Regression and Simulation Methods

Week 1: Welcome and Intro to R

Chris Oldnall, 10th October 2023

Welcome!
The Local Support Team

Chris Oldnall
MAC-MIGS 2021 Cohort
Causal Effect Estimation
in Population Genomics

Material + Weeks 1-5 Delivery

Clara Panchaud
MIGS 2021 Cohort
Statistical Ecology:
Animal Behaviour Models

Weeks 6-10 Delivery

Where to find us?

Online: Chat or Teams Channel

In Person: JCMB 5325 or
Around Bayes

When are tutorials?

Weeks 1-5: Tuesday 1-2pm

Weeks 6-10: TBD

What will be covered?

50% Theory
50% Programming (R)

A little bit more on me and teaching...

College of Arts, Humanities and Social Sciences (CAHSS)

CDCS
- Introduction to Python
- Introduction to R and RStudio
- Introduction to Statistics

College of Science and Engineering (CSE)

School of Mathematics
- Probability
- Statistics Year 2
- Statistical Methodology
- Incomplete Data Analysis
- Statistical Research Skills
- Linear Programming, Modelling and Solution

College of Medicine and Veterinary Medicine (CMVM)

School of Medicine
- Research and Evidence Based Medicine

The Usher Institute
- Introduction to Statistics
- Data Types and Data Structures in Python and R
- Introduction to Data Science

Masters Supervision
- Identifying and explaining key trends in accident and emergency activity data within Scotland, Hui Teoh [with Dr Kasia Banas]
- Advancing understanding of blood glucose control in type 1 diabetes, George McKay [with Prof Shareen Forbes and Dr Marta Vallejo]

Where to find everything...

SMSTC Website

- Course notes
- Course slides
- Exercises
- R Scripts from lectures

Teams Channel (or GitHub)

- R Notebook
- Tutorial slides
- 3 ⭐s and a 💥 wish
- Discussion spaces

github.com/chrisoldnall/SMSTC_RegAndSim_2023

What is 3 ⭐s and a 💥 wish?

- Effective learning tool for both you to reflect on what you actually learnt (and feel achieved) but also for us to get to understand what you still didn't get or wanted to know.

Format:

- *I learnt this…*
- *And this as well…*
- *Maybe I also developed this.*
- *What I do still wonder is this.*

- As a mini case-study in Week 10 we will use data collected from these to perform some analysis.

Course schedule...

Week	Topic	Lecture	Tutorial
1	Intro to R	N/A	Tuesday 10th October
2	Linear Models	N/A	Tuesday 17th October
3	Non Linear Regression	N/A	Tuesday 24th October
4	Likelihood	N/A	Tuesday 31st October
5	Inference with Likelihood	N/A	Tuesday 7th November
6	GLMs	Tuesday 14th November	TBD
7	Model Selection	Tuesday 21st November	TBD
8	Simulation Methods	Tuesday 28th November	TBD
9	Bootstrap Methods	Tuesday 5th December	TBD
10	Case Studies	Tuesday 12th December	TBD

An Introduction to R

You think you know it? Possibly not...

