Reproducible Research in Statistics

Jessica Minnier

June 16, 2016

Why reproducibility?

Claerbout's Principle

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generate the figures.
(Buckheit and Donoho 1995; De Leeuw 2001)
(Jon F. Claerbout is the Cecil Green Professor Emeritus of Geophysics at Stanford University. He was one of the first scientists to emphasize that computational methods threaten the reproducibility of research unless open access is provided to both the data and the software underlying a publication. (Claerbout and Karrenbach 1992, Wikipedia)

Schwab, Karrenbach, Claerbout:

It takes some effort to organize your research to be reproducible. We found that although the effort seems to be directed to helping other people stand up on your shoulders, the principal beneficiary is

NIH Requirements

NIH requirements (beginning Jan 2016)

“Enhancing Reproducibility through Rigor and Transparency”
`http://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-103.html`

1. *Scientific Premise*
    - “describe the general strengths and weaknesses of the prior research being cited by the investigator as crucial to support the application.”
    - experimental design/power of prior studies used for hypothesis generation, weaknesses include different populations/species, unblinded, not adjusting for confounders
2. *Rigorous Experimental Design*
3. *Consideration of Sex and Other Relevant Biological Variables*
    - “sex is a biological variable that is frequently ignored in animal study designs and analyses”
4. *Authentication of Key Biological and/or Chemical Resources*
5. *Implementation*

Journals unite to encourage reproducibility

NIH Principles and Guidelines for Reporting Preclinical Research

http://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research

NIH held a joint workshop in June 2014 with the Nature Publishing Group and Science on the issue of reproducibility and rigor of research findings

A video/slide presentation about this topic and how it applies to grant applications and peer review can be found here:
http://grants.nih.gov/grants/policy/rigor/NIH_Policy_Rigor_For_Reviewers/presentation.html

NIH Principles and Guidelines for Reporting Preclinical Research

- aim of facilitating the interpretation and repetition of experiments as they have been conducted in the published study
- journal should include policies for statistical reporting in information to authors

Literate Programming

Literate Programming

Literate programming is an approach to programming introduced by Donald Knuth in which a program is given as an explanation of the program logic in a natural language, such as English, interspersed with snippets of macros and traditional source code, from which a compilable source code can be generated. (Knuth 1984)

Examples: Sweave, knitr (for R); SASweave, Statrep (for SAS); StatWeave (for STATA)

This is knitr:

```r
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
```

Version Control

Version Control

"Version control systems (VCS), which have long been used to maintain code repositories in the software industry, are now finding new applications in science. One such open source VCS, Git, provides a lightweight yet robust framework that is ideal for managing the full suite of research outputs such as datasets, statistical code, figures, lab notes, and manuscripts. For individual researchers, Git provides a powerful way to track and compare versions, retrace errors, explore new approaches in a structured manner, while maintaining a full audit trail. For larger collaborative efforts, Git and Git hosting services make it possible for everyone to work asynchronously and merge their contributions at any time, all the while maintaining a complete authorship trail."
(Ram 2013)

Why Use Version Control?

Have you ever:

- Made a change to code, realised it was a mistake and wanted

Reproducibility in Practice

Peng (2009) in Biostatistics: Reproducible research and Biostatistics (the journal)

Authors should submit the following:

1. A “main” script which directs the overall analysis. This script may load data, other software, and call the necessary functions for conducting the analysis described in the article.
2. Other required code files, presumably called from the “main” script file.
3. External data or auxiliary files containing the analytic data sets or other required information.
4. A “target” file (or files) containing the results which are to be reproduced. Such a file could consist of an ASCII text file containing numerical results or a PDF file containing a figure. This will aid in the comparison of computed results with published results.

Although not required, authors are encouraged to use literate programming tools [. . .]

Goals for BSR (for CCSG)

Aim to improve transparency and reproducibility in the research and analyses performed by BSR members.

- Store all data and code in a central location we can easily locate files related to current and archived projects
  - We implement file naming conventions for all new projects
  - All reports, including tables or figures prepared for manuscripts, are saved in a final reports folder along with the source code to produce them.
- Generate reproducible reports for final results with knitr (for R code) and generate well-documented and easily reproducible results from SAS code.
- Version tracking system (git or another) whenever possible.
- Post commonly used code, R functions, and SAS macros to the online repository GitHub (`https://github.com/ohsu-knight-cancer-biostatistics`).
- When a BSR member is an author of a publication, any code used to produce published results is tested by another BSR member and either submitted with the publication or the

Resources

Recommended Books

Stodden, Victoria, Friedrich Leisch, and Roger D. Peng, eds. Implementing reproducible research. CRC Press, 2014.

Gandrud, Christopher. Reproducible Research with R and R Studio. CRC Press, 2013.

Xie, Yihui. Dynamic Documents with R and knitr. Vol. 29. CRC Press, 2013.

Resources (classes)

Karl Broman's class "Tools for Reproducible Research" at UWisconsin-Madison `http://kbroman.org/Tools4RR/`

"Reproducible Research" by Johns Hopkins on Coursera (Peng, Leek, Caffo) `https://www.coursera.org/learn/reproducible-research`

References and Resources (websites)

ROpenSci's "Reproducibility in Science" guide: `http://ropensci.github.io/reproducibility-guide/` including the reproducibility checklist `http://ropensci.github`