

MAS1802: Problem Solving II (Statistical Computing with R)

Dr Lee Fawcett

School of Mathematics & Statistics

Semester 2, 2016/2017

1.1 Housekeeping

MAS1802 – A tale of two halves:

- Weeks 1–5: Pencil-and-paper problem solving with Evgenios Kakariadis
- Weeks 6–11: Statistical computing with Lee Fawcett
- Week 12: Revision for the exam with Evgenios

These slides refer to the computing part only.

1.1 Housekeeping

MAS1802 – A tale of two halves:

- Weeks 1–5: Pencil-and-paper problem solving with Evgenios Kakariadis
- Weeks 6–11: Statistical computing with Lee Fawcett
- Week 12: Revision for the exam with Evgenios

These slides refer to the computing part only.

1.1 Housekeeping

Things we've changed since MAS1801:

- You didn't like how MatLab was self-taught: I have timetabled a weekly lecture for the computing part of MAS1802
- You wanted more support in the computer practicals: I have scheduled a weekly drop-in, two days before the assignment deadline, in the computer cluster
- You thought the project should have been more structured: The nature of this course requires some open-ended work, however the project for the computing part of MAS1802 will have more direction
- You wanted treats in class...

1.1 Housekeeping

Things we've changed since MAS1801:

- You didn't like how MatLab was self-taught: I have timetabled a weekly lecture for the computing part of MAS1802
- You wanted more support in the computer practicals: I have scheduled a weekly drop-in, two days before the assignment deadline, in the computer cluster
- You thought the project should have been more structured: The nature of this course requires some open-ended work, however the project for the computing part of MAS1802 will have more direction
- You wanted treats in class...

1.1 Housekeeping

Things we've changed since MAS1801:

- You didn't like how MatLab was self-taught: I have timetabled a weekly lecture for the computing part of MAS1802
- You wanted more support in the computer practicals: I have scheduled a weekly drop-in, two days before the assignment deadline, in the computer cluster
- You thought the project should have been more structured: The nature of this course requires some open-ended work, however the project for the computing part of MAS1802 will have more direction
- You wanted treats in class...

1.1 Housekeeping

Things we've changed since MAS1801:

- You didn't like how MatLab was self-taught: I have timetabled a weekly lecture for the computing part of MAS1802
- You wanted more support in the computer practicals: I have scheduled a weekly drop-in, two days before the assignment deadline, in the computer cluster
- You thought the project should have been more structured: The nature of this course requires some open-ended work, however the project for the computing part of MAS1802 will have more direction
- You wanted treats in class...

1.1 Housekeeping

Things we've changed since MAS1801:

- You didn't like how MatLab was self-taught: I have timetabled a weekly lecture for the computing part of MAS1802
- You wanted more support in the computer practicals: I have scheduled a weekly drop-in, two days before the assignment deadline, in the computer cluster
- You thought the project should have been more structured: The nature of this course requires some open-ended work, however the project for the computing part of MAS1802 will have more direction
- You wanted treats in class...

1.1 Housekeeping

Things we've changed since MAS1801:

- You didn't like how MatLab was self-taught: I have timetabled a weekly lecture for the computing part of MAS1802
- You wanted more support in the computer practicals: I have scheduled a weekly drop-in, two days before the assignment deadline, in the computer cluster
- You thought the project should have been more structured: The nature of this course requires some open-ended work, however the project for the computing part of MAS1802 will have more direction
- You wanted treats in class...

1.1 Housekeeping

Things we've changed since MAS1801:

- You didn't like how MatLab was self-taught: I have timetabled a weekly lecture for the computing part of MAS1802
- You wanted more support in the computer practicals: I have scheduled a weekly drop-in, two days before the assignment deadline, in the computer cluster
- You thought the project should have been more structured: The nature of this course requires some open-ended work, however the project for the computing part of MAS1802 will have more direction
- You wanted treats in class...

