The R Statistical Computing Environment: The Basics and Beyond

The R statistical programming language and computing environment has become the defacto standard for writing statistical software among statisticians and has made substantial inroads in the social sciences. R is a free, open-source implementation of the S language, and is available for Windows, Mac OS X, and Unix/Linux systems. There is also a commercial implementation of S called S-PLUS, but it has been eclipsed by R.

The basic R system is developed and maintained by the R Core group, comprising 20 members, many of them eminent in the field of statistical computing. The R Project for Statistical Computing is a project of the R Foundation, whose membership includes the R Core group and several other individuals.

A statistical package, such as SPSS or SAS, is primarily oriented toward combining instructions with rectangular case-by-variable datasets to produce (often voluminous) printouts. Such packages make routine data analysis relatively easy, but they make it relatively difficult to do things that are innovative or non-standard, or to add to the built-in capabilities of the package. In contrast, a good statistical computing environment also makes routine data analysis easy, but it additionally supports convenient programming; this means that users can extend the already impressive facilities of R. Statisticians and others have taken advantage of the extensibility of R to contribute nearly 4000 freely available “packages” of documented R programs and data to CRAN (the Comprehensive R Archive Network) <http://cran.r-project.org/web/packages/index.html> and many others to the Bioconductor package archive <http://www.bioconductor.org/>. As well, R is especially capable in the area of statistical graphics, reflecting the origin of S at Bell Labs, a centre of graphical innovation.

The first day of this workshop is meant to provide a basic overview of and introduction to R, including to statistical modeling in R – in effect, using R as a statistical package. The following three days pick up where the basic lectures leave off, and are intended to provide the background required to use R seriously for data analysis and presentation, including an introduction to R programming and to the design of custom statistical graphs, unlocking the power in the R statistical programming environment. Participants should bring their laptops to the workshop and should install R and RStudio in advance (see the instructions on the workshop website).

An outline of the workshop follows (with chapter references to Fox and Weisberg, An R Companion to Applied Regression, Second Edition):

Day 1. Getting started with R (Ch. 1); statistical models in R (Ch. 4, 5, & appendices)

Day 2. Data in R (Ch. 2); the basics of R programming (Ch. 8, Sec. 8.1-8.4)
Day 3. R programming, beyond the basics (Ch. 8, Sec. 8.5-8.10)

Day 4. R graphics (Ch. 7); building R packages

Course Web Site

Materials for the course will be deposited at
<http://socserv.mcmaster.ca/jfox/Courses/R-course-Berkeley/index.html>,
abbreviation <http://tinyurl.com/R-course-Berkeley>, which also has active links to
many of the resources described in this syllabus.

Acquiring R

More detailed instructions are on the workshop website at
<http://socserv.mcmaster.ca/jfox/Courses/R-course-Berkeley/R-install-
instructions.html>

Windows Users

You can download the R Windows installer from CRAN <http://cran.r-
project.org/bin/windows/base/> , or better from a CRAN mirror site near you
<http://cran.r-project.org/mirrors.html>; then double-click on the installer to install R
as you would any Windows software. You can subsequently download and install only
those packages that you want over the Internet from CRAN, via the Packages → Install
packages from CRAN menu in the RGui console.

Mac Users

A universal binary for Mac OS X 10.5 and higher is available from CRAN
<http://cran.r-project.org/bin/macosx/> , or better from a CRAN mirror site near you
<http://cran.r-project.org/mirrors.html>. Double-click on the downloaded file to install
R. You can then download and install packages over the Internet via the Packages &
Data → Packages Installer menu in R.app or R64.app console.

Linux/Unix Users

Precompiled binaries for popular Linux systems are available from CRAN
<http://cran.r-project.org/bin/linux/> (or better from a CRAN mirror site near you
<http://cran.r-project.org/mirrors.html>), or users can compile R from source. See
CRAN for details <http://cran.r-project.org/>.

RStudio

RStudio <http://www.rstudio.org/> is a free, open-source interactive development
environment (IDE) for R that installs easily on Windows, Mac OS X, and Linux systems
and works well “out of the box.” Though still under active development, RStudio in my
opinion provides a better interface to R than the standard Windows and Mac OS X interfaces. Among the many services that it provides RStudio includes a package manner that will allow you to install packages conveniently.

