Statistical Modelling and Computing

Exercises 1

Thursday 24 April, micro room 42, PINNI

The objective of the first micro session is to gain familiarity with WebCT and to try using R by repeating some basic commands given in the first few pages of the course. I will give a quick demonstration of some of these exercises at the beginning of the class.

1) First, login into WebCT and explore the various pages available. Not all of them will have anything on them but they soon will have. Read all of the messages on the discussion list. You should spend about 10 – 15 minutes at most doing this.

2) Look at the websites for R to look at the various notes on usage of R referred to on age 2. This is useful if you want to install R on your own machine, (10 minutes maximum).

3) Call up the R programme. You will find this under the Start menu under Programs and it is labelled Rgui (for R graphical user interface).

a) Find out what libraries are available on your system by typing `library()`.

b) If the MASS library (**M**odern **A**pplied **S**tatistics with **S**) is available then find out what is inside it (i.e. what statistical facilities or commands it has and what data sets are provided with it) with `library(help=MASS)`. Note that R thinks that upper case

LETTERS and lower case letters are different so you must type `MASS` and not `Mass` or `mass`. Note that the list of commands and data sets are all given in one single alphabetic list.

c) Open the MASS library with `library(MASS)`and find out what data sets are available to you with `data()`. You will see that they are listed first for those from the library MASS and then the standard ones that come with the base package. You will need to open the MASS library for almost all exercises in this course and checking what data sets are available is always a good idea before you try to open a particular data set since some have UPPER CASE letters in their names and others do not, e.g. data set `abbey` is all lower case but `Aids2` has a capital letter. R will not recognise `aids2` as the name of a data set.

d) try pressing the up arrow key ↑ and notice that you can retrieve commands that you issued — these can be edited and this saves a lot of time if you want to correct a small typing error in a complicated command.

4) Repeat all of the commands (***including the mistakes!***) given in the notes on pages 8 to 12, looking at the comments on these given on pages 13 – 15.

5) If you have time then use the help system to find out more about some commands and then read ahead in section 2.2 Graphical Summaries and try out some of the worked examples there on pages 23 & 24, and maybe further….

Statistical Modelling and Computing

Exercises 2 (and Assignment 1)

Friday 25 April, micro room 42, PINNI

Assignment 1: This first assignment consists of two parts. The first is to send the mail message referred to in Q1 below. The second is to send me the code and pictures obtained from doing questions 3 and 4. The easiest way of doing this is to cut and paste from the R windows into a Word document.

All assignments should be headed with your name &/or student number.

1) (**Assignment part 1**). Use WebCT to send me a message to say that you have completed all of Exercises 1. The message can say something like "I confirm that I have completed all the parts of yesterday's exercises and I found it very ……… to do them"

2) If there was anything you did not follow or understand in Exercises 1 then either

a) post a discussion message raising the query

or

b) send me a mail message and I will put the query up on the discussion board for others to see.

3) (**Assignment part 2a**). Look again at the data on InsectSprays: Try the following:

```
data(InsectSprays)
attach(InsectSprays)
xcount<- jitter(count)
par(mfrow=c(1,2))
```

```
boxplot(count)
rug(count,side=2)
boxplot(xcount)
rug(xcount,side=2)
```

The effect of `jitter(count)` is to separate out the individual observations by a very small amount so that you can see (for example) whether they are evenly spaced our or clustered together in small groups. Now produce a display with the rug plot of the jittered data on the boxplot of the actual counts.

4) (**Assignment part 2b**). Following the example on P32, generate 100 random numbers from a standard Normal distribution.

a) Display them in a histogram with a plot of the density of N(0,1) superimposed (as in P32/33).

b) Repeat the histogram with a kernel density estimate instead of the 'true' density, using

i) the default bandwidth

ii) bandwidth adjusted to a smaller value

iii) bandwidth adjusted to a larger value.

5) (Not for the assignment but if in doubt on how to do this then post a discussion messsage). How would you do kernel density estimation for half-normal data? i.e. data obtained by say `x<- abs(rnorm(100))` which gives just positive values? Experiment and see what happens if you use the method above.

Statistical Modelling and Computing

Exercises 3

Monday 28 April, micro room 42, PINNI

Optional 1: These questions do not constitute part of the formal assessment. If you would like to submit them to receive feedback, especially if you are in doubt on any parts and have not received an answer via the discussion board then you can submit them under the heading 'Optional 1' by Monday 5 May.