1.1 Housekeeping

Computing – General course structure:

Type of session	Day	Time	Place
Lecture	Monday	4–5	Bedson LT1
Practical	Friday	9–11	Herschel PC
Drop-in	Monday	11–12	Herschel PC
Office hour	Monday	1–2	My office (2.07)
<i>Deadline:</i>	<i>Wednesday</i>	<i>4pm/11.45pm</i>	<i>General office/online</i>

- This week: No Monday drop-in, no Friday practical; extra lecture!
- Next week: Monday drop-in → Compulsory introductory practical; no Wednesday deadline

1.1 Housekeeping

Computing – General course structure:

Type of session	Day	Time	Place
Lecture	Monday	4–5	Bedson LT1
Practical	Friday	9–11	Herschel PC
Drop-in	Monday	11–12	Herschel PC
Office hour	Monday	1–2	My office (2.07)
<i>Deadline:</i>	<i>Wednesday</i>	<i>4pm/11.45pm</i>	<i>General office/online</i>

- This week: No Monday drop-in, no Friday practical; extra lecture!
- Next week: Monday drop-in → Compulsory introductory practical; no Wednesday deadline

1.1 Housekeeping

Computing – General course structure:

Type of session	Day	Time	Place
Lecture	Monday	4–5	Bedson LT1
Practical	Friday	9–11	Herschel PC
Drop-in	Monday	11–12	Herschel PC
Office hour	Monday	1–2	My office (2.07)
<i>Deadline:</i>	<i>Wednesday</i>	<i>4pm/11.45pm</i>	<i>General office/online</i>

- This week: No Monday drop-in, no Friday practical; extra lecture!
- Next week: Monday drop-in → Compulsory introductory practical; no Wednesday deadline

1.1.1 Lecturer information: who am I?

- The module leader, and lecturer for the computing part of 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**

1.1.2 Lectures

- **Fewer lectures** for this course than for other modules (nature of the course)
- The calendar at the front of this handout contains a **full schedule** for the computing part of this course
- A summarised version of the notes will be used as slides - interspersed with quizzes and other really fun things. Pay attention!
- Notes, slides and practical sheets will be posted to **BlackBoard**, and a **dedicated webpage**, soon after the lecture/practical session
- Lectures will be recorded via **ReCap**
- Check your ncl e-mail regularly: class cancellations etc.

1.1.2 Lectures

- **Fewer lectures** for this course than for other modules (nature of the course)
- The calendar at the front of this handout contains a **full schedule** for the computing part of this course
- A summarised version of the notes will be used as slides - interspersed with quizzes and other really fun things. Pay attention!
- Notes, slides and practical sheets will be posted to **BlackBoard**, and a **dedicated webpage**, soon after the lecture/practical session
- Lectures will be recorded via **ReCap**
- Check your ncl e-mail regularly: class cancellations etc.

1.1.2 Lectures

- **Fewer lectures** for this course than for other modules (nature of the course)
- The calendar at the front of this handout contains a **full schedule** for the computing part of this course
- A summarised version of the notes will be used as slides - interspersed with quizzes and other really fun things. Pay attention!
- Notes, slides and practical sheets will be posted to **BlackBoard**, and a **dedicated webpage**, soon after the lecture/practical session
- Lectures will be recorded via **ReCap**
- Check your ncl e-mail regularly: class cancellations etc.

1.1.2 Lectures

- **Fewer lectures** for this course than for other modules (nature of the course)
- The calendar at the front of this handout contains a **full schedule** for the computing part of this course
- A summarised version of the notes will be used as slides - interspersed with quizzes and other really fun things. Pay attention!
- Notes, slides and practical sheets will be posted to **BlackBoard**, and a **dedicated webpage**, soon after the lecture/practical session
- Lectures will be recorded via **ReCap**
- Check your ncl e-mail regularly: class cancellations etc.

1.1.2 Lectures

- **Fewer lectures** for this course than for other modules (nature of the course)
- The calendar at the front of this handout contains a **full schedule** for the computing part of this course
- A summarised version of the notes will be used as slides - interspersed with quizzes and other really fun things. Pay attention!
- Notes, slides and practical sheets will be posted to **BlackBoard**, and a **dedicated webpage**, soon after the lecture/practical session
- Lectures will be recorded via **ReCap**
- Check your ncl e-mail regularly: class cancellations etc.

1.1.3 Computer practicals

- Lots of support - make good use of this!



