

DR LEE FAWCETT

MAS1802

PROBLEM SOLVING AND
COMPUTING II
(COMPUTING WITH R)

NEWCASTLE UNIVERSITY

Useful calendar:

Week beginning...	Mon	Tues	Thurs	Fri
4th March	×	Intro lecture 4-5 (Herschel LT1)	Intro lecture 11-12 (Bedson LT1)	Practical 1 11-1 (Herschel PC/ Barbara Strang 2.40)
11th March	Practical 1 DI 10-11 (Herschel PC)	Lecture 4-5 (Herschel LT1)	×	Practical 1 deadline: 11:00 Practical 2 11-1 (Herschel PC/ Barbara Strang 2.40)
18th March	Practical 2 DI 10-11 (Herschel PC)	Lecture 4-5 (Herschel LT1)	×	Practical 2 deadline: 11:00 Practical 3 11-1 (Herschel PC/ Barbara Strang 2.40)
25th March	Practical 3 DI 10-11 (Herschel PC)	Lecture 4-5 (Herschel LT1)	×	Practical 3 deadline: 11:00 Project 11-1 (Herschel PC/ Barbara Strang 2.40)



Week beginning...	Mon	Tues	Thurs	Fri
29th April	Project DI 10-11 (Herschel PC)	Lecture 4-5 (Herschel LT1)	×	Project 11-1 (Herschel PC/ Barbara Strang 2.40)
6th May	Bank holiday! ×	Lecture 4-5 (Herschel LT1)	×	Project 11-1 (Herschel PC/ Barbara Strang 2.40)
13th May	Project DI 10-11 (Herschel PC)	Project deadline: 16:00	×	×

Contents

<i>1</i>	<i>Introduction and Housekeeping</i>	5
<i>2</i>	<i>Introduction to R</i>	10
<i>3</i>	<i>Data Summaries</i>	19
<i>4</i>	<i>Control Statements and Functions</i>	27
<i>5</i>	<i>Random Number Generation</i>	35
<i>6</i>	<i>Monte Carlo Methods</i>	43

1

Introduction and Housekeeping

Housekeeping

THE COMPUTING PART of MAS1802 will run during the second half of semester 2 (unfortunately broken up by the Easter vacation!). In direct response to your feedback on the **MatLab** part of MAS1801, we will have weekly lectures, with printed notes, to support the computer practical sessions. Generally, a new topic will be introduced each week and will fit around the following schedule¹:

Type of session	Day	Time	Place
Lecture	Tuesday	4–5	Herschel LT1
Practical	Friday	11–1	Herschel PC/Barbara Strang 2.40
Drop-in	Monday	10–11	Herschel PC
Office hour	Monday	12–1	My office (Herschel 2.07)
<i>Deadline:</i>	<i>Friday</i>	<i>11am</i>	<i>General office/online</i>

¹ To kick-start things, we will have two lectures in the first week, with an additional lecture being held on Thursday 11am. The very first practical session will include an important introductory session in R.

Lecturer information

The module leader, and lecturer for the computing part of MAS1802, is Dr. Lee Fawcett. If you have any questions, comments or feedback on this course I can be contacted via email at lee.fawcett@ncl.ac.uk. More complicated questions about the course are best dealt with face-to-face.



Lectures

Given the nature of this course we will have fewer lectures than the other first year modules – as you might imagine, there will be a strong emphasis on practical skills in the computer cluster. I will hand out copies of the lecture notes and practical worksheets in class.

Slides will be used in lectures to accompany the notes. Lectures will also be recorded using the ReCap system. You can access all course materials, including the ReCap recordings, through Blackboard (BB) or via the dedicated course webpage:

<http://www.mas.ncl.ac.uk/~nlf8/teaching/mas1802/>

Computing practicals

During each computer practical, your general computer skills will be assessed. You are more likely to get high marks for the coursework if:

- You have attended the lecture;
- You don't panic, but read through the question, and the notes, carefully;
- You ask for help²;
- You are organised! For example, if:
 - your files are organised in a sensible manner – i.e. if you have a directory for MAS1802 and sub-directories for each practical;
 - your files are given convenient and meaningful names;
 - Your code “looks nice”.

