

MAS1343: Computational Probability and Statistics (with R)

Dr Lee Fawcett

School of Mathematics & Statistics

Semester 2, 2014/2015

Part I

Introduction and Housekeeping

1.1.1 Lecturer information: who am I?

- The lecturer for this module is Dr Lee Fawcett.
- If you have any questions, comments or feedback on this course I can be contacted via email.
- More complicated questions about the course are best dealt with face-to-face.

Timetable

- See your lecture notes for a full week-by-week schedule, including hand-in deadlines
- Generally, we meet twice per week:

Timetable

- See your lecture notes for a full week-by-week schedule, including hand-in deadlines
- Generally, we meet twice per week:
 - Mondays 9–11 :-(in LT3
 - Wednesdays 9–10 :-(in the Herschel PC cluster
- Occasionally, we will also meet on Thursdays at 3 in LT2.

1.1.2 Office hours

The scheduled office hour for this module is Thursday 3–5.

I also have the following office hours for MAS2317, so you can drop in then too: Friday 1–2.

In general, I am happy to see students anytime (if I'm free), except on Wednesdays and Friday after 3pm.

1.1.3 Lectures

- Your calendar at the front of your notes booklet contains a full schedule for this course.
- Check your ncl e-mail regularly for further information
- This course has:
 - Lectures - fewer than other modules
 - Problem classes: in lecture time - fewer than other modules
 - Tutorials: in lecture time - fewer than other modules
 - Drop-in sessions
 - Computing practicals - loads more than other modules
 - Feedback sessions: in lecture time

1.1.3 Lectures

Lectures will also be recorded. You can access the recordings through BB or at the following webpage:

<http://www.mas.ncl.ac.uk/~nlf8/teaching/mas1343/recap/>

The lecture slides will (eventually) be put on BB.

1.1.3 Lectures

Lectures will also be recorded. You can access the recordings through BB or at the following webpage:

<http://www.mas.ncl.ac.uk/~nlf8/teaching/mas1343/recap/>

The lecture slides will (eventually) be put on BB.

1.1.4 Assessment

- End of semester examination worth 60% of your overall mark. In order to pass this module you must get at least 35% in the exam.
- Assessment of course-work worth 40%:
 - A **single** CBA - due week 3;
 - Practicals;
 - Demonstrations of computer skills at the lab.

Practical Weightings

Not all practicals are worth the same marks:

Practical	1	2	3	4	5	6
Weight(%)	5	4	9	9	5	5

Table: Approximate coursework weightings. On average, 1% of coursework will take 1 hour (assuming you are familiar with your lecture notes). CBA 1 is worth 3% of your final mark.

1.1.5 Practicals

During each computer practical, your general computer skills will be assessed. In particular

- Are your files organised in a sensible manner? For example, do you have a directory for MAS1343 and sub-directories for each practical?
- Are your files named in a structured manner?
- Do you use the correct text editor?

To obtain these marks, you must attend at least five computer labs. These marks will form part of your final assessment.

1.1.5 Practicals

All computer practicals are held in the Herschel cluster.

1.1.5 Practicals

All computer practicals are held in the Herschel cluster. A list of other free computers in the University can be found:

<http://www.ncl.ac.uk/iss/teaching/clusters/usage.php>

1.1.5 Practicals

All computer practicals are held in the Herschel cluster. A list of other free computers in the University can be found:

<http://www.ncl.ac.uk/iss/teaching/clusters/usage.php>

and the mobile site is:

<https://m.ncl.ac.uk/iss/>

If you use a machine in a cluster and it is missing the necessary software, contact Dr Chris Graham.

1.2 What is **R**?

- **R** is a computer package that is widely used.
- For example Google, IBM, Shell and spread betting companies use **R** for statistical software development and data analysis.

1.2 What is **R**?

- **R** is a computer package that is widely used.
- For example Google, IBM, Shell and spread betting companies use **R** for statistical software development and data analysis.
- See the NY times article at <http://tinyurl.com/9m5qyh>.
- **R** uses a command line interface, though several graphical user interfaces are available.

1.2 What is **R**?

- **R** is a computer package that is widely used.
- For example Google, IBM, Shell and spread betting companies use **R** for statistical software development and data analysis.
- See the NY times article at <http://tinyurl.com/9m5qyh>.
- **R** uses a command line interface, though several graphical user interfaces are available.
- Why use a command line interface as opposed to Excel, say?

- **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 that can include mathematical symbols.

- **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 that can include mathematical symbols.
- In this class and MAS1343 we use **R**. Since you do so much **R** programming in this class, the **R** in MAS1343 will seem fairly easy.
- See BB for a list of recommended books

1.2.2 Previous knowledge

- This course is intended to teach you the basics of programming.
- **No** previous programming knowledge is assumed.
- Overall, the programming related aspects probably accounts for 50% of the final mark in this course - there are programming questions in the exam!
- It is crucial that you come to all the practicals for this course.

Course Summary

- The mathematics/statistics in this course is straightforward. Much easier than in other modules
- Each individual R command is also easy to understand

Course Summary

- The mathematics/statistics in this course is straightforward. Much easier than in other modules
- Each individual R command is also easy to understand
- **But** this module is usually rated as one of the most difficult. I will ask you to apply the knowledge you learn, and problem–solve.

1.3 Movie data set

- The internet movie database (IMDB), 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 about information imdb.com can be found online.
- Movies were selected for inclusion if they:
 - had a known length;
 - had been rated by at least one imdb user;
 - had an mpaa rating.

IMDB data set

- Title. Title of the movie.

IMDB data set

- Title. Title of the movie.
- Year. Year of release.

IMDB data set

- 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 '-1'.

IMDB data set

- 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 '-1'.
- Length. Length in minutes.

IMDB data set

- 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 '-1'.
- Length. Length in minutes.
- Rating. Average IMDB user rating.

IMDB data set

- 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 '-1'.
- Length. Length in minutes.
- Rating. Average IMDB user rating.
- Votes. Number of IMDB users who rated this movie.

IMDB data set

- 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 '-1'.
- 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 a 1,..., 10.

A big data set

- mpaa. MPAA rating.

A big data set

- mpaa. MPAA rating.
- Action, Animation,... 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 *Ab blaze* in Table 1.1.

A big data set

- mpaa. MPAA rating.
- Action, Animation,... 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 *Ab blaze* in Table 1.1.
- There are a total of 24 variables and 4847 films.
- This is **only a subset** of the data, the actual data set contains information on over 50,000 movies.

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	-1	5.7	43	4.5	...	14.5	PG	0	0	0	0	1	0	0
AKA	2002	123	-1	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	-1	5.0	46	14.5	...	24.5	R	0	0	0	0	0	0	0
Aberration	1997	93	-1	4.8	149	14.5	...	4.5	R	0	0	0	0	0	0	0
Abilene	1999	104	-1	4.9	42	0.0	...	24.5	PG	0	0	0	1	0	0	0
Ablaze	2001	97	-1	3.6	98	24.5	...	14.5	R	1	0	0	1	0	0	0
Abominable Dr	1971	94	-1	6.7	1547	4.5	...	14.5	PG-13	0	0	0	0	0	0	0
About Adam	2000	105	-1	6.4	1303	4.5	...	4.5	R	0	0	1	0	0	1	0

1.4 Module R package

This module has an associated R package. Installing this package is straightforward:

```
> install.packages("mas1343",  
+                 repos="http://R-Forge.R-project.org",  
+                 type="source")
```

To load the package, use

```
> library(mas1343)
```