Introduction to the R Statistical Computing Environment

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. 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, and is also associated with the Free Software Foundation.

A statistical package, such as SPSS, is primarily oriented toward combining instructions with rectangular case-by-variable datasets to produce (often voluminous) printouts. Statistical 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 (at the time of writing this syllabus) more than 5000 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 four lectures 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 assume here that the statistical content is familiar and emphasize the implementation of the various methods in R. 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 following five sessions 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 are encouraged to install R and bring their laptops to the lectures.
An outline of the classes follows (with chapter and references to Fox and Weisberg, *An R Companion to Applied Regression, Second Edition*):

1. Getting started with R and RStudio (Ch. 1)
2. Linear and generalized linear models in R (Ch. 4, 5)
3. 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*)
4. 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*)
5. Data and data management in R (Ch. 2)
6-8. Programming in R (Ch. 8)
9. R graphics (Ch. 7)

Course Web Site

Materials for the course will be deposited at
<http://socserv.socsci.mcmaster.ca/jfox/Courses/R/ICPSR-R-course/index.html>, abbreviation < tinyurl.com/ICPSR-R-course>, 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 course web site at
<http://socserv.socsci.mcmaster.ca/jfox/Courses/R/ICPSR/R-install-instructions.html>, with a link on the course 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 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 introduced *The R Series* <http://www.crcpress.com/browse/series/crcttherser>. 
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.

nonlinear models, generalized linear models, local-polynomial regression models, and generalized additive models. (The “White Book.”)


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)**

R packages in a number of application areas is available in various “CRAN task views”
<http://cran.r-project.org/web/views/>.