The Briefest of Introductions to R

Lawrence Hubert

Department of Psychology
The University of Illinois

A Programming Language and Software Environment for
Statistical Computing and Graphics

The R Programming Language

Originally developed by Ross Ihaka and Robert Gentleman in the late 1990's at the University of Auckland, New Zealand. It is now developed by the *R Development Core Team*

R is an implementation of the S programming language developed at Bell Laboratories by John Chambers and colleagues. Another implementation of S is in the commercially available product, S-Plus.

Extra Credit Question: What other languages and operating systems were developed at Bell Labs?

The name 'R' comes from (partly) the first name of the two original authors, and as a letter-play on the name 'S'.

The main place to get everything you need about R is:

http://www.r-project.org

and one of the CRAN mirrors (Comprehensive R Archive Network):

http://cran.r-project.org

R is widely used for statistical software development and data analysis; it could now be considered the *lingua franca* for statisticians, and is analogous to how MATLAB is viewed for the Engineering and Computer Science community.

R is "state of the art"; (almost all) statistical researchers
provide their "cutting edge" methods as R packages.

The source code is freely available under the GNU General
Public License; also, precompiled binary versions are available
(and free) for Windows, Mac OS X, and Linux.

Generally, R uses a command line interface, but several GUIs
are available. We will introduce one done by John Fox, called R
Commander, towards the end of our session.

If you need a GUI to do any of the analyses discussed, say, in
406/7, then R Commander is for you.

R supports a large variety of statistical and numerical
techniques in its (eight) base packages (e.g., the base stats
package); in fact, most standard methods (through all of
standard multivariate analysis) are already available in this
default installation.

R is also highly extensible through the use of packages —
user-submitted libraries for specific functions or specific areas of
study. We will list some later that are (most) relevant to
psychology.

Because of its S language lineage, R has stronger
object-oriented programming facilities than most statistical
computing languages.

The term "environment" is intended to characterize R as a fully planned and coherent system, rather than as an incremental accretion of very specific and inflexible tools; this is frequently the case with other data analysis software (e.g., SAS, SPSS, SYSTAT, ...)

R (like S) is designed around a true computer language; it allows users to add additional functionality by defining new functions. Much of the system is itself written in R, making it easy for users to follow the algorithmic choices made.

For computationally-intensive tasks, C, C++, and Fortran code can be linked and called at run time. Also, advanced users can write C code to manipulate R objects directly.

R is more than a statistics system. It is an environment within which statistical techniques are implemented.

R has its own Latex(-like) documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hardcopy.

Generally, a statistical package such as SPSS is oriented toward combining instructions and rectangular data sets to produce (voluminous) printout and graphs. Routine, standard data analysis is easy; innovative or nonstandard analysis is hard or impossible.

A programming environment is oriented toward transforming
one data structure into another. Programming environments,
such a R (or S), are extensible; standard data analysis is easy,
but so are innovation and nonstandard analysis.

One of R's strengths is its graphical capabilities, which produce
publication-quality graphs that include mathematical symbols.

Although R is mostly used by statisticians and other
practitioners requiring a complete environment for statistical
computation and software development, it could also be used
as a general matrix manipulation and calculation Toolbox.

Such a calculator usage is much like MATLAB (and the free
OCTAVE). This will not be our emphasis here, however.
Generally, I prefer MATLAB for this one-off calculator-type task
that deals explicitly with matrices.

Using R to do some matrix manipulations is fairly
straightforward, and comparable to how MATLAB does things.
You may need some of this facility, however, if you ever wish to
write your own functions, and, ultimately, your own packages.

Packages

The capabilities of R are extended by user-contributed *packages* (comparable to Toolboxes in MATLAB), allowing specialized statistical techniques, graphical devices, as well as programming interfaces and import/export capabilities to many external data formats.

A core set of packages are included with the installation of R, with over 1000 more available at a CRAN site.

Notable packages are listed along with comments on the official R Task View pages:

http://cran.r-project.org/web/views

Task views of particular interest are: Cluster; Environmetrics;
Multivariate; Psychometrics; Social Sciences

The Bioinformatics community has started a successful effort
to use R for the analysis of data from molecular biology
laboratories.

The Bioconductor Project, started in 2001, provides R packages
for the analysis of genomic data, such as Affymetrix and cDNA
microarray object-oriented data handling and analysis tools.

http://www.bioconductor.org

Jonathan Baron (U Penn, Psychology) has a very nice R help page. It gives a complete list of all packages, plus a search facility to look for what you might need:

http://finzi.psych.upenn.edu

Also, remember the R project page and the Cran mirrors:

http://www.r-project.org

http://www.cran.r-project.org

The *Journal of Statistical Software* is an on-line resource founded in 1996 by the American Statistical Association as a freely available and peer reviewed resource for statistical software and algorithms.

The site listed below for *JSS*, is devoted to an open-source philosophy. Thus, for both articles and code snippets, the source code is published along with the paper:

http://www.jstatsoft.org

Although code implementations can use different languages or computing environments, the emphasis is mainly on R, and, to a lesser extent, MATLAB.

