assumption valid?

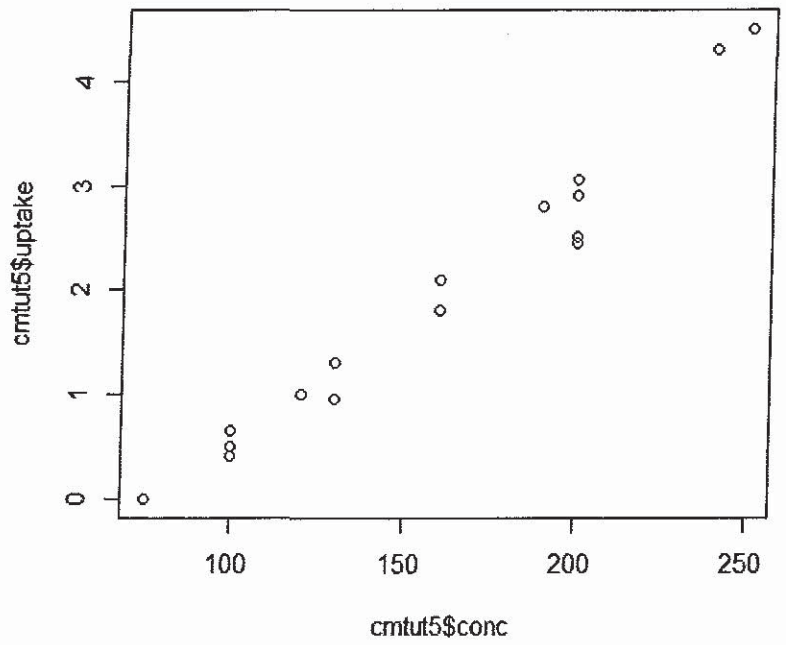
Is the assumption of constant variance valid?

Are there any outliers?

Any other comments?

Solutions

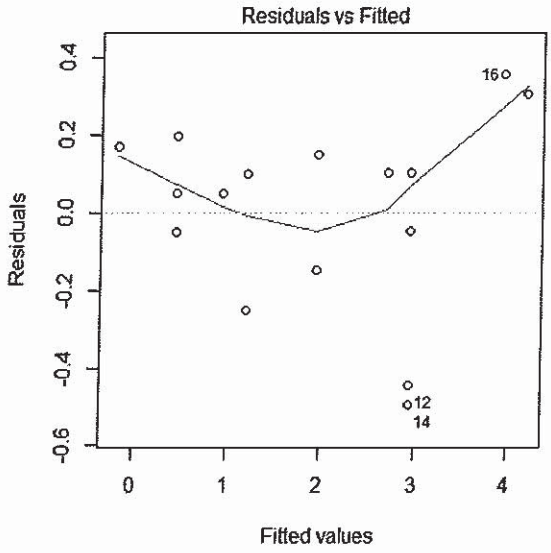
1.

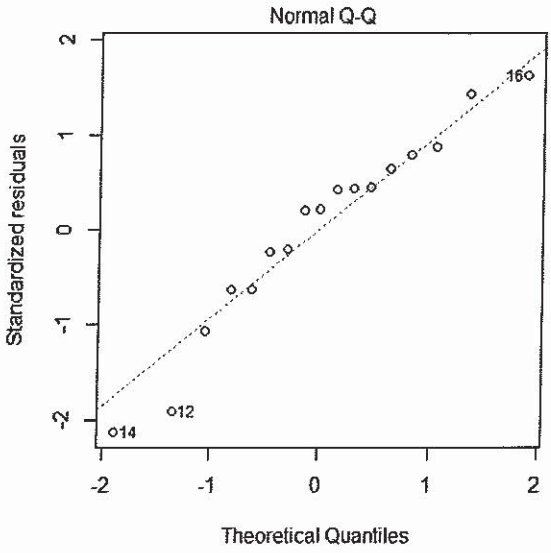



```
> summary(lm1)

Call:
lm(formula = uptake ~ conc, data = cmtut5)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49631 -0.14851  0.04928  0.15149  0.35590

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.042668   0.197917  -10.32 3.29e-08 ***
conc         0.024945   0.001182   21.11 1.44e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2457 on 15 degrees of freedom
Multiple R-squared:  0.9674, Adjusted R-squared:  0.9653
F-statistic: 445.6 on 1 and 15 DF,  p-value: 1.442e-12
```