1) Several new elements of the **R** language have been introduced in the last couple of sessions without specific comment. This is a good time to check these and note for further use. In particular:

a) Page 40, how to find the value of the y-coordinate of the density estimate at a specific x-value with `n=1` and `$y` in the call.

b) The idea that we can nest calls to commands and calculations, e.g. `lines(density(eruptions,……` and `t.test(A-B)` instead of doing these in two stages with `dens<-density(eruptions)` followed by `lines(dens)` or `D<- A-B ; t-test(D)` and note use of `;` to put two commands on one line.

c) Use of `qqnorm()` and `qqline()` to check normality on P47

d) Use of a specially written function to calculate the one-sample t-statistic on P61.

e) On page 61, the trick of sorting the simulated values and only printing out the first few hoping that the observed value t_{obs} would have a value in that range. This saves printing all 1000 [ordered] values but of course it will not always work.

2) Estimate the median (providing standard errors of the estimate) of the *waiting times* in the data set on the Old Faithful geyser eruptions, using

a) a kernel density estimate with a variety of appropriate bandwidths

b) bootstrapping

3) In the example on the shoes data on P61 the signs of the observed differences were randomly changed to provide a randomization test by randomly interchanging the labels A and B. perform a bootstrap test of whether the differences have mean zero by taking 1000 samples of size 10 with replacement from the observed differences –0.8, –0.6, ……,–0.3, calculating the t-statistic using the function given on p61 and seeing if the observed value of t = –3.3489 is surprisingly small in comparison with the bootstrap values of the t-statistic.

4) The correct analysis of the shoes data is to use the fact that the A and B samples are paired. However, as an exercise consider them as two separate samples and calculate the two-sample t-statistic by a function z(.,.)

```
z<-function(x,y)(mean(x)-mean(y))/sqrt(var(x)/length(x)+var(y)/length(y))
```

(check that `t.test(A,B)` and `z(A,B)` give the same answer) and then perform

a) a randomization test

b) a bootstrap test

Statistical Modelling and Computing

Exercises 4

Tuesday 29 April, micro room 42, PINNI

1) This class should be used to finish off the material in Chapter 3 and to check back on the material covered so far. In particular, several different *'computer intensive techniques'* (a rather old-fashioned name nowadays) were presented to test fro the difference between materials A and B in the *shoes* data. It is worth thinking carefully about just what these techniques test and what the differences between them are. If you are unsure then maybe you can encourage your colleagues to give their ideas on the discussion boards.

2) Chapter 4 has gone through a very large amount of material in very little detail. As well as the *hills* data set another set which has problems with outliers is the *stackloss* data set and this provides an opportunity to try out the methods for yourself on a different data set.

3) Another technique for robust regression is provided by the command `lqs()` which can be found by opening the `lqs library` (same name for command and library). This is worth investigating.

4) A suggestion of how to bootstrap linear models by extracting the residuals and resampling them to generate new data has been given and it is left to you to try an example. For example you can try to generate the bootstrap distribution of the estimate of the regression coefficient β. This has not been given in the course because it will not be followed up in later parts but if your special interest is in this direction then you can follow it. To be done efficiently you will need to write some functions to calculate the regression estimate of β.

5) If you have not yet tried Q5 on Exercises 2 now is the time to do so.

Exercises 5 (and Assignment 2)

Thursday 30 April, micro room 42, PINNI

Assignment 2: This assignment repeats some of the tasks in earlier Exercises on a new data set which has a slightly different structure. Submitted assignments (by Tuesday 13 May) should include all **R** code used to produce the answers (perhaps as an appendix) and there should be enough explanation (in English please) for me to understand what you have done. You may raise queries about the computing aspects via the discussion board of WebCT.

All assignments should be headed with your name &/or student number.

1) The data file cats in the MASS library gives the body weights and the heart weights of a total of 144 cats, 47 are Female and 97 are Male. This question will consider only the heart weights.

a) Provide separate kernel density estimates of the male and female heart weights.

b) For the male cats only, estimate the median heart weight using two different methods to calculate the standard error of the estimate.

c) Test the hypothesis that male and female cats have equal mean heart weights using

i) a two-sample t-test

ii) a bootstrap assessment of the two-sample t-statistic

iii) a randomization test

2) For the cat data referred to above, investigate the relationship between body weight and heart weight. In particular:

a) using simple regression techniques do there appear to be any outliers?

b) using scatterplot smoothing does there appear [informally] to be any evidence of a non-constant relationship between body and heart weights?

