R and Reproducible Research

Harrison Dekker
Library Data Lab
UC Berkeley
hdekker@berkeley.edu

Why promote reproducible research?

1. Allow authors to reproduce the results and figures in their research publications.
2. Aid verification of results by other researchers
3. Allow researchers to learn from and/or build on the work of others.
4. Build community.

What does R offer?

```
## Annette Dobson (1990) "An Introduction to Generalized
## Linear Models".
## Page 9: Plant Weight Data.

ctl <- c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
trt <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
group <- gl(2, 10, 20, labels = c("Ctl","Trt"))
weight <- c(ctl, trt)
lm.D9 <- lm(weight ~ group)
lm.D90 <- lm(weight ~ group - 1) # omitting intercept

anova(lm.D9)
summary(lm.D90)

# Source: http://stat.ethz.ch/R-manual/R-patched/library/stats/html/lm.html
```

Getting started

- Reproducible research “task view” on CRAN
    - Collection of packages to facilitate literate programming.
    - http://cran.r-project.org/web/views/ReproducibleResearch.html
- RStudio
    - Designed with rr in mind
    - UI facilitates literate programming (knitr, sweave), version control, file management.
    - Free, rpubs.com web publishing platform.

Literate programming

~~~~
# A Minimal Example for Markdown
This is a minimal example of using **knitr** to produce an
_HTML_ page from _Markdown_.

## R code chunks

```{r setup}
# set global chunk options: images will be 7x5 inches
opts_chunk$set(fig.width=7, fig.height=5)
```


Now we write some code chunks in this markdown file:

```{r computing}
x <- 1+1 # a simple calculator
set.seed(123)
rnorm(5)  # boring random numbers
```
~~~~

Beyond literate programming

- R data access tools
- rOpenSci
- R metadata tools

Dataset tools

- http://www.asdfree.com and https://www.github.com/ajdamico/usgsd
- R code packages for working numerous public data sets from major government survey programs.
- Well-documented code and great examples of using external databases to speed up R.
- Examples: American Community Survey, General Social Survey, Consumer Expenditure Survey, etc.

Data repository API wrappers

```
library(rnoaa)
out <- noaa(dataset = "NORMAL_DLY",
            station = "GHCND:USW00014895",
            datatype = "dly-tmax-normal",
            year = 2010, month = 4)
head(noaa_data(out))

                     date       dataType           station value atts
1 2010-04-01T00:00:00.000 DLY-TMAX-NORMAL GHCND:USW00014895   536    S
2 2010-04-02T00:00:00.000 DLY-TMAX-NORMAL GHCND:USW00014895   540    S
3 2010-04-03T00:00:00.000 DLY-TMAX-NORMAL GHCND:USW00014895   545    S
4 2010-04-04T00:00:00.000 DLY-TMAX-NORMAL GHCND:USW00014895   549    S
5 2010-04-05T00:00:00.000 DLY-TMAX-NORMAL GHCND:USW00014895   554    S
6 2010-04-06T00:00:00.000 DLY-TMAX-NORMAL GHCND:USW00014895   558    S
```

# *Source:* http://ropensci.org/packages/rnoaa.html

R data API wrappers exist for many scientific data repositories across disciplines:

- Ecological and evolutionary biology
- Climate
- Genomics
- Earth science
- Economics
- Links to packages at: http://ropensci.org/blog/2013/09/11/taskview/

rOpenSci

- small group of ecologist/R-developers working on a unified framework for connecting researchers with data.
- Work includes api’s for accessing scientific literature
- http://ropensci.org

R metadata tools

- Descriptive metadata is essential for data re-use, but is often given low priority in the research publication process.
- Data Documentation Initiative (DDI) is a well-established metadata specification for the behavioral and social sciences.
- r2ddi is an R package in development to help generate DDI metadata from data files of various formats.
- https://github.com/mhebing/r2ddi