Statistical Methods for Causal Inference in Observational and Randomized Studies

Mark J. van der Laan[1], Maya L. Petersen[1], Sherri Rose[2]

[1]University of California, Berkeley School of Public Health

[2]Johns Hopkins Bloomberg School of Public Health

laan@berkeley.edu · mayaliv@berkeley.edu · srose@jhsph.edu

stat.berkeley.edu/∼laan/

works.bepress.com/maya_petersen/

drsherrirose.com

targetedlearningbook.com

September 26, 2011

DAY ONE: OPTIONAL LAB

Introduction to R

R Resources

Many free resources on R online. A few include:

- The Comprehensive R Archive Network: cran.r-project.org
  - cran.r-project.org/manuals.html
  - cran.r-project.org/faqs.html
- R Short Courses:
  - https://mywebspace.wisc.edu/ratkovic/R_Short_Course/2010_R_Short_Course.html
  - https://sites.google.com/site/undergraduateguidetor/
  - http://scc.stat.ucla.edu/mini-courses
  - http://www.biostat.jhsph.edu/~ajaffe/rseminar.html
- Quick R (http://www.statmethods.net/), R tips for people who program in SAS, SPSS, and STATA

The slides for this lecture have been adapted from a short course **"Statistics with R for Biologists"** given by James H. Bullard[1], Kasper Daniel Hansen[2], and Margaret Taub[2] in July 2008.

http://wiki.biostat.berkeley.edu/~bullard/courses/T-berkeley-08/

[1] currently at Pacific Biosciences

[2] currently at Johns Hopkins BSPH

Background

- R is an open source version of the S language.
- R was written and released initially in 1995 by Robert Gentleman and Ross Ihaka.
- S was developed in 1976 by John Chambers at Bell Labs.

Background

- R has existing functions and tools, but also allows the user to implement and code new functions.
- New packages are added to CRAN frequently, and R is updated twice a year.

Installing R: cran.r-project.org

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for MacOS X
- Download R for Windows

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2011-07-08): R-2.13.1.tar.gz (read what's new in the latest version).
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing

Programming Environments

Command-line interface

- Can run simple code by typing directly into the command-line:
  e.g., `?glm`
  e.g., `3657/3`
- For most analyses, you will want a way to save your code so that you can rerun the code and also use multi-line code, e.g., `for` loops.

Programming Environments

- Xcode
- ESS (emacs speaks statistics)
- TextMate
- Notepad++

R Help / Commenting

- `help(glm)` or `?glm`:
  help for the `glm` function
- `library(help="stats")` or `help(package="stats")`:
  help for the `stats` package
- # is the comment symbol

Example Data Sets in R

- `data()`: This command lists all available data sets.
- Useful for running examples given in help files and testing code.

```
> data(Titanic)
> require(graphics)
> mosaicplot(Titanic, main = "Survival on the Titanic")
```

Examples: Vectors

```
> v1 <- 1:5
> v2 <- runif(5)
> v3 <- sample(c("A", "B", "C"), size=5, replace = TRUE)
> v4 <- v3 %in% c("A", "B")
```

Examples: Vector-Related Functions

```
> seq(1, 10, by = 2)
> seq(0, 10, along.with = c(1:51))
> seq(0, 10, length.out = 51)
> rep(1:5, 5)
> rep(1:5, 1:5)
> rep(1:5, each = 2)
> paste("chr", 1:23)
> paste(LETTERS[1:5], rep(1:5, each = 5), sep = "")
```

NA, -Inf, Inf, NaN

- `NA`: missing data
- `-Inf`/`Inf`: infinity
- `NaN`: Not a number

```
> sum(c(2, 3, NA, 6))
> 5/0
> 0/0
> -5/0
> c(2, 3, NA, 0)/c(3, 0, 5, 0)
> 0 * Inf
```

Examples: Matrices

```
> m1 <- matrix(1:6, nrow = 3, ncol = 2)
> m2 <- matrix(1:6, nrow = 3, ncol = 2, byrow = TRUE)
```

Reading and Saving Files

Reading

- `read.table`
- `scan`

Saving

- `write.table`
- `save`

TOMORROW

Using `SuperLearner` and `tmle` in R.