Introduction to Data Analysis with Normal Linear Models and R

- *Where to start: in the lab session you should* ***start at Section 4*** *(introductory session, about the trees data). Useful to look at the earlier sections at some stage.*
- *There is plenty of other material to practise on: e.g. all of the R examples from lectures.*

1 Introduction

This practical Lab aims to familiarize you with the R core commands needed to carry out exploratory data analysis, fitting normal linear models, and doing goodness of fit in readiness for doing the assessed practicals later on in the term. It is not intended to be a coherent lesson in R-programming. There is some basic R for you to work through in your own time if you haven't seen R before and would like a bit more background:

`http://www.stats.ox.ac.uk/~laws/SB1/Rbasics.R`

and

`http://www.stats.ox.ac.uk/~laws/SB1/RbasicsSolutions.R`

If you would like a more detailed introduction to R-programming then I recommend you work through *An Introduction to R* which you can obtain here:

`https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf`

There are many commands you might find useful and this session does not give examples of them all. When preparing your answers to the first assessed practical (which starts in week 5 and must be handed-in by the deadline at the start of week 8) you will find the R-examples associated with the lectures useful also. This session ends with a 'mock' practical and is in no way assessed.

You can download a copy of R for your own computer from `https://cran.r-project.org/`. There are versions available for Windows, Linux, and Mac. While you're there, you should also download some of the R documentation. Follow the *Manuals* link in the left hand column under the heading *Documentation*. There you will see *An Introduction to R*. It provides a good reference at a level suitable for new users. There are many other useful resources here.

Much of this introductory practical is borrowed from chapters 1 and 2 of *Modern Applied Statistics with S-PLUS* (MASS) by Venables and Ripley.

At some point (certainly not yet) you may need to download an install an R-package you don't have on your computer: see `http://www.stats.ox.ac.uk/~laws/SB1/Packages.html`.

To download and install a package using the standard RGui (not RStudio), you can follow these steps. From the R menu bar, click on "Packages", then on "Install package(s)...". Choose UK (Bristol) as CRAN mirror site, then the required package (for example `alr4` used in the first lecture example), and click ok. Then, to use a function or data object from the package in a given R session, you must first load the package by typing the `library()` command (for example, `library(alr4)`). In RStudio it is even easier: click on the "Tools" menu and then "Install Packages..." and type in the name of the package – you probably can't install a package on one of the lab machines (and you won't need to).

2 Getting Started

Start R. Under windows find R in the Start menu and click on it.

3 Getting Help

R includes an on-line help facility that can be invoked from the command line. For example, to get information on the function `var` use the command `help(var)`. This command can be abbreviated "`?`".

```
?var
```

For a feature specified by special characters and in a few other cases (one is "`function`"), the argument must be enclosed in single or double quotes. For example, to get help using the assignment operator `<-` use one of the following

```
help("<-")
?"<-"
```

Now practice using the help facility. For instance, what do the functions `abs`, and `pt` do? If you can't guess what `length`, `sqrt`, `mean`, and `var` do look them up too. The help facility works for non-function objects as well (as long as their creator has provided documentation). Use `library(MASS)` and `help(chem)`, or `?chem`, to see the help file describing the `chem` object in the MASS library.

The `help` function works well if you know the name of the function that you want help with. If you don't know the name of the function you want help with then the situation is a little trickier. R provides a function `help.search` that can be used to search for keywords in help files. Why not use `help(help.search)` to learn how to use it? The help file for `help.search` is rather complicated so scroll down to the bottom and see if you can figure it out from the examples. Use `help.search` to find the S function for fitting *linear models*. Alternatively, try using Google. Which way is easier?

Finally, sooner or later you are going to do something that makes R get stuck. The two most common ways to get stuck are endless loops (or really long computations that you didn't intend) and unbalanced parenthesis. When you are in the endless loop situation the prompt looks like this

```
 while(TRUE) {}
-
```

In this situation press Esc (in Windows) or control-c (in Linux) to cancel the currently running command and return the command prompt. When you have unbalanced parenthesis the prompt looks like this

```
> t <- sqrt(n) * (mean(x) - mu / std.dev(x)
+
```

The way out of this is to type a bunch of right parenthesis, hit return (which generates a syntax error), then type the command again but with the parenthesis in the right places. Typing Esc (in Windows) or control-c (in Linux) will also return you to the command prompt.

4 Introductory R Session

Some of the expressions which appear below will not be familiar: this is deliberate and you should access the Help facility to find out what they do and how they are used.

Consider how you would the analysis below. A detailed solution can be found here

`http://www.stats.ox.ac.uk/~laws/SB1/RnotesForTreesDataAnalysis.R`

Consider the `trees` data from lectures which are built into R. Let us look at fitting a simple normal linear model (NLM).

1. Inspect the data. Plot the data.
2. Fit the simple normal linear model `Volume ~ Girth + Height` (i.e. without the log-transformation of covariates and response that we did in lectures).
3. Test for an effect due to height.
4. Make a couple of diagnostic plots (i.e. plots involving residuals, etc) to examine the plausibility of the NLM: `Volume ~ Girth + Height`
5. Briefly state your conclusions.

5 Example of Assessed Practical-style problem

Frogs of four species had their oxygen consumption measured at two temperatures and two exercise levels. There were two frogs of each species at each temperature, and each of the two was measured both at rest and during forced exercise. Think of *Oxygen Consumption* as the response, and the rest of the variables as potentially explanatory.

The data are set out in the following way.

Variable	Description
Subject	1-16
Species	1-4
Temperature	Low or High
Rest	Oxygen consumption (ml O2/g/hr) at rest
Exercise	Oxygen consumption during exercise

These data and description are available at

`http://www.statsci.org/data/general/frogs.txt`

and on the course webpage `http://www.stats.ox.ac.uk/~laws/SB1/data/frogs.txt`

The data come from: Zar, J. H. (1996). Biostatistical Analysis, Fourth Edition. Prentice-Hall International, Upper Saddle River, New Jersey. Exercise 14.5.

1. Perform exploratory data analysis and give a brief summary of the data.

   What is the problem? How will we use the data to answer it? What are the variables, are any changes needed to prepare the data for analysis? You should uncover the orthogonal design somewhere indicating consequences.

2. Model the relation between *Oxygen Consumption* and the available explanatory variables using a normal linear model for the response, or some function of the response. You should consider possible interactions between the explanatory variables. Carry out outlier analysis and model selection, clearly describing and motivating each step.

Build and fit model. In your report be completely clear what models you fit and how you carry out tests for model selection; give mathematical expressions for the model with all quantities defined and all assumptions explicit; the orthogonal design relevant here. Carry out outlier analysis; check model goodness of fit - analysis should look at QQ-plots, plot fitted v. studentised resid, and cooks-distances (topic covered next lecture) at least.

3. Comment on your findings.

   Which variables matter and what does it mean? Are there any problems (eg overfitting/Data dredging).

Consider how you would answer this. A detailed solution can be found here

`http://www.stats.ox.ac.uk/~laws/SB1/RnotesForFrogsDataAnalysis.R`

6 What to do next

In week 5 we will have a practical session like this one which will kick off your Assessed Practical. You might like to see an example of a report like the one we expect from you. See

`http://www.stats.ox.ac.uk/~laws/SB1/Week3PracticalSampleReport.pdf`

and the associated file of R-commands

`http://www.stats.ox.ac.uk/~laws/SB1/RnotesForRatsDataAnalysis.R`