IN THIS COURSE, most of the computer practical sessions are held in the Herschel PC cluster³, and all machines in the cluster have the necessary software and packages installed. However, you can use computers in other parts of the University while doing your coursework. A list of free computers can be found at:

<http://m.ncl.ac.uk/itservice/#clusfree>

If you use a machine in a cluster and it is missing the necessary software, contact Christian Lawson-Perfect: christian.perfect@ncl.ac.uk.

Office hours

The scheduled office hour for this module is Monday 12–1pm; I will always be free at this time, in my office, to see students who need extra help with the work. I also have office hours for other courses which are usually quiet (see noticeboard outside my office), so feel free to drop in then too. In general, I am happy to see students anytime (if I'm free), except on Wednesdays and Friday after 2pm.

Assessment

- End of semester examination worth **40%** of your overall mark. This will be based on material from the pencil-and-paper part of the course with Evgenios.
- Assessment of coursework worth **60%** of your overall mark:
 - Weekly computing exercises with Lee (10%)
 - Group computing project with Lee (30%)
 - Problem-solving (pencil-and-paper) exercises with Evgenios (20%)

The following deadlines relate to the coursework for the computing part of this module:

- Practical 1: **Friday 15th March, 11:00am** (online submission)

We won't try to catch you out!

² As in MAS1801, there will be lots of help available in computer practicals: Dr. Fawcett, Dr. Cowburn, Dr. Henderson and Dr. Newman will all be on-hand, as will several Ph.D students:



³ Although, due to the size of the class, some of you will be in a smaller satellite cluster for the main Friday practical

- Practical 2: **Friday 22nd March, 11:00am** (hard-copy submission to the General Office)
- Practical 3: **Friday 29th March, 11:00am** (online submission)
- Group computing project: **Tuesday 14th May, 4:00pm** (hard-copy submission to the General Office)

For the group work you will be in the same groups as you were allocated in the problem solving part of the course⁴.

⁴ Although I might change my mind on this one...

Quiz questions

Last summer, we wrote lots of quiz questions in Numbas to help develop your computing skills for this part of the module. These quiz questions are written within the usual computer-based assessment (CBA) framework – if you're interested, go to:

www.mas.ncl.ac.uk/~nlf8/basicR

Late work policy

It is not possible to extend submission deadlines for coursework in this module and no late work can be accepted. For details of the policy (including procedures in the event of illness etc.) please look at the School web site

<http://www.ncl.ac.uk/maths-physics/students/resources>

What is R?

R is a computer package that is widely used for statistical software development and data analysis⁵. R uses a command line interface, though several graphical user interfaces are available. The system provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and others) and graphical techniques.

⁵ For example Google, IBM, Shell, Thomas Cook, Facebook. The recent NY times article at <http://tinyurl.com/9m5qyh> is worth a read.

R is highly extensible through the use of user-submitted libraries for specific functions or specific areas of study. A particular strength of R is its graphical facilities, which produce quality graphs (better than MatLab!) that can include mathematical symbols. Although R is mostly used by statisticians and other practitioners requiring an environment for statistical computation and software development, it can also be used as a general matrix calculation toolbox with comparable benchmark results to many other software packages (e.g. MatLab!). Why is it called R? There are a few reasons, one of which is that it's named after the two people who developed the software: Ross Ihaka and Robert Gentleman, both from Auckland University in New Zealand.

You will use R throughout your degree at Newcastle and the best part is, it's **completely free to download to your own computer!**