$$uptake = -2.04 + 0.025 \times conc$$

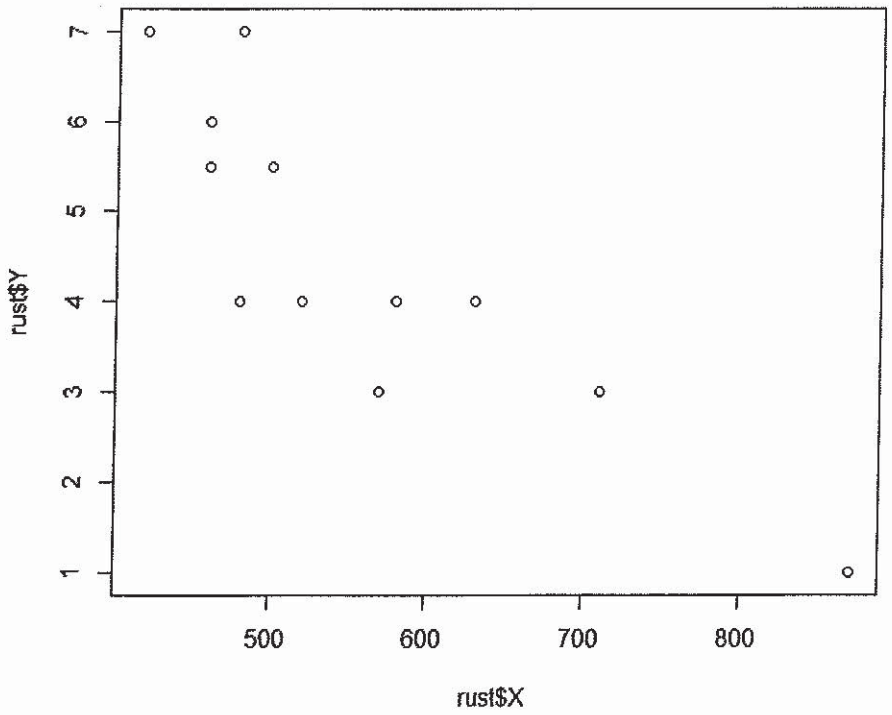
Concentration has a highly statistically significant relationship with uptake ($p<0.001$). A 1 unit increase in concentration is associated with a 0.025 unit increase in uptake.





No major issues with residual plots

2. Regression analysis of the Rust data.



From the scatter plot there appears to be a negative relationship between x and y (as x increases then y decreases). The relationship appears to be fairly linear.

```
> lm2<-lm(Y~X,data=rust)
> anova(lm2)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
X          1 25.9491 25.9491  30.347 0.0002586 ***
Residuals 10  8.5509  0.8551
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(lm2)

Call:
lm(formula = Y ~ X, data = rust)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4168 -0.4003  0.2846  0.3523  1.5832

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.156686   1.237510   9.015 4.07e-06 ***
X           -0.011958   0.002171  -5.509 0.000259 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9247 on 10 degrees of freedom
Multiple R-squared:  0.7521, Adjusted R-squared:  0.7274
F-statistic: 30.35 on 1 and 10 DF,  p-value: 0.0002586
```

The F-test in the ANOVA is testing whether there is a linear relationship between x (negative degree days) and y (disease index), or more specifically:

Null hypothesis: $\beta = 0$, that the slope parameter in the model = 0, i.e. it is not needed and the model can be reduced to $y = \alpha$, there is no relationship between disease index and negative degree days.

Alternative hypothesis: $\beta \neq 0$, that the slope parameter in the model does not equal 0, and that the x variable explains a significant amount of the variation in the y variable.

The p-value for this test is <0.001, therefore we conclude that there is strong statistically significant evidence that the slope parameter does not equal 0 and that we should keep x in our model.

The output also states that 72.7% of the variation in disease index is explained by our model (negative degree days) (calculated as $1 - \frac{\text{Residual MS}}{\text{Total MS}}$).

The reported "standard error of the observations" is the unexplained variation in our data, that is the variation in y that is not explained by our model (this is $\sqrt{\text{Residual MS}}$).

