

Programming with R/Advanced R

Dries Debeer & Benjamin Becker

18. + 19. March 2021

FDZ Spring Academy

Who are we?

Dries Debeer

Researcher at KU Leuven

scDIFtest, permimp, eatATA

dries.debeer@kuleuven.be

Benjamin Becker

Researcher at IQB (Statistics
Department)

eatGADS, eatDB, eatATA,
pisaRT

b.becker@iqb.hu-berlin.de

Who are you?

1. Specific interests/motivation for this workshop?
2. Previous knowledge and experience?
 - with R
 - with other statistic software
 - with other programming languages

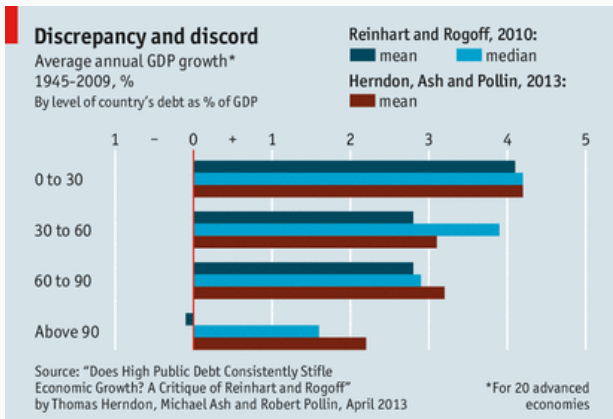
Motivation

- Being more efficient in your research
 - Save time and nerves
 - Avoid errors and bugs
 - High transfer effect to all projects
- Successful collaborations (with your future self?)
- Syntaxes as part of paper submissions

Two of your worst enemies

- Past Self
 - Is the biggest enemy in existence
 - Did not document anything
 - Uses a completely different style of writing code than yourself
- Future Self
 - Has the memory of a goldfish
 - Will have zero understanding for your current brilliance

Motivation



The Sunday Telegraph Sunday 17 May 2020

Coronavirus

Selling behind lockdown was a reliable buggsy mess, claim experts

Data that predicted 500,000 could die in UK unless extreme measures were taken are impossible to replicate, say scientific teams

Science

By Hannah Ireland and Peter Dinkley
The Covid-19 modelling that used little as its lockdown, missing the economy and leaving millions out of work has been criticised by experts.

Prof Nick Ferguson's Imperial College computer coding was derided as "totally unsuitable" by leading experts, who warned it was "something you wouldn't state your case".

The model, credited with forcing the Government to close bars and introduce a nationwide lockdown, is a "buggy mess, where loads more than a level of eight last pasts than a busy travel piece of programming", said David Forth, the co-founder of British data technology company Wavelength.

"In our commercial reality, we would not agree for developing code like this and any business that relied on it to produce software for sale would likely go bust."

The comments are likely to impact a row over whether the UK was right to go into lockdown, with conflicting claims suggesting people may have already acquired substantial herd immunity and Covid-19 may have hit Britain earlier than first thought.

Scientists have also been split on the likely rate of Covid-19 which has resulted in vastly different models.

To make any significant impact has been attributed to Imperial model, which placed the likely rate higher than others and predicted 500,000 in the UK could die without a lockdown.

It was said to have prompted a dramatic change in government policy causing businesses, schools and restaurants to be shut immediately in March. The Bank of England has predicted that the economy could take a year to return to normal, after its worst recession in more than 80 years.

The Imperial model works by using code to simulate transport links, population size, social networks and health care provision, to predict how coronavirus would spread. However, questions have emerged over whether the model is accurate, after researchers released its code, which in its original form was "dozens of lines" down speed over more than 10 years.

In its initial form the code was unimpressive, developers claimed, with some parts looking "like they were made

'In our commercial reality, we would fire anyone for developing code like this'

'Any business that relied on it to produce software for sale would likely go bust'

'It looks more like a bowl of angel hair pasta than a finely tuned piece of programming'

'The early 2000s were yet another confirmation that their modelling approach was flawed to the core'

Chinese translated from PerlScript, an old coding language, according to John Carroll, a US developer, who helped clean the code before it was published.