[Note that intentionally this question is left open ended and you may decide for yourself how to interpret it, e.g. whether or not to separate male and female data etc]

Statistical Modelling and Computing

Exercises 6

Monday 5 May, micro room 42, PINNI

1) Several further new elements of the **R** language have been introduced in the last couple of sessions, again without specific comment. Working through and repeating the analyses given in the notes is one way of making yourself familiar with these and it is good to look at the `help` system to check details. The `help` system is not always as clear as it should be and you should use the micro sessions to check out any uncertainties quickly. If they arise later then you should raise them on the WebCT discussion boards. Particular items are:

a) various commands for changing the internal structure (e.g. creating vectors or matrices such as `c(.)`, `rbind(.)`, `cbind(.)`, `as.matrix(.)`.

b) commands for creating factors and repetitions of values such as `factor(.)`, `rep(.)`.

c) interactive commands such as `identify(.)`, `tree.snip(.)`.

d) use of the many and various options inside commands, e.g. `plot(dataset, type="n")` produces a plot with no points in it — why is this very useful??

2) Look at the data set `fgl` from the MASS library Try analysing this using `pca` and `lda`. try doing some of the randomization assessments of correct classification given in the chapter 5 and maybe also some of the tree-based methods. When using `lda`, try the effect of including `method='t'` or `method='mve'` which provides a robust analyses by using a robust estimate of covariance.

Statistical Modelling and Computing

Exercises 7

Tuesday 6 May, micro room 42, PINNI

1) Review the most recent material presented and in particular check that there are no further urgent matters relating to **R** computing that are best asked in person.

2) Construct a neural network to classify the following one-dimensional data U into two categories A and B: Training data for A={3.2, 4.7, 8.2, 2.9, 6.2}; training data for B={9.2, 11.1, 9.7, 16.2}; unknown test data to be classified U={8.4, 8.6, 8.7, 8.8, 9.0}

a) using 2 units

b) using 3 units

c) Is there any difference in the performance of the networks in a) & b)?

d) Which do you think is better?

3) Construct a neural network to classify the following two-dimensional data U into two categories A and B: Training data for A={ (3.1, 2.4), (3.3, 2.7), (2.9, 2.6), (3.5, 3.1), (4.7, 3.9)};
Training data for B={ (2.4, 3.1), (2.7, 3.3), (2.6, 2.9), (3.1, 3.5), (3.9, 4.7)}; unknown test data to be classified U={(3.1,3.1), (2.4,2.4), (3.5, 2.0)} using as few units as you think can provide a satisfactory solution.

Assignment 3

Thursday 30 April, micro room 42, PINNI

> **Assignment 3:** Submitted assignments (by Tuesday 20 May) should include all **R** code used to produce the answers (perhaps as an appendix) and there should be enough explanation (in English please) for me to understand what you have done. You may raise queries about the computing aspects via the discussion board of WebCT.

Data set `crabs` (which is provided in the `MASS` library) consists of measurements on 200 crabs. The crabs are of two types "B" and "O" (Blue & Orange) coded in variable `sp` and there are 50 Males and 50 Females of each, coded in variable `sex`. Five measurements have been made on each crab, coded as variables `FL, RW, CL, CW,` & `BD`. To form a matrix of the measurements you can do `crabmtx<-cbind(FL, RW, CL, CW, BD)` and to form a factor with 4 levels to code for male and female blue and orange crabs you can do

`group<-c(rep("BM",50),rep("BF",50),rep("OM",50),rep("OF",50)).`

There are three potential problems of interest:

a) can you discriminate between Blue and Orange crabs?

b) can you discriminate between Male and Female crabs?

c) can you discriminate the 4 categories of M & F Blue & Orange crabs?

This assignment is to investigate these questions using some of the techniques in the later chapters of this course (i.e. `lda`, `tree` and `nnet` etc.)

Specific tasks and questions to start are

1) what 'raw' correct classification rate is provided by `lda()` in each of the four cases (i.e. use `predict` to classify the same data as used to construct the `lda` object)

2) if you take a random sample of half the data (half of each category) and calculate the lda object and then estimate the correct classification rate by predicting the categories of the other half of the data what answers do you get — maybe try this for a few different samples (i.e. repeat it 2 or 3 times). Note, here the commands `crabsamp.lda<-lda(sp~crabmtx[samp])` and `crabsamp.ld <- predict(crabsamp.lda, crabs[-samp,])` may be useful.

3) can you answer both of the above questions using other methods such as tree(.) and nnet(.)? (Do not answer just yes or no! — show some attempt at doing it).

It is well-known that whereas the Iris Data is particularly nice`FL, RW, CL, CW,` & `BD`, everything works very well on it, the Crabs Data is particularly awkward and difficult and you may get results which look very different from those on the Iris Data. If in doubt you should raise your queries on WebCT. Don't worry too much about giving ideas to others — in return you will get more ideas back.