- Attendance closely monitored
- You are more likely to get high marks for the coursework if:
 - You have attended the lecture: In practicals, I will assume you have been to the lecture on Monday; this is reflected in what is expected of you in the practicals
 - You don't panic! Read the notes carefully, and type code **slowly**
 - You ask for help
 - You try to have fun
 - You are organised!

1.1.3 Computer practicals

- Lots of support - make good use of this!



- Attendance closely monitored
- You are more likely to get high marks for the coursework if:
 - You have attended the lecture: In practicals, I will assume you have been to the lecture on Monday; this is reflected in what is expected of you in the practicals
 - You don't panic! Read the notes carefully, and type code **slowly**
 - You ask for help
 - You try to have fun
 - You are organised!

1.1.3 Computer practicals

- Lots of support - make good use of this!



- Attendance closely monitored
- You are more likely to get high marks for the coursework if:
 - You have attended the lecture: In practicals, I will assume you have been to the lecture on Monday; this is reflected in what is expected of you in the practicals
 - You don't panic! Read the notes carefully, and type code **slowly**
 - You ask for help
 - You try to have fun
 - You are organised!

1.1.3 Computer practicals

- Lots of support - make good use of this!



- Attendance closely monitored
- You are more likely to get high marks for the coursework if:
 - You have attended the lecture: In practicals, I will assume you have been to the lecture on Monday; this is reflected in what is expected of you in the practicals
 - You don't panic! Read the notes carefully, and type code **slowly**
 - You ask for help
 - You try to have fun
 - You are organised!

1.1.3 Computer practicals

- Lots of support - make good use of this!



- Attendance closely monitored
- You are more likely to get high marks for the coursework if:
 - You have attended the lecture: In practicals, I will assume you have been to the lecture on Monday; this is reflected in what is expected of you in the practicals
 - You don't panic! Read the notes carefully, and type code **slowly**
 - You ask for help
 - You try to have fun
 - You are organised!

1.1.3 Computer practicals

- Lots of support - make good use of this!



- Attendance closely monitored
- You are more likely to get high marks for the coursework if:
 - You have attended the lecture: In practicals, I will assume you have been to the lecture on Monday; this is reflected in what is expected of you in the practicals
 - You don't panic! Read the notes carefully, and type code **slowly**
 - You ask for help
 - You try to have fun
 - You are organised!

1.1.3 Computer practicals

- Lots of support - make good use of this!



- Attendance closely monitored
- You are more likely to get high marks for the coursework if:
 - You have attended the lecture: In practicals, I will assume you have been to the lecture on Monday; this is reflected in what is expected of you in the practicals
 - You don't panic! Read the notes carefully, and type code **slowly**
 - You ask for help
 - You try to have fun
 - You are organised!

1.1.4 Office hours

- Monday 1-2pm
- Any other office hours (e.g. MAS8306, ACC1012, ACC1053)
- Or just give me a knock (usually not free Wednesday afternoon or Friday after 2pm)

1.1.4 Office hours

- Monday 1-2pm
- Any other office hours (e.g. MAS8306, ACC1012, ACC1053)
- Or just give me a knock (usually not free Wednesday afternoon or Friday after 2pm)

1.1.4 Office hours

- Monday 1-2pm
- Any other office hours (e.g. MAS8306, ACC1012, ACC1053)
- Or just give me a knock (usually not free Wednesday afternoon or Friday after 2pm)

1.1.5 Assessment

- End of semester **examination** worth **40%** of your overall mark. This will include material from the pencil-and-paper part of the course after Easter
- Assessment of **course-work** worth **60%** of your overall mark:
 - 'Weekly' computing exercises (10%: 3@ 3.33% each)
 - Group computing project (30%)
 - Problem-solving exercises (20%)

The weekly computer exercises will be due on the following dates, all of which are Wednesdays:

22nd March (23.45); 26th April; 3rd May (23:45).

The group computing project will be due in by no later than 4pm, Wednesday 17th May.

1.1.5 Assessment

- End of semester **examination** worth **40%** of your overall mark. This will include material from the pencil-and-paper part of the course after Easter
- Assessment of **course-work** worth **60%** of your overall mark:
 - ‘Weekly’ computing exercises (10%: 3@ 3.33% each)
 - Group computing project (30%)
 - Problem-solving exercises (20%)

The weekly computer exercises will be due on the following dates, all of which are Wednesdays:

22nd March (23.45); 26th April; 3rd May (23:45).