The parameter estimates gives us our model:

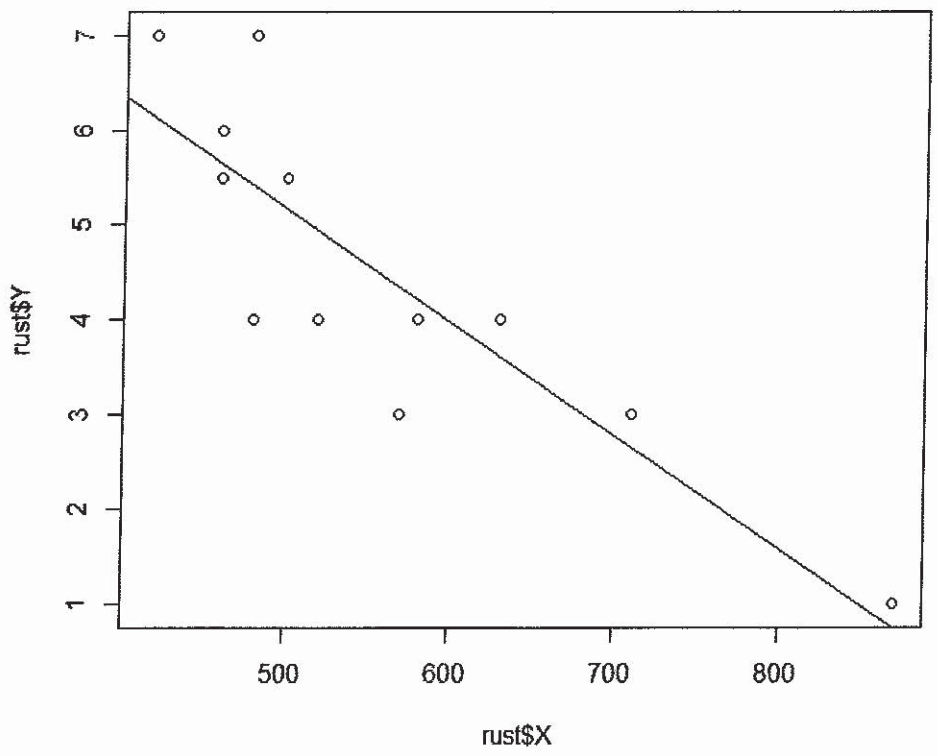
$$y = 11.16 - 0.01196x$$

The interpretation of the intercept (11.16) is not relevant to our data as this relates to y when x = 0, our range of x is between 420 and 870. The interpretation of the slope, -0.01196, is that this is the estimated change in disease index for an additional negative degree day.

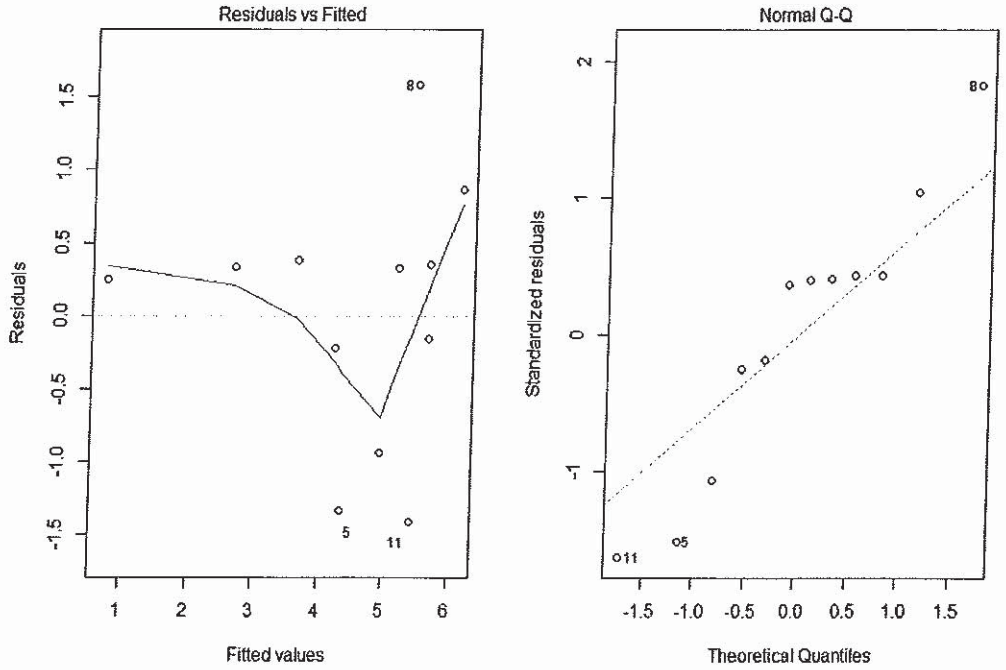
We are given standard errors of the estimates, along with the t-tests testing whether each parameter = 0 or not.

