Efficient statistical consulting using R
Workflow for data analysis projects

Peter Baker

School of Public Health, University of Queensland, Herston, Australia. p.baker1@uq.edu.au

Keywords: Data analysis, DRY (don't repeat yourself) workflow, version control, make, project management

Abstract

Many researchers and statistical consultants are drowning in data. A generation ago, computer processing power and storage were limited and so considerable time was spent formulating strategies to efficiently manipulate data and focus on the analysis of relatively small data sets or subsets of larger ones. Today, the process of managing larger data analysis projects is throwing up considerable challenges. Given today's powerful computing environments it would seem that things should have become easier but it appears that due to larger data sets and competing time demands many data analysts face problems with organising their workflow. The ideas presented in this tutorial follow Long [5] which provides a useful guide to managing workflow for data analysis in large projects using *STATA* on Windows. However, Long's approach concentrates on manual methods. Efficient computing solutions are available using programming tools like *R* functions, *make*[6] or version control[4] but are not widely understood.

This tutorial provides a hands-on introduction to strategies for the workflow of research data management and data analysis. I will demonstrate a systematic way to employ computing tools such as *R*, *git* and *make* to assist in this process. These tools will be incorporated into the hands on exercises.

R is ideal for automating repetitive tasks since, in addition to thousands of built-in functions, it is easy to write your own. Such functions can be used for diverse tasks like setting up similar directory structures and R syntax files for similar projects, checking data against code books or meta data, repeating analyses on similar variables and extracting results for reproducible reporting to *HTML*, *Word* or *pdf* via **Rmarkdown** and *pandoc*. In addition, *Make* keeps track of the dependencies in the process and so allows the workflow to be broken down into smaller chunks like reading, checking, transforming and analysing data. Only the updated steps in this workflow are re-run. Also, *git* safely allows analysts to keep previous versions of *R* scripts or **Rmarkdown** documents without having multiple versions of files cluttering up the work area.

Finally, while specific strategies are provided, I will outline how these may be easily modified to suit individual taste and circumstances.

Aims and Learning Objectives

After completion of this tutorial, participants will be able to:

1. identify systematic strategies to plan a data analysis project,
2. employ standard *R* scripts, *git* and *make* (or *makepp*) for a simple data analysis project, and
3. be aware of approaches for more complex data analysis projects by extending the strategies for a simple project.

Target Audience

This tutorial should be of interest to statistical consultants in government and industry, researchers and students managing and conducting data analysis of research data and *R* practitioners in general.

Background knowledge and requirements

Background knowledge

Basic use of *R* for data manipulation will be assumed and briefly revised. However, simple *R* functions for assisting project set up, data checking, data cleaning and reporting will be provided as will `Makefiles` and recipes for using *git*.

Some familiarity with the `apply` family of functions, the package **plyr** and the pipe operator (`%>%`) from the **magrittr** package would prove advantageous as would use of *make* and *git* but this will not be assumed. Experience in writing and/or modifying simple *R* functions would also be advantageous.

Laptop setup

Participants should have a recent version of *R* installed on their laptop. They should also install recent version of *RStudio*, *Emacs Speaks Statistics* or favorite editor. However, when not using *RStudio*, they should have working copies of **Rmarkdown** and **knitr**. Windows users should install *Rtools* in order to install *gnu make* and users of other operating systems will need to install appropriate development tools.

In addition, several *R* packages should be installed including **plyr** and **magrittr**.

Appropriate instructions will be provided at least one month prior to the tutorial.

Why the tutorial topic is of interest

Researchers and statistical consultants often spend a great deal of time managing data and performing data analysis for a number of research projects. Often these projects have much in common and an efficient workflow strategy can greatly aid the process. This tutorial will provide an introduction to this important area of practical data analysis and links to further topics.

Tutorial Outline

Each session will be about half an hour and consist of lectures interspersed with hands on exercises. Overall, approximately half the time will be hands on using data supplied.

Table 1: Suggested Morning Schedule: Efficient statistical consulting using R tutorial.

Start	End	Session
9:00	9:30	Introduction and overview of data analysis workflow: planning, organising, documentation, execution, freezing files
9:30	10:00	Using code books, checking and cleaning data in R
10:00	10:30	R functions for automating data handling and analysis
10:30	11:00	Coffee and tea break
11:00	11:20	make: reproducing steps in the process
11:20	11:40	git: version control - why and how
11:40	12:10	Putting it all together: a simple data analysis project
12:10	12:30	Summary and close

Biography

Dr Peter Baker is a Statistical Consultant and Senior Lecturer, Epidemiology and Biostatistics, School of Public Health, University of Queensland, Herston, QLD. Australia. http://researchers.uq.edu.au/researcher/2181. Prior to 2007, Peter worked as a statistical consultant and researcher at CSIRO for 20 years. He has been using R since the late 1990s both as his main tool in statistical consulting and for teaching introductory and advanced Biostatistics courses in the School of Public Health. He has authored or coauthored several R packages[3, 2, 1] and regularly runs one or two day courses *Introduction to R* for the UQ School of Public Health and has run specialised R courses for University of Queensland School of Agriculture and Food Sciences and also the Australian Centre for Air Pollution Research.

References

[1] Baker, P. (2010). *polySegratio: Simulate and test marker dosage for dominant markers in autopolyploids*. R package version 0.2-3.

[2] Baker, P. J. (2012). *polySegratioMM Bayesian mixture models for marker dosage in autopolyploids*. R package version 0.6-2.

[3] Barnett, A. G., P. Baker, and A. J. Dobson (2012, June). Analysing seasonal data. *The R Journal 4*(1).

[4] Loeliger, J. and M. McCullough (2012, August). *Version Control with Git: Powerful tools and techniques for collaborative software development* (2nd ed.). O'Reilly Media, Inc.

[5] Long, J. S. (2009). *The Workflow of Data Analysis Using Stata*. StataCorp LP.

[6] Mecklenburg, R. (2004). *Managing Projects with GNU Make* (3rd ed.). O'Reilly Media, Inc.