The group computing project will be due in by no later than 4pm, Wednesday 17th May.

1.1.5 Assessment

- End of semester **examination** worth **40%** of your overall mark. This will include material from the pencil-and-paper part of the course after Easter
- Assessment of **course-work** worth **60%** of your overall mark:
 - ‘Weekly’ computing exercises (10%: 3@ 3.33% each)
 - Group computing project (30%)
 - Problem-solving exercises (20%)

The weekly computer exercises will be due on the following dates, all of which are Wednesdays:

22nd March (23.45); 26th April; 3rd May (23:45).

The group computing project will be due in by no later than **4pm, Wednesday 17th May.**

1.1.6 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`

- I will make lectures more fun by including quiz questions here, too!

1.1.6 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`

- I will make lectures more fun by including quiz questions here, too!

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?

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?

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?

1.2 What is R?

- **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
- R's syntax is similar to that used in MatLab
- R has loads of **built-in functions** for statistical analysis, but can also do lots of basic mathematical operations too
- R is **completely free** (unlike MatLab), as is the Rstudio software which we will use in this course
- Named after Ross Ihaka and Robert Gentleman (Auckland University, NZ) who developed the software
- Industry-standard statistical software, and used right throughout your degree

1.2 What is R?

- **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
- **R**'s syntax is similar to that used in MatLab
- **R** has loads of **built-in functions** for statistical analysis, but can also do lots of basic mathematical operations too
- **R** is **completely free** (unlike MatLab), as is the Rstudio software which we will use in this course
- Named after Ross Ihaka and Robert Gentleman (Auckland University, NZ) who developed the software
- Industry-standard statistical software, and used right throughout your degree

1.2 What is R?

- **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
- **R**'s syntax is similar to that used in MatLab
- **R** has loads of **built-in functions** for statistical analysis, but can also do lots of basic mathematical operations too
- **R** is **completely free** (unlike MatLab), as is the Rstudio software which we will use in this course
- Named after Ross Ihaka and Robert Gentleman (Auckland University, NZ) who developed the software
- Industry-standard statistical software, and used right throughout your degree

1.2 What is R?

- **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
- **R**'s syntax is similar to that used in MatLab
- **R** has loads of **built-in functions** for statistical analysis, but can also do lots of basic mathematical operations too
- **R** is **completely free** (unlike MatLab), as is the Rstudio software which we will use in this course
- Named after Ross Ihaka and Robert Gentleman (Auckland University, NZ) who developed the software
- Industry-standard statistical software, and used right throughout your degree

1.2 What is R?

- **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
- **R**'s syntax is similar to that used in MatLab
- **R** has loads of **built-in functions** for statistical analysis, but can also do lots of basic mathematical operations too
- **R** is **completely free** (unlike MatLab), as is the Rstudio software which we will use in this course
- Named after Ross Ihaka and Robert Gentleman (Auckland University, NZ) who developed the software
- Industry-standard statistical software, and used right throughout your degree

1.2 What is R?

- **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
- **R**'s syntax is similar to that used in MatLab
- **R** has loads of **built-in functions** for statistical analysis, but can also do lots of basic mathematical operations too
- **R** is **completely free** (unlike MatLab), as is the Rstudio software which we will use in this course
- Named after Ross Ihaka and Robert Gentleman (Auckland University, NZ) who developed the software
- Industry-standard statistical software, and used right throughout your degree

1.2.1 Accessing R

- R is installed on all University machines
- In this class we will be using Rstudio, which is an R IDE
- Both R and Rstudio are free, so you can install them on you own computer. See:
<http://www.ncl.ac.uk/maths/students/teaching/installingr/>

1.2.1 Accessing R

- R is installed on all University machines
- In this class we will be using Rstudio, which is an R IDE
- Both R and Rstudio are free, so you can install them on you own computer. See:
<http://www.ncl.ac.uk/maths/students/teaching/installingr/>

1.2.1 Accessing R

- R is installed on all University machines
- In this class we will be using Rstudio, which is an R IDE
- Both R and Rstudio are free, so you can install them on you own computer. See:
<http://www.ncl.ac.uk/maths/students/teaching/installingr/>