We should also routinely request for additional statistics such as confidence intervals for the parameter estimates.

A graph of the fitted model can be obtained using abline. The plot shows that the model fits our data relatively well.



Model checking plots show that the assumptions behind the model are possibly not appropriate, particularly the assumption of constant variance. However we must keep in mind the small sample size when interpreting these plots.


Residuals vs Fitted
Residuals
Fitted values
8
5
11
Normal Q-Q
Standardized residuals
Theoretical Quantiles
8
5
11

STATISTICAL
SERVICES CENTRE

Experimental Design & Analysis for John Innes Centre

28 February -1 March 2017

For enquiries about our short course programme
or commissioned courses, please contact:

Training Co-ordinator
Statistical Services Centre
Harry Pitt Building
University of Reading
Whiteknights Road
RG6 6FN

Email statistics-courses@lists.reading.ac.uk
Tel +44 (0)118 378 6408

www.reading.ac.uk/ssc

University of Reading

Practical 3 - R

Linear Models

Objectives

- To investigate the order of fitting of terms in a multiple regression model.
- To learn how R is used to analyse non-orthogonal data
- To understand how R output from non-orthogonal data structures may be interpreted

1. Plankton data

- The plankton data from Session 3 are stored in file **ExpDesign.xls**, sheet **S3_plankton**. Import this dataset into R repeat the analysis that was discussed in the lecture. Use the following script.

```
> mod1<-lm(y~x1+x2)
> anova(mod1)
```

What terms were statistically significant?

- Now reverse the order of explanatory variables in the dialogue box, i.e. fit terms in the order x2 x1.

What results are the same as before and which ones are different?

Can you explain why some are the same and some are different?

- If you had to choose between the following models, which one would you choose and why?

(i) x1 and x2

(ii) x1 only

(iii) x2 only

2. Tomato Experiment (one last time)

1. Preparing the data

- The data from Session 2c are available in the file **ExpDesign.xls**, sheet **P3_TomExp2**. ~~Either use data from this worksheet or your own data from Experiment 2 during this practical. If you choose the latter option, ensure your data frame includes columns for block, heat, light, variety and yield, with the first four of these as factors.~~

2. Initial Analysis

- Now use the –lm- function to conduct a non-orthogonal analysis of your dataset. Store results in an object using an assignment.

  As terms to be fitted, indicate ~ **block** + **heat*light*var**. Note that the order in which the block and the treatment factors appear matters. To correctly interpret the ANOVA F-tests for treatment factors and their interactions, we need to look at their effects **after** adjusting for block differences.

- Obtain an ANOVA table for the above model: is it the same as you had before?

- Note that if you request a summary of the fitted model, you get parameter estimates but no predicted means for each factor separately.

  To obtain adjusted means for heat, for example, use the following script:

```
> lsmeans(mod,~heat)
```

  An estimate of the difference between the two predicted means is and a corresponding 95% CI are found using:

```
library(multcomp)
comps<-glht(mod,lsm(pairwise~heat))
summary(comps,test=adjusted(type="none"))
confint(comps)
```

  Note if heat had more than 2 levels we would need to amend the last line in the R script to

```
tm<-qt(0.975,mod$df.residual); tm
confint(comps,calpha=tm)
```

- Comment on the results.

3. Further Analysis

We often want to go further in the analysis to fully reflect the factorial structure of treatments, instead of treating each unique combination of factor levels as a set of unrelated treatments.

Suppose there was a statistically significant two-way interaction, say between heat and light, giving four unique combinations. Ignoring the factorial treatment structure and conducting all six pairwise comparisons would require an adjustment for multiplicity. Instead it is best to convey the meaning of a statistically significant interaction (a.k.a. effect modifier) graphically as follows.

First assign the adjusted means to an object:

```
> adm<-lsmeans(mod,~heat*light); adm
```

Then plot these in a line plot with light on the *x* axis using:

```
> lsmip(adm,heat~light)
```

The function name –lsmip- stands for "least squares means interaction plot".

Read the online help file for the lsmip function, to try finding how to illustrate graphically a statistically significant three-way interaction.

This final example is intended for later practice.