Title	Year	Length	Budget	Voting statistics					Movie genre							
				Rating	Votes	r1	...	r10	mpaa	Action	Animation	Comedy	Drama	Documentary	Romance	Short
A.k.a. Cassius	1970	85	NA	5.7	43	4.5	...	14.5	PG	0	0	0	0	1	0	0
AKA	2002	123	NA	6.0	335	24.5	...	1.5	R	0	0	0	1	0	0	0
Alien Vs. Pred	2004	102	45000000	5.4	14651	4.5	...	4.5	PG-13	1	0	0	0	0	0	0
Abandon	2002	99	25000000	4.7	2364	4.5	...	4.5	PG-13	0	0	0	1	0	0	0
Abendland	1999	146	NA	5.0	46	14.5	...	24.5	R	0	0	0	0	0	0	0
Aberration	1997	93	NA	4.8	149	14.5	...	4.5	R	0	0	0	0	0	0	0
Abilene	1999	104	NA	4.9	42	0.0	...	24.5	PG	0	0	0	1	0	0	0
Ablaze	2001	97	NA	3.6	98	24.5	...	14.5	R	1	0	0	1	0	0	0
Abominable Dr	1971	94	NA	6.7	1547	4.5	...	14.5	PG-13	0	0	0	0	0	0	0
About Adam	2000	105	NA	6.4	1303	4.5	...	4.5	R	0	0	1	0	0	1	0

Table 1.1: The first ten rows of the movie data set. **Credit:** This data set was initially constructed by Hadley Wickham at <http://had.co.nz/>.

Accessing R

R is installed on all University machines. In this class we will be using Rstudio, which is an R IDE⁶. Both R and Rstudio are free, so you can install them on you own computer⁷. See

<http://www.ncl.ac.uk/maths-physics/students/resources/>

for more details.

Previous computing knowledge

This course is intended to teach you the basics of programming. **No** previous programming knowledge is assumed, although the **MatLab** part of MAS1801 should have helped your general skills. It is crucial that you come to all the practicals for this course; it is also vital that you let me/other staff know, in the practical sessions, if you're not sure about anything. Burying your head in the sand will not help. Neither will crying, or getting angry⁸.

Recommended R textbooks

On the course webpage I have a list of suitable books on R programming. Since you will be using R throughout your degree, then it may be worthwhile buying a good R textbook. A really good, free-to-download book written by Venables and Smith, can be found here:

<https://cran.r-project.org/doc/manuals/R-intro.pdf>

Movies dataset

The internet movie database, (IMDB) <http://imdb.com/>, is a website devoted to collecting movie data supplied by studios and fans. It claims to be the biggest movie database on the web and is run by *Amazon*. More information imdb.com can be found online:

http://imdb.com/help/show_leaf?about

including information about the data collection process

http://imdb.com/help/show_leaf?infosource

IMDB makes their raw data available at <http://uk.imdb.com/interfaces/>.

⁶ Integrated development environment (IDE).

⁷ Versions of R and Rstudio are available for Windows, Apple Mac and Linux.

⁸ Please don't do this.



Movies are selected for inclusion if they had a known length, had been rated by at least one IMDB user and had an MPAA (*Motion Picture Association of America*) rating. Amongst other things, the dataset contains the following fields:

- **Title**: Title of the movie.
- **Year**: Year of release.
- **Budget**: Total budget in US dollars. If the budget isn't known, then it is stored as 'NA'.
- **Length**: Length in minutes.
- **Rating**: Average IMDB user rating.
- **Votes**: Number of IMDB users who rated this movie.
- **r1-10**: Percentage(to nearest 5%) of users who rated this movie 1, ..., 10
- **mpaa**: MPAA rating.
- **Genre**: Action, Animation, Comedy, Drama, Documentary, Romance, Short. Binary variables representing if movie was classified as belonging to that genre. A movie can belong to more one genre. See for example the film *Ablaze* in Table 1.1.

There are a total of 23 variables and 4988 films. The first few rows are given in Table 1.1. We will use this dataset to illustrate the concepts covered in this class.

This is **only a subset** of the data, the actual data set contains information on over 50,000 movies.

1.4 Useful R package

We have written an add-on R package for the module which will allow easy access to many interesting datasets (including the movies dataset). Installing this package is straightforward, and will make loading datasets used in lectures really easy:

```
> install.packages("ncldata",
+                   repos="http://R-Forge.R-project.org",
+                   type="source")
```

To load the package, use

```
> library(ncldata)
```

Then to load the movies dataset, for example, type:

```
> data(movies)
```

Alternatively, datasets can be downloaded directly from the course webpage/BB page (see practical work this/next week). We will explore the movies dataset in Chapter 2.

When typing in code from the lecture notes, **never** include the + symbols (unless they're actually part of a calculation!). This just indicates, in the notes, that the code has 'wrapped' onto another line.