The Language

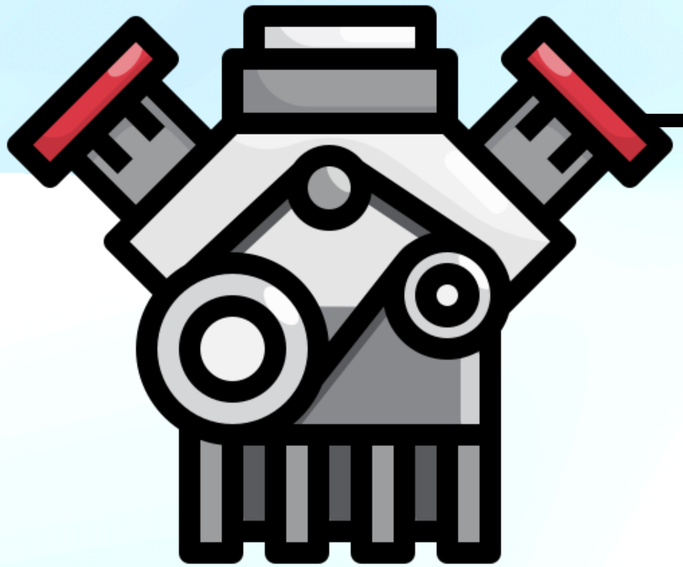
The Interpreter

RStudio®

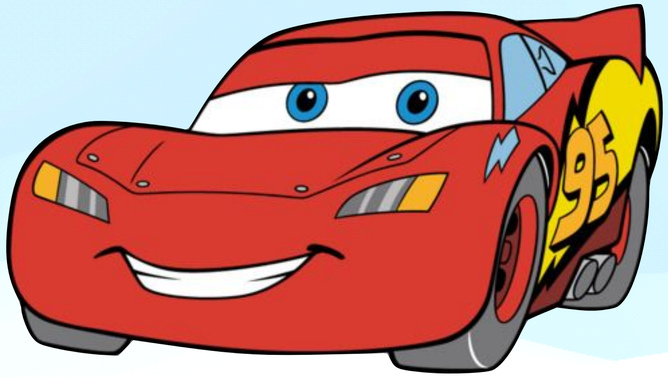
Visual Studio Code

Noteable™

The Engine

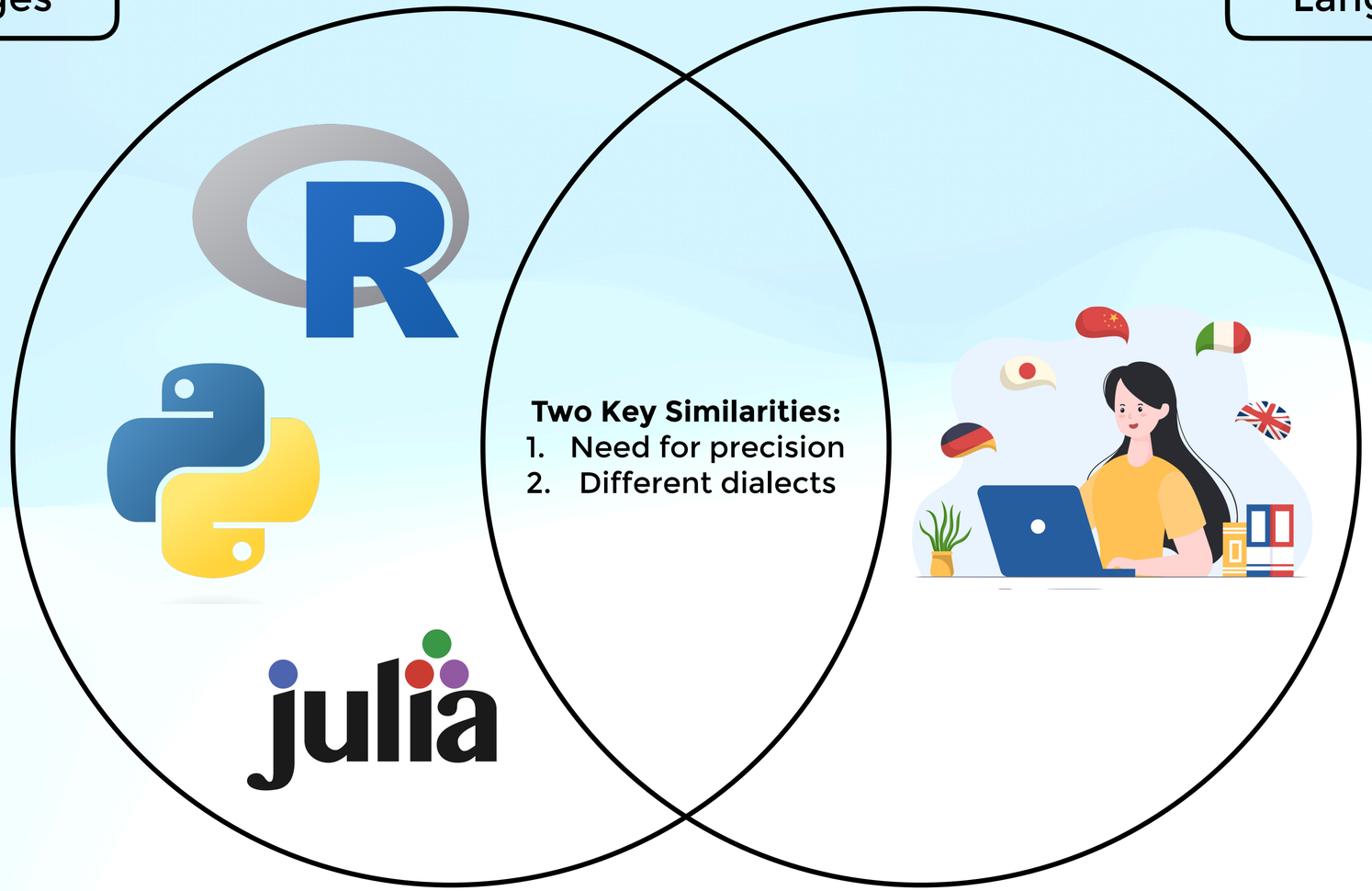


The Car



Programming Languages

Spoken Languages



$ and [ ]

Base R

or

Tidyverse

%>%

1.12 Exercises: Lecture 1

1–1. The dataframe `cats` in the `MASS` package contains measurements of the body weights and heart weights of a sample of cats. Plot and model the relationship between these two variables. Does it help to use a log scale? (These data are described and analysed in detail in the book by Venables & Ripley (2002a), so you might like to look there if you get stuck.)

1–2. Try different methods of clustering on the `iris` data. Look at the help file to identify what the options are.

1–3. Load the `crabs` data from the `MASS` package. Look at the information available (`?crabs`). The last five variables refer to different morphological measurements. Is there evidence from the morphological data alone of a division into two forms? The species are identified by the variables `sp` (Blue or Orange). Can we construct a rule to classify a crab of unknown species? How effective is this rule?

1–4. Plot the `rodent` data (log(Speed) against log(Mass)) and fit a simple linear regression, as we did in the lecture. (Remember that you might like to omit the porcupine.) Use the `predict` function to find the fitted values and standard errors. Then use the `polygon` function to shade areas corresponding to confidence intervals and prediction intervals for each value of log(Mass). (Hint: you might find the function `rev` useful when thinking about the input to `polygon`.)

1–5. Create a matrix which evaluates the function $10\sin(x^2+y^2)/(x^2+y^2)$ over the range (-10, 10) in x and y. Try plotting this with `persp`, `image` and `contour`.