3. Strawberry trial data (described in Session 3)

The data are stored in sheet **P3_Straw** in the file **ExpDesign.xls**. Use either the ANOVA or the Regression approach to fit the model for the trial as designed, i.e. as a RCBD (randomised complete block design). What are your conclusions?

- You now now notice (using ideas we will introduce in later sessions) that the plot number is also a blocking factor. See the field layout on slide 11 of Session 3 to understand what this means.

  You could therefore use the plot number as an additional blocking factor. However, it results in a very unbalanced design.

- Convert the variable **plot** into a factor. Then obtain frequency counts (with the margins) for **Block** and **Genotype** first to check the design is balanced; then repeat for **Plot** and **Genotype**.

- Use a linear model to model the variability in the yield response. Specify **Block** + **Plot** as the nuisance blocking factors + **Genotype** as the treatment.

  What are your conclusions now?

Outline solutions

1. Plankton data

```
> mod1<-lm(y~x1+x2)
> anova(mod1); summary(mod1)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x1         1 10.915 10.9147  12.301 0.003484
x2         1 13.511 13.5109  15.227 0.001595
Residuals 14 12.422  0.8873

Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.49822 -0.44595  0.09693  0.39407  2.12059

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.338488   0.628992  -2.128  0.05159
x1           0.011778   0.006932   1.699  0.11142
x2           0.009077   0.002326   3.902  0.00159

Residual standard error: 0.942 on 14 degrees of freedom
Multiple R-squared:  0.6629,    Adjusted R-squared:  0.6147
F-statistic: 13.76 on 2 and 14 DF,  p-value: 0.0004949

> mod2<-lm(y~x2+x1)
> anova(mod2); summary(mod2)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value   Pr(>F)
x2         1 21.8642 21.8642 24.6416 0.000208
x1         1  2.5613  2.5613  2.8867 0.111415
Residuals 14 12.4220  0.8873

Call:
lm(formula = y ~ x2 + x1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.49822 -0.44595  0.09693  0.39407  2.12059

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.338488   0.628992  -2.128  0.05159
x2           0.009077   0.002326   3.902  0.00159
x1           0.011778   0.006932   1.699  0.11142

Residual standard error: 0.942 on 14 degrees of freedom
Multiple R-squared:  0.6629,    Adjusted R-squared:  0.6147
F-statistic: 13.76 on 2 and 14 DF,  p-value: 0.0004949
```

Table of regression coefficients is unchanged regardless of the order of fitting predictors, as all estimates are adjusted for all other terms in the regression model, which remain unchanged. However, *x1* and *x2* are correlated (not orthogonal), so the order of fitting will affect the F test results.

The order of fitting predictors changes the hypothesis tests:

A regression model with *x1* only needs adding *x2*, as the latter p-values < 0.05.

But a regression model with *x2* only, does not need adding *x1*, as the latter p-value > 0.05.

So when using *x2* for predicting *y*, addition of *x1* is not "statistically significant", but may be required on scientific grounds.

2. **Strawberry trial**

RCBD model

```
> rcbd<-lm(yield~block+genotype)
> anova(rcbd)
Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value  Pr(>F)
block      3  1.217  0.4058  0.1996 0.89547
genotype   7 29.085  4.1550  2.0436 0.09698
Residuals 21 42.697  2.0332
```

No evidence of a genotype effect (p-value = 0.097), adjusting for block.

Analysis as unbalanced design

```
> table(block,plot)
     plot
block 1 2 3 4 5 6 7 8
    1 1 1 1 1 1 1 1 1
    2 1 1 1 1 1 1 1 1
    3 1 1 1 1 1 1 1 1
    4 1 1 1 1 1 1 1 1
```

Effects of block and genotype are orthogonal, i.e. can be evaluated independently; provided no other terms are fitted in the linear model, order of fitting these 2 effects does not change the respective hypothesis tests.

```
> table(plot,genotype)
    genotype
plot E F G M P R S V
   1 2 0 1 0 0 0 0 1
   2 0 1 0 0 1 0 1 1
   3 0 0 0 2 0 2 0 0
   4 0 1 1 0 1 0 1 0
   5 0 0 2 0 1 0 1 0
   6 1 1 0 1 0 0 0 1
   7 1 1 0 0 0 0 1 1
   8 0 0 0 1 1 2 0 0
```