1.2.2 Previous computing knowledge

- This course is intended to teach you the basics of programming in **R**
- No previous programming knowledge is assumed, although MatLab should have helped your general computing 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*

1.2.2 Previous computing knowledge

- This course is intended to teach you the basics of programming in **R**
- **No** previous programming knowledge is assumed, although MatLab should have helped your general computing 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*

1.2.2 Previous computing knowledge

- This course is intended to teach you the basics of programming in **R**
- **No** previous programming knowledge is assumed, although MatLab should have helped your general computing 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*

1.2.2 Previous computing knowledge

- This course is intended to teach you the basics of programming in **R**
- **No** previous programming knowledge is assumed, although MatLab should have helped your general computing 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*

1.2.3 Recommended R textbooks

- At <http://goo.gl/nzQTK>, I have constructed a list of suitable books (with comments) on R programming
- Since you will be using R throughout your degree, then it may be worthwhile buying a good R textbook...
- ... however, a really good, **free-to-download** book, written by Venables and Smith, can be found here:

www.cran.r-project.org/doc/manuals/R-intro.pdf

1.2.3 Recommended R textbooks

- At <http://goo.gl/nzQTK>, I have constructed a list of suitable books (with comments) on R programming
- Since you will be using R throughout your degree, then it may be worthwhile buying a good R textbook...
- ... however, a really good, **free-to-download** book, written by Venables and Smith, can be found here:

www.cran.r-project.org/doc/manuals/R-intro.pdf

1.2.3 Recommended R textbooks

- At <http://goo.gl/nzQTK>, I have constructed a list of suitable books (with comments) on R programming
- Since you will be using R throughout your degree, then it may be worthwhile buying a good R textbook...
- ... however, a really good, **free-to-download** book, written by Venables and Smith, can be found here:

www.cran.r-project.org/doc/manuals/R-intro.pdf

1.3 Movie dataset

- 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 information about 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.

1.3 Movie dataset

- **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

1.3 Movie dataset

- **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

1.3 Movie dataset

- **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

1.3 Movie dataset

- **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

1.3 Movie dataset

- **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

1.3 Movie dataset

- **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

1.3 Movie dataset

- **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

1.3 Movie dataset

- **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 *Ablaze* 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

1.3 Movie dataset

- **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 *Ablaze* 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

1.3 Movie dataset

- **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 *Ablaze* 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 Useful 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)
```

Part II

Introduction to R

2.1 A simple R session

At its most basic, R can be used as a calculator – for example, for multiplication and division...

```
> 5*5  
[1] 25  
> 10.2/6  
[1] 1.7
```

...and more 'advanced' operations:

```
> 2^3  
[1] 8
```

```
> exp(1.5)  
[1] 4.481689
```

```
> log(10)  
[1] 2.302585
```

```
> 4 %% 3  
[1] 1
```

...and more 'advanced' operations:

```
> 2^3  
[1] 8
```

```
> exp(1.5)  
[1] 4.481689
```

```
> log(10)  
[1] 2.302585
```

```
> 4 %% 3  
[1] 1
```

...and more 'advanced' operations:

```
> 2^3  
[1] 8
```

```
> exp(1.5)  
[1] 4.481689
```

```
> log(10)  
[1] 2.302585
```

```
> 4 %% 3  
[1] 1
```

2.1.1 Assignment operations

- In the computer practicals we will use assignment, i.e. $x = 5$.
- However, computer assignment is different from typical mathematical assignment.
- For example,

```
> x = 5  
> x = x + 1  
> x  
[1] 6
```

- Notice that when we type `x = 5`, R doesn't display or print any output to the screen.
- If we want to see what value has been assigned to the variable, we type `x`.
- An alternative is to surround the expression with brackets.
- For example

```
> (x = 2*x)
[1] 12
```

- You can also use the `<-` operator for assignment. This is, for almost all situations, identical to the `=` operator.

Question 1

What is the value of x?

```
> x = 5  
> x = 2*x + 1  
> x
```

- 1 5
- 2 11
- 3 12

Question 1

What is the value of x?

