PROGRAMMING IN R

The S statistical programming language and computing environment has become the de-facto standard among statisticians and has made substantial inroads in the social sciences. A statistical software “package,” such as SPSS, makes routine data analysis relatively easy, but makes 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 additionally supports convenient programming.

This two-week course focuses on programming in R, a free, open-source implementation of S. (A commercial implementation of S called S-PLUS is still available, but has been eclipsed by R.) Some previous exposure to R is assumed, but not necessarily to programming in R, nor to programming in general. The course is project-oriented: In addition to small daily problems, each participant is encouraged to complete a substantial programming project in R, either working individually or as part of a collaborative team.

An outline of daily sessions follows (with, where appropriate, chapter and section references to Fox, 2002):

June 23: Review of data analysis using R (Ch., Sec. 2.1-2.2, Ch. 3-5); projects: getting started.

June 24: Basic R programming (Sec. 8.1-8.4, 8.6).

June 25: Drawing custom graphs in R (Ch. 7).

June 29: Data structures in R (Sec. 2.3-2.4); projects: status reports and problems.

June 30: Object-oriented programming in R (Sec. 8.4-8.5).

July 1: Improving and debugging R programs.

July 2: Writing R statistical modeling functions.

July 3: Building and writing R packages; projects: status reports and problems.
Projects

Because programming is primarily a learn-by-doing activity, I strongly encourage all participants to complete a substantial R programming project. As mentioned, projects may be undertaken by individuals or by teams. Good projects should be doable during at most the first month of the summer program, should be of interest to the programmers, and should be within the programmers’ statistical competence.

A project may entail writing a single function (e.g., to fit a particular type of statistical model); or, more likely, several related functions; or, even better, a package containing related functions, associated documentation, and possibly data. Because there are more than 1700 packages on CRAN (and more than 300 in the Bioconductor package repository), it may be difficult to find entirely novel projects. Although it would be desirable to write a novel program, the principal criterion for projects is that they provide useful learning experiences.

In addition to some class time devoted to projects, I and the TAs will have regular office hours to assist participants, and the TAs will convene additional help sessions.

Course Web Site

Materials for the course will be deposited at

BIBLIOGRAPHY

There has been an explosion of sources on R, with many titles of the form “X with R.” Moreover, many statistics books not specifically focused on R contain R code and examples. The following bibliography focuses instead on books that deal primarily with R programming, R graphics, and closely related topics.

An exception is my book, J. Fox, An R and S-PLUS Companion to Applied Regression, Sage, 2002. Although the focus is not on programming, and the book is a bit dated, many of the examples and much of the code that I use in this course are derived from it. The course is meant to be self-contained, and you do not have to read this or any other book to take it; you may, however, find relevant sections of the book, indicated in the course outline, to be useful. A web site for the book, with scripts for the examples, several online appendices, and some other R-related materials may be found at <http://socserv.mcmaster.ca/jfox/Books/Companion/index.html>. The book is associated with the car package for R.

Another useful general source that is not focused on programming is W. N. Venables and B. D. Ripley, Modern Applied Statistics with S, Fourth Edition, Springer, 2002. It, too, is slightly dated, but it covers many of the topics in this course, and has a wider range than Fox (2002). Venables and Ripley’s text is associated with the MASS, nnet, class, and spatial packages, which are part of the standard R distribution.
Manuals

R is distributed with a complete set of manuals, which are also available at the CRAN (Comprehensive R Archive Network) web site, <http://cran.r-project.org/manuals.html>.

A manual for Trellis Graphics (also useful for the lattice package in R) is at <http://cm.bell-labs.com/cm/ms/departments/sia/doc/trellis.user.pdf>.

Programming in R/S


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.


R. Gentleman, R Programming for Bioinformatics, Boca Raton: Chapman and Hall, 2009. A thorough, though at points relatively difficult, treatment of programming in R, by one of the original co-developers of R and a founder of the related Bioconductor Project (which develops computing tools for the analysis of genomic data). Don’t let the title fool you: Most of the contents of the book are quite general.


the definitive treatment of writing software in the various versions of S-PLUS and R; now somewhat dated, particularly with respect to R.

**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.


**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.
Other Sources (Some Free)