Effect of plot and genotype are not independent, so the order of fitting changes the respective hypothesis test. It is customary to adjust treatment for nuisance factors (i.e. aspects that either cannot be manipulated, or are not under evaluation), so fit blocking factors first and genotype after.

```
> mod1<-lm(yield~block+plot+genotype)
> anova(mod1)
Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value    Pr(>F)
block      3  1.217  0.4058  0.6364   0.60387
plot       7 51.020  7.2886 11.4294 7.76e-05
genotype   7 11.835  1.6907  2.6512   0.05713
Residuals 14  8.928  0.6377
```

Only weak evidence of a genotype effect (p-value = 0.057), adjusting for block and plot. As not significant at 5% level, further comparisons of the genotypes are not pursued.

Practical 5 – R

Linear models with Both Categorical and Continuous Predictors

Part 1

Firstly, we consider the apple data from the example on slide 3 of Session 5.

1. Start with slide 4 and import the apple data, which is stored in the sheet **P4_apples** in the workbook **ExpDesign.xls**. Attach the data frame object.
2. Produce the graph shown in slide 4, which should indicate that a parallel-line model is likely to be needed. We need a contributed package from CRAN, so use:

```
> install.packages('car'); library(car)
> scatterplot(postyield~preyield|treat,smoother=0)
```

3. Fit the model using:

```
> mod1<-lm(postyield~preyield+treat)
> summary(mod1); anova(mod1)
```

The output gives the model estimates and the ANOVA table as shown in slide 7.

4. Now check if an interaction is needed in the model. The model may be specified using **~treat*preyield** above. What are your conclusions? See slide 6. Are they consistent with the graph produced earlier?

6. Is the model any different if the pre-yield covariate is fitted after treat, i.e. fitting terms in the order: **treat**, **preyield** and then their interaction? [Hint: Use **~treat+preyield+treat:preyield**.]

7. Now return to the lm object storing results of the **parallel-lines model**. Does the order of fitting of **preyield** and **treat** matter? What results (F-value and p-value) should you present to summarise whether or not there is a difference between the treatments?

8. To reproduce results shown in slide 9, obtain predicted treatment means with

```
> library(lsmeans)
> lsmeans(mod1,~treat)
```

At which value of **preyield** are the above predicted means formed?

Obtain predicted treatment means for specific values of pre-treatment yield, for example, **preyield** values of 5, 10, and 15. Do this using the following script:

```
> prog<-list(treat=c('C','P','Q'),preyield=c(5,10,15))
> lsmeans(mod1,'treat',by='preyield',at=prog)
```

9. Write down the assumptions for your model. Check these assumptions visually using residual plots:

```
> par(mfrow=c(1,2),pch=19); plot(mod1,1:2)
```

Are the assumptions plausible?

10. Finally, compare all treatment levels to each other, i.e. conduct all pairwise comparisons. Use a Bonferroni multiplicity adjustment to keep the family-wise significance controlled at 5%. Summarise your findings.

```
> library(multcomp)
> comps<-glht(mod1,lsm(pairwise~treat))
> summary(comps,test=adjusted(type='bonferroni'))
```

A set of Bonferroni adjusted simultaneous 95% confidence intervals is obtained with:

```
> bt<-qt(1-0.05/(3*2),20)
> confint(comps,calpha=bt)
```

The results are interpreted as follows: there is a 95% chance that these intervals all simultaneously contain the true mean difference corresponding to the three pairwise comparisons.

Part 2

A varietal trial was conducted on soybean to assess yield. Four varieties were tested, laid out as a randomised block design. Variety 1 is used traditionally by farmers, while 2, 3 and 4 were varieties developed under a new breeding programme. There was canker infestation throughout the experimental area, which varied from plant to plant. It was believed that the varieties did not differ in their susceptibility to canker. The researcher felt that the variety comparison could be improved by adjusting for canker infestation. The data are presented below, yield measured in bushels/acre and canker infestation as percent of stem infestation.

Block	Variety 1		Variety 2		Variety 3		Variety 4	
	C.I.(X)	Yield(Y)	C.I.(X)	Yield(Y)	C.I.(X)	Yield(Y)	C.I.(X)	Yield(Y)
1	19.3	20.3	10.1	28.3	4.3	26.7	18.0	25.1
2	29.2	17.7	34.7	20.7	48.2	14.7	30.2	20.1
3	1.0	26.7	14.0	26.0	6.3	25.0	7.2	24.9
4	6.4	25.3	5.6	34.1	6.7	29.0	8.9	29.8

The data are available in 4 columns, stored in the sheet **P5_soya** in **ExpDesign.xls**.

canker:	Canker infestation; the *x* variable
yield:	Yield; the *y* variable of interest
block:	Block number
treat:	Treatment (variety) number

First convert both block and treat to factors.

