The R Statistical Computing Environment: The Basics and Beyond

The R statistical programming language and computing environment has become the de-facto standard for writing statistical software among statisticians and has made substantial inroads in the social sciences -- it is now possibly the most widely used statistical software in the world. R is a free, open-source implementation of the S language, and is available for Windows, Mac OS X, and Unix/Linux systems.

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, and is also associated with the Free Software Foundation.

A statistical software package, such as SPSS, is primarily oriented toward combining instructions, possibly entered via a point-and-click interface, with rectangular case-by-variable datasets to produce (often voluminous) output. Such packages make it easy to perform routine data analysis tasks, but they make it relatively difficult to do things that are innovative or nonstandard -- or to extend the built-in capabilities of the package.

In contrast, a good statistical computing environment makes routine data analysis easy and also supports convenient programming. R fulfills both of these requirements, and users can readily write programs that add to its already impressive facilities. Thousands of R add-on packages, freely available on the Internet in the Comprehensive R Archive Network (CRAN) <http://cran.r-project.org/web/packages/>, and many others in the Bioconductor package archive <http://www.bioconductor.org/>, extend the capabilities of R to almost every area of statistical data analysis. R is also particularly capable in the area of statistical graphics.

The first two days of this workshop are 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. I will also show you how to use RStudio, a sophisticated front-end to R, which includes support for “literate programming” to create documents that mix R code with explanatory text, encouraging reproducible research.

The material scheduled for day two is flexible, and I encourage participants to contact me with requests for topics to cover. (The topics that I’ve included in the course syllabus for day two – mixed-effects models, survival analysis, structural-equation models – should be read as suggestions.) Some caveats: (1) If you’d like me to cover another specific topic, please contact me sufficiently in advance of the workshop to prepare the necessary materials. (2) I’ll try to select topics that are of broad interest, to more than one participant. (3) I understand that, unlike the remainder of the workshop, not all day-two
topics will be of interest to all participants. (4) Of course, I’m limited in what I can competently teach by my knowledge and expertise.

Learning even a bit of R programming, will greatly increase your ability to manage and analyze data using R. The final three days pick up where the basic material leaves 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 of the workshop (see the instructions on the workshop website).

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

Day 1. Getting started with R (Ch. 1); linear and generalized models in R (Ch. 4, 5); literate programming and reproducible research using R and RStudio.

Day 2. Mixed-effects models and repeated-measures ANOVA and MANOVA with the *car, nlme, and lme4* packages (on-line appendices on *Multivariate Linear Models* and *Mixed-Effects Models*); Survival (“event-history”) analysis and structural-equation models with the *survival* and *sem* packages (on-line appendices on *Cox Regression for Survival Data* and *Structural-Equation Models*).

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

Day 4. R programming, beyond the basics (Ch. 8, Sec. 8.5-8.10)

Day 5. R graphics (Ch. 7); building R packages

**Workshop Web Site**

Materials for the workshop 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 and RStudio**

I recommend that you use R with the RStudio interactive development environment (IDE), which provides a sophisticated programming editor, and tools for literate programming for reproducible research, R package management, R package creation, and many other useful features. Both R and RStudio are free software, available on the Internet. Instructions for installing R and RStudio for Windows, Mac OS X, and Linux systems are on the workshop web site at <http://socserv.mcmaster.ca/jfox/Courses/R-course-Berkeley/R-install-instructions.html>, with a link on the workshop home page.
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 <http://www.springer.com/series/6991> and Chapman and Hall/CRC’s The R Series <http://www.crcpress.com/browse/series/crctherser>.

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 available on the web at <http://cm.bell-labs.com/cm/ms/departments/sia/doc/trellis.user.pdf>.

A great deal of information about using the RStudio interactive development environment is available on the RStudio website at <https://support.rstudio.com/hc/en-us> (see under “Documentation”).

Programming in R


designer of the S language and a member of the R Core group of developers. Not an easy read. (The “Green Book.”)

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.

H. Wickham, *Advanced R*. Boca Raton FL: Chapman and Hall/CRC, 2015. Hadley Wickham has contributed a number of widely used R packages (such as *ggplot2* for graphics and *plyr* for data manipulation) and is associated with RStudio. As the name
implies, you may (and should!) be interested in reading this book after you’ve learned the basics of R programming. A related volume by Wickham, *R Packages*, Sepastopol CA: O’Reilly, 2015, is (as its name implies) about how to write R packages. Wickham’s approach to R programming and package-writing is sometimes idiosyncratic but always carefully considered and interesting. The websites for the books <http://adv-r.had.co.nz/> and <http://r-pkgs.had.co.nz/> provide access to the text. Hadley Wickham is a member of the R Foundation.

Xie, Y., Dynamic Documents with R and knitr. Boca Raton FL: Chapman and Hall/CRC, 2013. Yihui Xie describes the use of his `knitr` package for creating LaTeX documents with embedded executable R code. This package also provides the basis for R Markdown in RStudio.

**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)**

See the publications list on the R web site <http://www.r-project.org/doc/bib/R-publications.html>. *The R Journal* <http://journal.r-project.org/>, the journal of the R Project for Statistical Computing, and its predecessor *R News* <http://www.r-project.org/doc/Rnews/index.html>, are also good sources of information, as is the *Journal of Statistical Software* <http://www.jstatsoft.org/>, an on-line American Statistical Association journal dominated by coverage of R packages. Information about R packages in a number of application areas is available in various “CRAN task views” <http://cran.r-project.org/web/views/>. The RStudio web site <https://www.rstudio.com/> is a good source of information both on using the RStudio IDE and on other topics, such as R Markdown.