1–6. The `sm` package contains a function `sm.density` which can be called in the form `sm.density(x, add = TRUE, display = "slice")` to add the contours of a two-dimensional density estimate to a plot of `x[,2]` versus `x[,1]`. Use the `aircraft` data from this package and select the six measurements for all aircraft in period 3. (The data can be made available by issuing the instruction `provide.data(aircraft)`.) Use `pairs` to produce a scatterplot matrix with two-dimensional density estimate contours added within each panel. (Hint: use the `panel` argument of `pairs`, which allows you to define your own function to do the plotting within each panel.)

1–7. Generate some data from a normal distribution and plot a histogram with a vertical scaling which makes it an estimate of the underlying probability density function (i.e. the area of the histogram is 1 - see the help information for `hist`). Add a curve which corresponds to a fitted normal distribution.

1–8. Understanding the distribution of extreme values is an important feature of some modelling problems, especially in environmental settings. Write some code which will repeatedly simulate from a normal distribution and then plot a histogram of the maxima from all the samples?

1–9. Look at the help information for `locator`. Now create some code which will draw a scatterplot of some simple regression data of your choice, with the fitted least squares regression line added, but then allow a position on the plot to be clicked. A new point is drawn on the scatterplot and the least squares regression line for the augmented data is also drawn.

These exercises are **not** 'An Intro to R'.

You should be able to feel confident... by Week 10... to do these.

Tutorial 1 - Saying hello to R (Solutions).

Christopher A Oldnall

Welcome to the Regression and Simulation methods module. This is the first script for the first tutorial. This is a prime opportunity to learn a lot more about data science and how to programme in R so please make full use of this.