Plot **yield** versus **canker** infestation for each **variety**. Comments on the resulting graph, e.g. is there a linear relationship?

Do you think the relationship is the same for each variety?

Fit a linear model to assess if the canker infestation and treatment have a significant effect on the yield, accounting for the block-to-block variability. Also assess the interaction between canker infestation and treatment.

Investigate and explain the differences in the model output which come about by changing the order in which the terms are added to the model. Which model makes the most sense for this example where the researcher is looking to determine if there are significant differences between treatments, having adjusted for any effects caused by block and canker infestation?

Obtain adjusted treatment mean yields for each of the varieties using results from the currently fitted model. Note the results below. Obtain also the raw (unadjusted) means for each variety. Informally compare both sets of means: are they different? If so, why?

Variety	Adjusted Means	Raw Means
1		
2		
3		
4		

(e) Compute the differences in means, their associated standard errors and p-values for the comparisons between all pairs of varieties. Use the Bonferroni correction to adjust for multiplicity, and obtain corresponding simultaneous 95% confidence intervals for the true differences.

Write down your conclusions for reporting your results in a scientific journal.

(f) Finally, conduct a residual analysis and comments on your findings.

Outline solutions

1. Refer to example in Session 5.

2.

```
> mod1<-lm(yield~block+canker*treat)
> anova(mod1)
Analysis of Variance Table

Response: yield
             Df  Sum Sq Mean Sq F value    Pr(>F)
block         3 262.140  87.380 41.4420 0.0005904
canker        1  27.446  27.446 13.0170 0.0154144
treat         3  59.259  19.753  9.3684 0.0170552
canker:treat  3   2.252   0.751  0.3560 0.7877126
Residuals     5  10.542   2.108
```

No statistically significant interaction between canker and treatment, so the latter can be dropped and the linear model re-fitted, giving

```
> mod2<-lm(yield~block+canker+treat)
> anova(mod2)
Analysis of Variance Table

Response: yield
          Df  Sum Sq Mean Sq F value    Pr(>F)
block      3 262.140  87.380  54.636 1.132e-05
canker     1  27.446  27.446  17.161  0.003242
treat      3  59.259  19.753  12.351  0.002266
Residuals  8  12.794   1.599
```

Statistically significant treatment effect.

Adjusted treatment means:

```
> lsmeans(mod2,~treat)
 treat   lsmean        SE
 1     22.02346 0.6398964
 2     27.40987 0.6329286
 3     24.06399 0.6338537
 4     25.10268 0.6328653
```

Raw treatment means:

```
> describeBy(yield,treat,mat=TRUE)[,c(2,4:6,15)]
   group1 n   mean       sd       se
11      1 4 22.500 4.217424 2.108712
12      2 4 27.275 5.552402 2.776201
13      3 4 23.850 6.316381 3.158190
14      4 4 24.975 3.960955 1.980478
```

Latter set of means are unadjusted for the effect of either block or canker, and their precision is not based on a common estimate of residual standard deviation across all treatments, so they do not correspond to the linear regression model we fitted to obtain the accumulated ANOVA table.

```
> comps<-glht(mod2,lsm(pairwise~treat))
Note: df set to 8
> summary(comps,test=adjusted(type='bonferroni'))

         Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = yield ~ block + canker + treat)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
1 - 2 == 0  -5.3864     0.9031  -5.965  0.00202
1 - 3 == 0  -2.0405     0.9055  -2.254  0.32561
1 - 4 == 0  -3.0792     0.9029  -3.411  0.05530
2 - 3 == 0   3.3459     0.8944   3.741  0.03418
2 - 4 == 0   2.3072     0.8942   2.580  0.19567
3 - 4 == 0  -1.0387     0.8944  -1.161  1.00000
(Adjusted p values reported -- bonferroni method)
```

