Statistical Modelling and Computing

Dr Nick Fieller

Department of Probability & Statistics

University of Sheffield

UNIVERSITY
OF TAMPERE

visiting

University of Tampere

2000/01

Statistical Modelling and Computing

0. Introduction

0.1 Books and Websites

Venables, W. N. & Ripley, B. D. (1999) ***Modern Applied Statistics with S-PLUS****,* (3rd Edition), Springer.

This is the main course book. The software (including versions in R) and datasets used in this book are available from various websites such as

http://www.stats.ox.ac.uk/pub/MASS3

A full list of sites can be found at

http://www.stats.ox.ac.uk/pub/MASS3/sites.html

This course will use many of the data sets and functions from the MASS library.

Nolan, D. & Speed, T. P. (2000), ***Stat Labs: Mathematical Statistics Through Applications.*** Springer. Support material is available at:

http://www.stat.Berkeley.edu/users/statlabs

This book is recommended for additional reading.

Ripley, B.D. (1996) ***Pattern Recognition and Neural Networks***. Cambridge University Press.

This book provides much fuller details of neural nets from the practical statistical point of view.

Much of the course will be focused around the computing system R which provides various statistical facilities including high quality graphics. It is an open source system and is available free. It is 'not unlike' the expensive commercial package S-PLUS 2000, the prime difference is that R is command-line driven without the standard menus and dialog boxes in S-PLUS. Otherwise, most code written for the two systems is interchangeable.

The sites from which R and associated software (extensions and libraries) and manuals can be found are listed at

http://www.ci.tuwien.ac.at/R/mirrors.html

The nearest ones are at

http://cran.dk.r-project.org (in Denmark)

and

http://cran.uk.r-project.org (in Bristol, UK)

Free versions of full manuals for R (mostly in PDF format) can be found at any of these mirror sites. There is also a wealth of contributed documentation. Particularly useful are:

Using R for Data Analysis and Graphics by John Maindonald (PDF [702kB], 106 pages). Many of the topics in this course are covered in these notes.

R for Beginners by Emmanuel Paradis, (PDF [152kB], 31 pages). This provides a useful introduction to R. The notes are translated from the original version in French (but not always very accurately).

R reference card by Jonathan Baron, (PDF [58kB], LaTeX [5kB], 1 page)

These can be consulted online during R sessions or downloaded and printed to take away.

0.2 Objectives

The overall objective of this course is to provide an introduction to some of the techniques of modern statistical methodology. An integral part of modern statistical analysis is directed towards understanding data, discovering structure in it and making inferences about the wider world. Applied Statistics is not a subset of mathematics, though mathematics is a useful tool in developing statistical methods and techniques, just as it is a useful tool in the various forms of engineering. In some ways, this course regards applied statistics as '*data engineering*' — this includes actually doing practical things with data. Inevitably, some attention has to be given to the computational side and there will be some pointers to the mathematical aspects.

A great revolution in statistical practice occurred with the development of the language S and later the development of S-PLUS.

(R is essentially the same language as S-PLUS but is free)

This integrated computing system has allowed the statistical community to extend traditional methods and to try out new techniques to provide new ways of investigating practical statistical problems. Often these are based not on mathematical development but on more intuitive ideas. This course aims to give a flavour of this new approach to statistical thinking and an introduction to implementing them in practice.

0.3 Outline of Course

1. Overview of S-PLUS and R:– how does it work and what can it do.

2. Exploratory Data Analysis:– standard summary descriptions and plots, robust summaries, improved alternatives to histograms.

3. Classical Univariate Statistics:– revision and implementation of one and two sample tests, analysis of variance, bootstrap and permutation methods.

4. Linear Statistical Models:– classic linear regression and diagnostics. Robust methods, smooth regression and additive models.

5. Multivariate Methods:– multivariate EDA, principal components and biplots, discrimination and classification, cluster analysis.

6. Tree-based Methods:– Classification and Regression Trees, trees for decision making.

7. Neural Networks:– use for classification and regression problems.