**Installing the car Package**

For this course, you'll want to install the `car` package associated with the *R Companion to Applied Regression*; use the command

```r
install.packages("car")
```

or install via the menus in the Windows or Mac OS X versions of R or via the packages tab in RStudio.

**Selected Bibliography**

Publishers of statistical texts have been producing a steady stream of books on R. Of particular note is Springer's *Use R!* series of brief paperbacks on various R-related topics [http://www.springer.com/series/6991], several titles of which I've listed below. Recently, Chapman and Hall, which has published a number of books on R, has also announced *The R Series.*

**Basic Texts**


Alternatively (or additionally), more advanced students may wish to use W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S* as a principal source. Bill Venables is a member of the R Foundation, and Brian Ripley is a member of the R Core group.

**Manuals**

R is distributed with a set of manuals, which are also available at the CRAN web site [http://cran.r-project.org/manuals.html].

A manual for S-PLUS Trellis Graphics (also useful for the lattice package in R) is also available on the web at [http://cm.bell-labs.com/cm/ms/departments/sia/doc/trellis.user.pdf].

**Programming in R**


J. M. Chambers, *Software for Data Analysis: Programming with R*. New York: Springer, 2008. Chambers’s newest book ranges quite widely, and emphasizes a deep understanding of the R language, along with object-oriented programming, and links between R and other software. Some topics are unusual, such as processing text data in R.


W. N. Venables and B. D. Ripley, *S Programming*. New York: Springer, 2000. A companion volume to *Modern Applied Statistics with S*, and at the time of its publication the definitive treatment of writing software in the various versions of S-PLUS and R; now somewhat dated, particularly with respect to R. Brian Ripley is a member of the R Core group of developers, and Bill Venables is a member of the R Foundation.

**Statistical Computing in R**

The following three books treat traditional topics in statistical computing, such as optimization, simulation, probability calculations, and computational linear algebra, using R (although the coverage of particular topics in the books differs). All offer introductions
to R programming. Of these books, Braun and Murdoch is the briefest and most accessible.


**Graphics in R**


P. Murrell and R. Ihaka, “An approach to providing mathematical annotation in plots.” *Journal of Computational and Graphical Statistics*, 9:582-599, 2000. One of the unusual and very useful features of R graphics is the ability to include mathematical notation. This article explains how. Paul Murrell and Ross Ihaka are both members of the R core group.


**Data Management**

P. Spector, *Data Manipulation with R*. New York: Springer, 2008. Data management is a dry subject, but the ability to carry it out is vital to the effective day-to-day use of R (or of any statistical software). Spector provides a reasonably broad and clear introduction to the subject.
(Highly) Selected Statistical Methods Programmed in R

Also see the package listing on CRAN <http://cran.r-project.org/web/packages/index.html> and the various CRAN “task views” <http://cran.r-project.org/web/views/index.html>.

R. S. Bivand, E. J. Pebesma, and V. Gómez-Rubio, Applied Spatial Data Analysis with R, New York: Springer, 2008. There is a strong community of researchers in spatial statistics developing R software, much of which is described in this book, including the basic sp package, which provides R classes for spatial data. Roger Bivand is a member of the R Foundation.


R. Koenker, Quantile Regression. Cambridge: Cambridge University Press, 2005. Describes a variety of methods for quantile regression by the leading figure in the area. The methods are implemented in Koenker's quantreg package for R.
C. Loader, *Local Likelihood and Regression*. New York: Springer, 1999. Another text on nonparametric regression and density estimation, using the **locfit** package. Although the text is less readable than Bowman and Azzalini, the **locfit** software in very capable.


S. van Buuren, *Flexible Imputation of Missing Data*, Boca Raton FL: CRC Press, 2012. There are several packages in R for multiple imputation of missing data; this book largely describes the **mice** (multiple imputation by chained equations) package.

W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S, Fourth Edition*. New York: Springer, 2002. An influential and wide-ranging treatment of data analysis using S. Many of the facilities described in the book are programmed in the associated (and indispensable) **MASS**, **nnet**, and **spatial** packages, which are included in the standard R distribution. This text is more advanced and has a broader focus than the **R Companion**. Brian Ripley is a member of the R Core group of developers.


**Other Sources (Some Free)**