Yet, the problems appear to go much deeper than messy coding. Many have criticised that it is almost impossible to replicate the same results from the same data, using the same code.

Scientists from the University of Edinburgh said they got different results when they used different ma-

chine, and even in some cases using the same machines. "There appears to be a bug in either the compiler or in use of the network file. If we attempt two completely identical runs, only varying in that the second should use the network file produced by the first, the results are quite different", the Edinburgh researchers wrote on the GitHub website.

After a discussion with a GitHub developer, it was provided. It is said to be one of a number of

bugs discovered within the system. GitHub developers said that the model was "dozens of lines" of code, and "would give you different results depending on whether you had different code."

It has prompted questions from scientists, who say "models must be capable of passing the basic scientific test of providing the same results given the same initial set of parameters". Otherwise, there is simply no way of knowing whether they will be reliable.



explore predictions under different assumptions, and with different interventions, is incredibly powerful."

Like the Imperial code, a rival model by Prof Imperial College at the University of Oxford works on a so-called "SIR approach" in which the population is divided into those that are susceptible, infected and recovered. However, while Prof Gupta assumed that 0.1 per cent of infected people would die, Prof Ferguson worked on a 1 per cent. That led to a dramatic reversal in government policy from attempting to limit "herd immunity" to a full-on lockdown.

Concrete over Prof Ferguson's model have been raised, with Dr Fernando Sanchez, the VP of architecture at Wavelength, saying his track record did not inspire confidence. In the early 2000s, Prof Ferguson's models incorrectly predicted up to 100,000 road case deaths, according to Sanchez from both his and his own team's work.

"The facts from the early 2000s are not just another confirmation that their modelling approach was flawed to the core," says Dr Sanchez. "We don't know for sure if the same model/code was used, but we clearly see their methodology wasn't rigorous then and surely hasn't improved now."

A spokesman for Imperial's Covid-19 team said: "The Government has never relied on a single disease model to inform decisions making. As has been previously stated, decision making during lockdown was based on a consensus view of the scientific evidence, including several modelling studies by different academic groups."

Multiple groups using different models concluded that the pandemic would overwhelm the NHS and cause unacceptable high mortality in the absence of effective social distancing measures. Within the Imperial team, we use several models of differing levels of complexity, all of which produce consistent results. We are working with legitimate academic groups and technology companies to develop and further document the simulation code used in the Imperial model, and the partisan reviews of a few clearly discredited models."

"Epistemology is not a branch of computer modelling and the continuous arrival of lockdowns rely not on any mathematical model but on the scientific consensus that Covid-19 is highly transmissible with an infection fatality rate exceeding 0.1 per cent in the UK."

Concept of Technical Debt

- We write (messy) code for data cleaning/analyses
- We decide on data sets/models/graphs/tables/...
- We try to publish it, get a major revision
- We need to rerun some analyses
- We pay for our messy code

Goals of this workshop

- Better practical R skills
- Better theoretical understanding of R (and programming)
- Different framing: R as a programming language

R Objects (Recap)

Clean Code

Iteration

Understanding Functions

Writing Functions

Object Oriented Programming (S3)

Version Controlling (Git + Github)

R Ressources

- Advanced R Ed. 1 (<http://adv-r.had.co.nz/>)
- Advanced R Ed. 2 (<https://adv-r.hadley.nz/>)
- R Inferno (https://www.burns-stat.com/pages/Tutor/R_inferno.pdf)
- R Packages (<https://r-pkgs.org/>)
- Clean Code (<https://enos.itcollege.ee/~jpoial/oop/naited/Clean%20Code.pdf>)

Thank you for your attention!

Thank you for your attention!

Questions? Remarks?