Reproducible Computing

Continuing Education Course, JSM 2019 Biometrics Section 2019-07-27
8:30 am - 5 pm

Abstract

Success in statistics and data science is dependent on the development of both analytical and computational skills.

This workshop will cover:

- Recognizing the problems that reproducible research helps address.
- Identifying pain points in getting your analysis to be reproducible.
- The role of documentation, sharing, version control, automation, and organization in making your research more reproducible.
- Introducing tools to solve these problems, specifically R, RStudio, RMarkdown, git, GitHub, and make. - Strategies for scaling these tools and methods for larger more complex projects.

Workshop attendees will work through several exercises and get first-hand experience with using relevant tool-chains and techniques, including R/RStudio, literate programming with R Markdown, automation with make, and collaboration and version control with git/GitHub.

Schedule (Tentative)

Time	Activity
08:30 - 09:00	Welcome
09:00 - 09:40	Literate programming
09:40 - 10:15	Naming & Organization
10:15 - 10:30	*Coffee break*
10:30 - 12:30	Version control with Git and GitHub
12:30 - 14:00	*Lunch break*
14:00 - 14:30	Scaling reproducible projects
14:30 - 15:15	Introduction to make
15:15 - 15:30	*Coffee break*
15:30 - 16:30	make in action
16:30 - 17:00	Parting remarks

Welcome, literate programming, and naming

- Recognize the problems that reproducible research helps address and identify pain points in getting your analysis to be reproducible.
- The role of documentation, sharing, automation, and organization in making your research more reproducible.
- Literate programming with R Markdown
    - Introduce the data: World Cup!
    - Hands on activity: Updating an analysis when the source data changes
- Naming best practices

Organization and version control with Git and GitHub

- Project organization
    - File and folder organization for projects
    - Naming conventions
- What is Git and version control?
- Git in GitHub
    - Define git vocabulary (commit, fork, pull request, repository, commit message).
    - Demonstrate ability to navigate through a Github repository main page.
    - Define the difference between a directory and a repository.
    - Create a repository on GitHub.
    - Demonstrate ability to commit changes to text files with a commit message.
    - Evaluate repository History.
    - Create a pull request to someone else's repository.
- Git in RStudio
    - Define git vocabulary (push, fork, local repository, remote repository)
    - Fork remote repository from Github into RStudio.
    - Navigate through the basics of using git in RStudio.
    - Push local repository from RStudio to Github.
    - Demonstrate the ability to host code from RStudio to Github.
- usethis Package

Scaling reproducible projects + Make

- Practical example - Scottish lip cancer
    - Reproducible R Markdown document with "full Bayesian analysis" including data munging, EDA, model fitting and analysis.
- Caching as a solution to scaling
    - Build your own cache: Saving your own results with `save()` vs. `saveRDS()`
    - R Markdown caching: `cache = TRUE`
- Using make to automate and scale
    - Introduce make
    - Review make syntax
    - Introduce hands on exerciseß

Computing requirements

An R + RStudio computing environment will be provided for all students via RStudio Cloud. All that will be needed the day of the event is a laptop and a Google Account that can be used for authentication.

Instructor

Colin Rundel (http://www2.stat.duke.edu/~cr173/) - University of Edinburgh, Duke University

Colin recently started as a Lecturer in the School of Mathematics at the University of Edinburgh. Prior to this position he was an a Assistant Professor of the Practice in the Department of Statistical Science at Duke University. He has developed and taught a number of Statistical Computing courses for undergraduate, master's and Ph.D. levels students. His pedagogical and research interests are in the area of statistical computing, data science, and spatial statistics.

Acknowledgements

- Data Carpentry (https://github.com/datacarpentry)'s modules on naming, organization, and version control.
- Happy git with R (http://happygitwithr.com/) by Jenny Bryan (http://github.com/jennybc).

License