Literate Statistical Programming
An Introduction using R and RStudio

Andres Martinez & Michael Clark

Center for Social Research

May 14, 2015

Literate Statistical Programming

The term has many aliases, including:

- Reproducible research (RR)
- Replicable science (RS)
- Reproducible (data) analysis (RDA)
- Dynamic data analysis
- Dynamic report generation
- Literate (data/statistical) analysis

LSP and RDA are here used interchangeably. In my view these are different from RR/RS.

Reproducible Data Analysis

The ultimate standard for strengthening scientific evidence is replication.

RDA requires:

- Data
- Code
- Clear documentation (of data and code)
- Standard means of distribution

Literate Statistical Programming

- Think of a report (e.g., journal article, blog, research paper/memo) as a single stream of human-readable text and machine-readable code
- This is not quite the same as having a commented script file in, say, R or Stata and certainly not the same as having scripts and report files living separate lives

Literate Statistical Programming

May be conceived as a stream of code chunks and human-readable text chunks:

- Code chunks:
  - load and prepare data
  - compute a result
  - create a table or plot
- Human-readable text chunks:
  - Describe the data
  - Explain analysis
  - Present a result

Literate Statistical Programming

The programs or streams of text and code can then be:

- 'Weaved' to produce human-readable documents
- 'Tangled' to produce machine-readable documents

The basic idea is to combine

- a machine-readable programming language
- a human-readable documentation

Tools

There is a growing number of (awareness about?) open source tools that facilitate LSP. We will discuss:

- **R**: software programming language / environment for statistical computing / graphics
- **R Studio**: integrated development environment (IDE) for R
- **Sweave**: R functionality
- **knitR**: R functionality/package (subsumes Sweave)
- **LaTeX**: document preparation system and markup language
- **HTML**: standard markup language for web pages (HyperText Markup Language)
- **Markdown**: plain text formatting syntax easily convertible to HTML
- **pandoc**: (universal?) document converter

Tools

Sweave:

- Original system in R designed to do RDA
- Focus mainly on LaTeX (which some find difficult to learn)
- Lacks features like caching, multiple plots per chunk, support for multiple programming languages
- Development mostly stalled

Tools

knitR:

- More recent (package)
- Inspired by Sweave, builds on its functionality
- Possible to use with other programming languages
- Supports a variety of documentation languages (LaTeX, Markdown, HTML)
- Frequently updated, actively developed (young developer)

See `http://yihui.name/knitr/`.

Examples

A couple to get you started:

- A web page using Markdown
- A report using LaTeX

Setup in RStudio:

- May want to create a new project (say, using a new directory in NetFile)
- Set weaving to be done by knitR (Tools, Options, Sweave)
- Install knitr (if not already): install.packages("knitr")

What is Markdown?

- "A plain text formatting syntax designed so that it can optionally be converted to HTML using a tool by the same name" (Wikipedia)
- From R Studio:
  - a simple markup language designed to facilitate authoring web content easy
  - a format that enables easy authoring of reproducible web reports from R
  - rather than writing HTML and CSS code, Markdown enables the use of a syntax much more like plain-text email
  - combines the core syntax of Markdown (an easy-to-write plain text format for web content) with embedded R code chunks that are run so their output can be included in the final document

Markdown in RStudio

- RStudio greatly facilitates the combination of R with Markdown (R is effectively used as a Markdown implementation)
- Combination achieved via the inclusion of R code chunks within a R Markdown file (.Rmd or .rmd), as opposed to a Markdown file (.md)
- The process involves 2 major steps:
  - Weaving the R Markdown file (.Rmd) into a plain Markdown file (.md) — accomplished by the package knitR
  - Converting the markdown files into an HTML document — accomplished by the package markdown

See `http://www.rstudio.com/ide/docs/r_markdown`.

LaTeX in RStudio

- Much of what we have said about Markdown applies
- Combination (of R and TeX) achieved via the inclusion of R code chunks within a R NoWeb file (.Rnw or .rnw), as opposed to TeX file (.tex)
- The process involves 2 major steps:
  - Weaving the R NoWeb file (.Rnw) into a plain TeX file (.tex) — accomplished by knitR
  - Converting the .tex file into, say, a .pdf file.
- R code chunks opened and closed differently

Presentation, examples, and some useful resources

Presentation and related materials:
`http://www3.nd.edu/~amarti38/RDA.zip`

A nice example of what's possible:
`https://micl.shinyapps.io/texEx/texEx.Rmd`

Useful resources:

- Help from within RStudio and from RStudio.com
- http://yihui.name/knitr/
- http://stackexchange.com/