Introduction to the R Statistical Computing Environment

The R statistical computing environment has become the de-facto standard for writing statistical software among statisticians and has made substantial inroads in the social sciences and more generally – it is now possibly the most widely used statistical software in the world. R is a free, open-source implementation (and extension) of the S language, which was originally developed at Bell Labs in the mid-1970s. R is available for Windows, macOS, 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 most of the R Core group along with a number other individuals, and is also associated with the Free Software Foundation. See the R web site at [https://www.r-project.org/](https://www.r-project.org/).

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) printed output. Although they may include limited programming capabilities, 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. R is also particularly capable in the area of statistical graphics.

More than 14,000 R add-on packages (as of May 2019), freely available on the internet in the Comprehensive R Archive Network (CRAN) [https://cran.r-project.org/web/packages/index.html](https://cran.r-project.org/web/packages/index.html), and many others in the Bioconductor package archive [http://www.bioconductor.org/](http://www.bioconductor.org/), aimed primarily at researchers in bioinformatics, extend the capabilities of R to almost every area of statistical data analysis. R packages are also available from other, less curated, sources, including from Github [https://github.com](https://github.com).

These lectures provide a broad introduction to the R statistical computing environment. The first four lectures constitute 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 more or less familiar and emphasize the implementation of the various methods in R. I will also show you how to use the R Commander, a simple graphical user (“point-and-click”) interface (“GUI”) for R, and RStudio, a sophisticated but easy to use interactive development environment (IDE) for R, which includes support
for “literate programming” to create documents that mix R code with explanatory text, encouraging reproducible research.

Learning even a bit of R programming, however, will greatly increase your ability to manage and analyze data using R. The following five sessions pick up where the basic lectures leave off, and cover the background required to use R seriously for data management, 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 strongly encouraged to install R and to bring their (fully charged) laptops to the lectures.

The overall objective is to provide some facility in the use of R, to a level that enables participants to employ the software for assignments and projects in other Summer Program courses as well as in their own work, and to provide a foundation for learning more about R and statistical computing.


1. Getting started with R and the R Commander (Ch. 1, Secs. 1.2-1.3, 1.5-1.7)
2. Workflow in R, with RStudio and R Markdown (Ch. 1, Secs. 1.1, 1.4)
3. Linear, generalized linear, and mixed-effects models in R: basics (Chs. 4, 6, 7)
4. Plotting, testing, and checking statistical models (Chs. 6, 8)
5. Data and data management in R (Ch.2)
6, 7, and 9. Introduction to programming in R, parts 1, 2, and 3 (Ch. 10)
8. R graphics (Ch. 9)

The lectures are largely modular, so you should be able to attend those that interest you. Lectures 1 and 2 are basic, however, and necessary for much of the material in the other lectures. Lecture 5 is a prerequisite for lectures 6, 7, and 9.

The lectures are based mostly on material in Fox and Weisberg, *An R Companion to Applied Regression, Third Edition*, and partly on materials prepared specifically for the lectures.

**Lecture Series Web Site**

Materials for the lecture series will be deposited at <http://socserv.socsci.mcmaster.ca/jfox/Courses/R/ICPSR-R-course/>,
Acquiring R and RStudio

As mentioned, I strongly encourage participants to work along with me during the lectures, using the materials provided on the lecture-series website. I recommend that you use R with the RStudio IDE, which provides a sophisticated programming editor and tools for literate programming for reproducible research, R package management, R package creation, debugging, and many other useful features. Both R and RStudio are free software, available on the internet.

Detailed instructions for installing R, RStudio, and some other optional software, for Windows, macOS, and Linux systems, are on the lecture-series web site <http://socserv.socsci.mcmaster.ca/jfox/Courses/R/ICPSR/R-install-instructions.html>, with a link on the lecture-series home page.

The R Commander, used only briefly in the first lecture, is installed as a standard R package, although under macOS, it’s also necessary to install the XQuartz windowing system, which, like R and RStudio, is free software; details are on the installation web page for the lecture series.

Selected Bibliography

Publishers of statistical texts have been producing a steady stream of books on R that threatens to become a deluge. 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 Text

The principal source for this workshop is J. Fox and S. Weisberg, An R Companion to Applied Regression, Third Edition (Sage, 2019). Additional materials are available on the web site for the book <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>, abbreviation <tinyurl.com/rcompanion>, including several appendices (on multivariate linear models, survival analysis, Bayesian regression modeling, and more). The book is associated with the car and effects packages for R. It should not be necessary to read the book to understand the lectures.

Manuals

R is distributed with a set of manuals, which are accessible from RStudio, and which are also available at the CRAN web site <https://cran.r-project.org/manuals.html>.
A manual for S-PLUS Trellis Graphics (also useful for the \texttt{lattice} package in R) is available on the web at
\url{https://www.albany.edu/acc/courses/acc522fall2007/lecturenotes/trellis.usermanual.pdf}.

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

\textbf{Programming in R}


J. M. Chambers, \textit{Software for Data Analysis: Programming with R}. New York: Springer, 2008. This 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.


H. Wickham, *Advanced R*. Boca Raton: CRC/Chapman and Hall, 2015. Hadley Wickham has contributed a number of very widely used R packages (such as ggplot2 for graphics and dplyr 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 free access to the texts.


**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, if somewhat dated, introduction to the subject.

H. Wickham and G. Grolemund. *R for Data Science*. Sebastopol CA: O’Reilly, 2017. This book has quite a wide focus, touching on subjects such as statistical modeling,
statistical graphics, and reproducible research with R Markdown, but its real strength is in data management using various R packages in the “tidyverse” created by Hadley Wickham and his colleagues at RStudio. A website for the book is at <http://r4ds.had.co.nz/> and includes the text; also see <http://tidyverse.org/>.

(Highly) Selected Statistical Methods Programmed in R

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


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.


P. Mair, *Modern Psychometrics with R*. Cham, Switzerland: Springer, 2018. Shows how to use R for a wide range of standard psychometric methods, such as classical test theory, item-response theory, factor analysis, and structural equation modeling. The book provides an overview of the topics covered that’s broader than it is deep.

R. McElreath, *Statistical Retrinking: A Bayesian Course with Examples in R and Stan*. Boca Raton: CRC/Chapman and Hall, 2016. The title is reasonably descriptive of this very readable introduction to modern Bayesian methods. The use of R and Stan in the book is somewhat idiosyncratic, employing the author’s rethinking package, which is
freely available but not from CRAN. Stan implements state-of-art Hamiltonian Monte Carlo methods for drawing samples from posterior distributions, and may be accessed through R via the **rstan** package (which is on CRAN – see, e.g., the on-line appendix to Fox and Weisberg’s *R Companion* on Bayesian regression in R).


S. van Buuren, *Flexible Imputation of Missing Data*, Boca Raton: 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. A high-quality, influential, and wide-ranging treatment of data analysis using S. Many of the facilities described in the book are programmed in the associated (and very useful) **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*. I once considered the MASS book the best moderately advanced reference on statistical data analysis in S and R. The book is still useful, but it is showing its age.


Other Sources (Some Free)

- The home page of the R web site <https://www.r-project.org> has links to a variety of resources.
• *The R Journal* <https://journal.r-project.org/>, the journal of the R Project for Statistical Computing, and its predecessor *R News* <https://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/>, a free-access online and “high-impact” 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” <https://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 (see the link for “Resources”).