STATISTICS DENMARK

Introduction to the software R

Anne Vinkel
Office of Methods and Analysis

Agenda

- A birds-eye view introduction to R
    - What is R?
    - The R community
    - The R workflow, using Rstudio
- R in official statistics
- Will talk minimally about concrete programming practises

What is R?

What is R?

- R is ~~a~~ the programming language and environment for statistical computing and for graphics
- Based on the object-oriented language S (1975)
- GNU project (100% free software)
- Managed by the R Foundation for Statistical Computing, Vienna, Austria.
    - Core team with a strong focus on correctness, stability and backwards compatibility
- Extensible by user-developed “packages”

Package development

- A “package” is a collection of functions and data made by other R users
- About eight packages supplied with base R
- More than 15.000 packages on the official archive CRAN
  - To be accepted on CRAN, must fulfill criteria relating to documentation and stability
- There’s a package for everything
  - but “buyer beware”

Why is R so popular?

"… R may be `riding the wave of data science' that is currently washing over industry, government and academia. R probably was simply there at the right place at the right time – offering a wide range of statistical and data handling functionality with a convenient and programmable interface for free when data science took off as a field. Additionally, the choice to publish R as an open source tool was an important cornerstone for its success."

Kowarik & van der Loo 2018

The R community

- Huge, active online community
  - StackExchange, Twitter
- Conferences, meet-ups (UseR!, EARL, uRoS,…)

R for official statistics

- R is beginning to be accepted by official statistical bodies
- An active community interested in using R for official statistics, keeping track of and developing packages for use in e.g.
  - Complex survey design
  - Editing and visual inspection of microdata
  - Imputation
  - Statistical disclosure control
  - Seasonal adjustment and forecasting

See the task view for official statistics: https://cran.r-project.org/web/views/OfficialStatistics.html

R for official statistics

Selected statistical offices using R (Templ & Todorov, 2016)

- Statistics Austria
- Statistics Netherlands
- National Statistical Office, UK
- National Statistical Institute, Romania
- United Nations Industrial Development Organisation

Advantages and disadvantages of using R for official statistics

Disadvantages:

- Relatively steep learning curve for non-programmers
- Not always good at handling large data sets
- Not always user-friendly
- Perceived risk of using software that is not backed by commercial support

Advantages and disadvantages of using R for official statistics

Advantages:

- Free
- Provides flexible way of reading, manipulating and writing data
- Availability of recent statistical methodology
- Easy to integrate presentation of results into workflow
- No commercial support, but an active community

The R workflow

What is R?

What is Rstudio?

What is RStudio?

- Integrated development environment for R
- Aggregates all convenient information and procedures into one single place
- Allows you to work in projects
- Manages your code with highlighting
- Gives extra functionality (produce reports in html, doc, pdf, webpages,…)
- Allows for integration with version control routines, such as Git.

A demonstration

- Write commands directly in the console
- Or write code in the editor and submit with Ctrl + Enter
- Data can be read in from any format (excel, txt, SAS, Oracle, …)
- Output can be saved to disk and used in reports
    - Or reports can be generated in Rstudio using Rmarkdown

Working code-centrically

- A single file of code, or a series of code files stored in the same place produce all your results
- Separates data from results
- Can be run non-interactively
- Allows for reproducibility
- Not unique to R

R in Statistics Denmark

The R workflow

To recap:

- **Three separate things:** The data, the results and the code producing the results
- The code:
  - Reads in data
  - Performs checks, transformations and analyses
  - Creates output
    - Possibly in the form of Rmarkdown reports
- Place the code, the data and the results in sensible places on your computer

In an official statistics environment

- Need to store data centrally and safely
  - Everybody works on the same data
  - Access to data is restricted to those who need it
- Need for disclosure control
  - Make sure personal data stays in the restricted zones
  - Check who is looking at individual data and why
- More computational power needed

The setup in Statistics Denmark

- Data lives in databases
- Analyses are
  - produced by code which is saved in a version control system
  - run on a secure server
- Output is stored in databases or routed through a secure channel allowing for routine control checks
- A whitelist of approved packages is curated by IT

Further good practises

- Introductory courses to all employees
- Collection of organisational know-how, e.g. a wiki
- Establishment of support team
- Encouragement for developing packages for internal use - and for publishing these for the use of others

Templ & Todorov 2016

A demonstration

- Data lives in a database, can be read in using SQL syntax or R syntax
- Analyses done with code files
- Data can be uploaded to database
- Reports generated using Rmarkdown, output through the dueslag

In summary

- R is a suitable tool (among others) for official statistics
- The use by official statistical bodies is likely to increase