```
> x = 5  
> x = 2*x + 1  
> x  
[1] 11
```

- 1 5
- 2 11
- 3 12

Question 2

What is the value of x?

```
> x = 5  
> y = 6  
> x = x + 2*y  
> x + 5  
> x
```

- 1 5
- 2 10
- 3 17
- 4 22

Question 2

What is the value of x?

```
> x = 5  
> y= 6  
> x = x + 2*y  
> x + 5  
[1] 22  
> x  
[1] 17
```

- 1 5
- 2 10
- 3 17
- 4 22

Question 3

What is the value of x?

```
> x = 5
```

```
> y = 6
```

```
> x + 5
```

```
> x
```

- 1 5
- 2 10
- 3 17
- 4 22

Question 3

What is the value of x?

```
> x = 5  
> y = 6  
> x + 5  
[1] 10  
> x  
[1] 5
```

- 1 5
- 2 10
- 3 17
- 4 22

2.1.2 Data types

R has a variety of data types:

```
> (v = TRUE)
[1] TRUE
```

```
> (w = "fred")
[1] "fred"
```

```
> (x = 5.0)
[1] 5
```

2.1.2 Data types

R has a variety of data types:

```
> (v = TRUE)
[1] TRUE
```

```
> (w = "fred")
[1] "fred"
```

```
> (x = 5.0)
[1] 5
```

2.1.2 Data types

R has a variety of data types:

```
> (v = TRUE)
[1] TRUE
```

```
> (w = "fred")
[1] "fred"
```

```
> (x = 5.0)
[1] 5
```

And some “special” data types:

```
> (y = 5/0)
[1] Inf
> (z = y-y)
[1] NaN
```

Type	Example 1	Example 2	Example 3	Example 4
Doubles	2	3.1242	-45.6	4e-10
Logicals/Boolean	TRUE	FALSE		
Characters	"FRED"	"x"	"Male"	"TRUE"
Infinity	Inf	5/0		

Table: Summary of data types in R.

Another important data type in R is NA. This is used to represent missing values. A list of data tables is given in Table 1.1.

Question 4

What type of variable is x?

```
> x = "1"
```

- 1 Character
- 2 Double
- 3 Logical

Question 4

What type of variable is x?

```
> x = "1"
```

- 1 Character
- 2 Double
- 3 Logical

2.2 The R workspace

Once you create a variable, R stores that variable in memory for reuse. You can view available variables with the `ls()` command:

```
> ls()  
[1] "movies"
```

To delete a variable in R, we use the `rm` function. For example,

```
> x = 0  
> y = 1  
> z = 2  
> ls()  
[1] "movies" "x"      "y"      "z"
```

```
> rm(x)
```

```
> ls()  
[1] "movies" "y"      "z"
```

To delete a variable in R, we use the `rm` function. For example,

```
> x = 0  
> y = 1  
> z = 2  
> ls()  
[1] "movies" "x"      "y"      "z"
```

```
> rm(x)
```

```
> ls()  
[1] "movies" "y"      "z"
```

To delete a variable in R, we use the `rm` function. For example,

```
> x = 0  
> y = 1  
> z = 2  
> ls()  
[1] "movies" "x"      "y"      "z"
```

```
> rm(x)
```

```
> ls()  
[1] "movies" "y"      "z"
```

The nuclear option

We can remove everything in the workspace using:

```
> rm(list=ls())
```

2.3 Vectors

- Vectors are the most basic of all data structures, but are used in almost all R code.
- A R vector contains n values of the same type, where n can be zero.
- For example

```
> c(0, 1, 2, 3, 4, 5)
[1] 0 1 2 3 4 5
> (my_first_vec = c(0, 1, 2, 3, 4, 5))
[1] 0 1 2 3 4 5
> (my_second_vec = c("Male", "Female", "Male"))
[1] "Male"    "Female"  "Male"
```

In R, when we type:

```
> x = 5  
> y = "Fred"
```

We have actually created a vector of doubles and characters (of length one).

There are special functions in R to determine type of a variable:

```
> x = 5
```

```
> is.double(x)  
[1] TRUE
```

```
> is.character(x)  
[1] FALSE
```

```
> is.vector(x)  
[1] TRUE
```