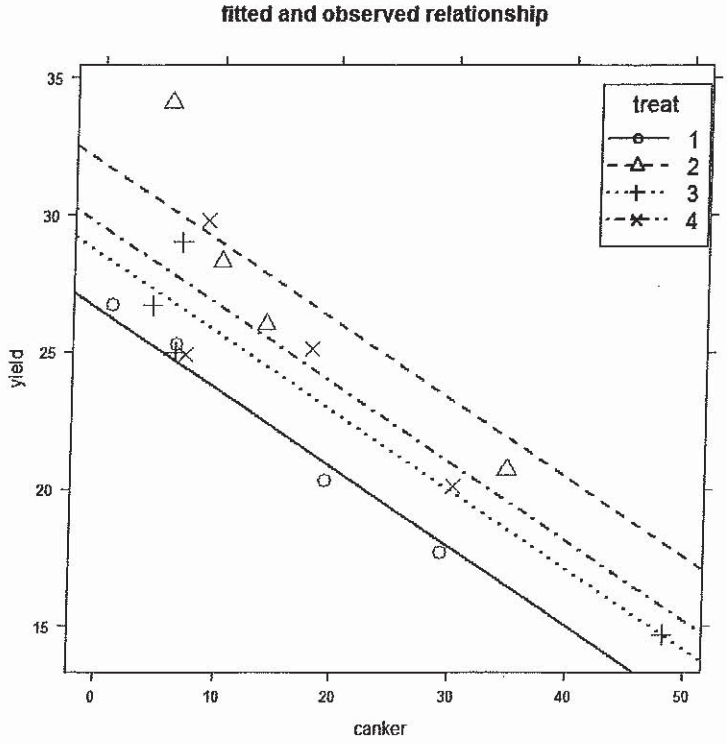
Two pairs out of six are statistically significantly different: 1 and 2, 2 and 3.

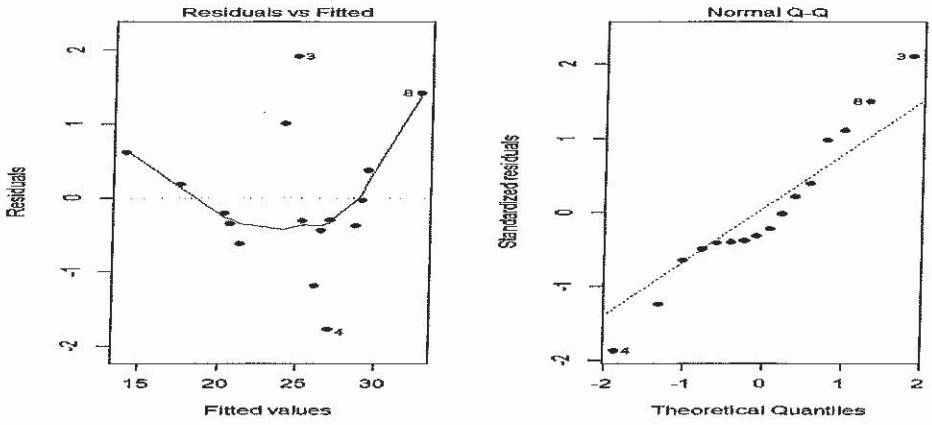
In other words, treatments 2 & 4, 1 & 3 are not statistically significantly different.

```
> bt<-qt(1-0.05/(6*2),8); bt
[1] 3.478879
> confint(comps,calpha=bt)

         Simultaneous Confidence Intervals

Fit: lm(formula = yield ~ block + canker + treat)

Quantile = 3.4789
95% confidence level


Linear Hypotheses:
           Estimate lwr      upr
1 - 2 == 0 -5.38641 -8.52805 -2.24476
1 - 3 == 0 -2.04053 -5.19059  1.10953
1 - 4 == 0 -3.07921 -6.22015  0.06172
2 - 3 == 0  3.34588  0.23443  6.45732
2 - 4 == 0  2.30719 -0.80374  5.41813
3 - 4 == 0 -1.03868 -4.15023  2.07286
```

95% confident that the above intervals simultaneously contain the true value of all six pairwise differences. Only the first and third intervals do not contain zero, i.e. mean of treatment 2 is higher than that of treatment 1 by 5.4 units, with 95% CI (2.245; 8.528) and of that of treatment 3 by 3.35 units, with 95% CI (0.234; 6.457).

The currently fitted model is shown below:





If concerned by the non-straight line appearance of the normal probability plot and by the fan-like appearance of the residual vs fitted plot, could try modelling log(yield) instead of yield. But sample size is not particularly large, so might conclude assumptions hold roughly.