Throughout we will use a package called 'tidyverse'. This is a 'superpackage' which contains lots of other packages with lots of functions. If you haven't installed this already then you need to go ahead and run the line below. You only need to install a package once.

```
# install.packages("tidyverse")
```

You will notice in your console that the section above appeared in a 'grey' area. This is because this is an R Markdown document. This is represented by the file extension .Rmd. Meanwhile there are also R Scripts (represented by .R extensions). Typically when writing code for publishable purposes or for software we have a series of R Scripts, however R markdown files are becomming more flavoursome when examining unqiue datasets since a neater overview can be given, it also allows for a neat PDF (or other document type) to be an output showing text, code and code output.

Having installed tidyverse earlier, we must still load in the package for our system to use it. We do that with the 'library' command - this is a built in R function. Let's do that now below, noting we do not need to use the quotes to load it in:

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

You may have additionally noticed that to create a chunk of code in an R Markdown document we must use two sets of 3 's with a set of curly braces following the first set of 3 with 'r' written in it. This is telling the interpreter that this is a block of R code.

Exercise 1

Now that you know the basics of how to install and load packages, have a go at writing a code block below to install and load the package called "MASS". This will be the data set we will work with for the rest of this tutorial.

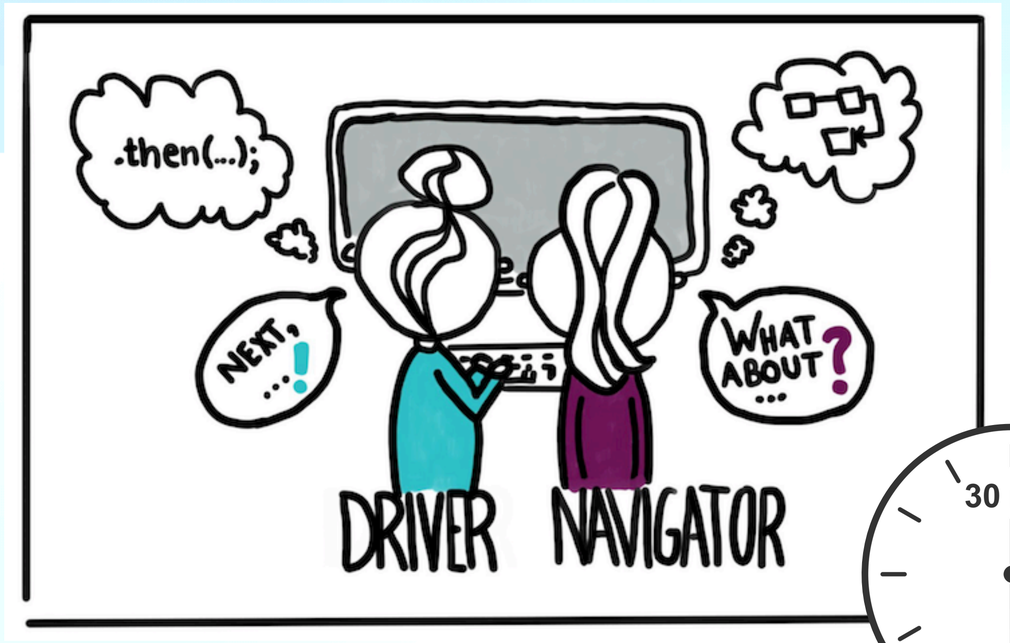
Your Answer:

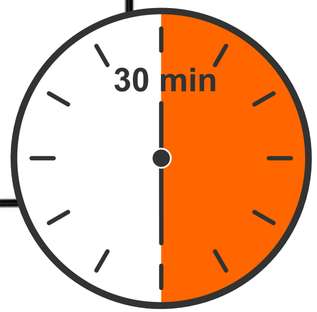
Instead each week there will be an R notebook with altered and more crafted exercises (both theoretical and programming).

(adapted from the exercises in the SMSTC course + other resources)

Rest of the tutorial...

- In pairs work on the first notebook. This is called paired programming! We will continue to use this approach even for theoretical problems.





Rest of the week...

From 1-5:

- Create a matrix which evaluates the function $10\sin(x^{2}+y^{2})/(x^{2} + y^{2}) over the range (-10, 10) in x and y.

From 1-7:

- Generate some data from a normal distribution and plot a histogram to give an estimate of the underlying probability density function.

From 1-8:

- Understanding the distribution of extreme values is an important feature of some modelling problems, especially in environmental settings. Write some code which will repeatedly simulate from a normal distribution and then plot a histogram of the maxima from all the samples.