Several other good sources of R related material:

R News was (now, *The R Journal*) the newsletter of the R
project, and features short to medium length articles covering
topics that might be of interest to users or developers of R. It
is all free, and on-line in pdfs:

http://journal.r-project.org

Bill Revelle (Northwestern, Psychology) maintains the
Personality Project, and a very good introduction to R for
psychological research:

http://www.personality-project.org/r/r.guide.html

See the handout from Rnews (December, 2007) on the Psychometrics Task View; and meta: An R Package for Meta-Analysis.

See the handout on Zelig (Everyone's Statistical Software) —

Extra Credit Question: Who was Zelig and what actor played him?

Multilevel Modeling: multilevel and nlme:

http://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf

Neural Networks: nnet and AMORE (A MORE flexible neural network package)

coin: A computational framework for conditional inference (i.e., think exact tests; permutation tests; randomization tests)

tsp: Infrastructure for the Traveling Salesperson Problem

tree: Classification and regression trees

seriation : Row/column object ordering in proximity matrices

sem: Structural equation models

rimage: Image processing module for R

mlica: Maximum likelihood implementation of Independent Components Analysis

lsa: Latent semantic analysis (using document-term matrices)

labdsv: Ordination (i.e., scaling) and multivariate analysis for ecology

gap: Genetic analysis package

fmri: Analysis of fMRI experiments

AnalyzeFMRI: Functions for analysis of fMRI data sets stored
in the ANALYZE or NIFTI format

ecodist: Dissimilarity-based functions for ecological analysis
(contains nmds for non-metric multidimensional scaling)

eba: Elimination-by-aspects models

lp: Interface to Lp_solve for linear/integer programs

Rglpk: Linear and mixed integer programming solver using GLPK (GNU Linear Programming Kit)

clValid: An R package for cluster validation (also, see the CRAN Task View on Cluster Analysis and Finite Mixture Models)

clue: cluster ensembles (contains a lot of my work on fitting ultrametrics and additive trees by least-squares iterative projection)

qgen: Quantitative genetics using R

psyphy: Functions for analyzing psychophysical functions

smacof: multidimensional scaling (from Jan de Leeuw)

psych: Procedures for personality and psychology research
(from Bill Revelle)

ade4: Analysis of ecological data: Exploratory and Euclidean
methods in environmental sciences

anacor: Simple and canonical correspondence analysis

amap: Another multidimensional analysis package

acepack: ace and avas for selecting regression transformations

e1071: Packages from Vienna, including svm (support vector machines) and lca (latent class analysis)

cba: Clustering for business analytics (includes order.optimal for ordering the leaves of a tree — one of my research areas as well)

MASS: Package for the text *Modern Applied Statistics with S* (contains isoMDS to carry out Kruskal's non-metric multidimensional scaling)

Packages for Social Network Analysis

sna: Tools for social network analysis

network: Classes for relational data

latentnetHRT and latentnet : Latent position and cluster models for networks

inetwork: Network analysis and plotting

statnet: A suite of packages for network analysis:

http://csde.washington.edu/statnet

ergm: Exponential-family models for networks

degreenet: Models for skewed count distributions relevant to
networks

Also, see STOCNET at

http://stat.gamma.rug.nl/snijders/

The is an open software system for the advanced statistical
analysis of social networks (primarily, the work of Tom
Snijders).

Gnumeric

The Gnumeric spreadsheet is part of the GNOME (Linux) desktop environment. GNOME is a project to create a free, user friendly desktop environment for Linux.

The goal of Gnumeric is to be the best possible spreadsheet. It will import existing Microsoft Excel files, and it won't screw-up simple statistical computations (as Excel routinely does; Microsoft seems incapable of coming up with a quality product — this is continually documented in the statistical literature).

There is an open-source build for Windows, as well as the various Linux versions. See:

http://www.gnome.org/projects/gnumeric

Gnumeric can even do linear and mixed integer programming.

It is an explicit friend of R.

One strong characteristic of R is the great help system. Try:

help(t.test)

help.search("cluster")

Also, the various manuals (involving a huge number of pages) are available in pdfs when you install R. The various contributed packages also come with various kinds of help documentation.

Before we can use R or any of the contributed programs, we need to get our data (usually, a subject by variable rectangular matrix) into what is called a *data frame*.

Here are a number of ways of doing this:

1) Keyboard input

2) Reading data into a data frame from a textfile

3) Using the spreadsheet-like data editor in R

4) Importing data from some spreadsheet format (e.g., *.xls), or from SPSS (e.g., *.sav)

5) Accessing data that are already in R libraries

We have our community data in a textfile called
communitydata.txt , with verbatim contents as follows:

accidents vehicles police
a 1 4 20
b 4 10 6
c 5 15 2
d 4 12 8
e 3 8 9
f 4 16 8
g 2 5 12
h 1 7 15
i 4 9 10
j 2 10 10

```
accidents = c(1,4,5,4,3,4,2,1,4,2)

comm.labels = c('a','b','c','d','e','f','g', 'h','i','j')

vehicles = scan()

4 10 15 12 8 16 5 7 9 10

police = scan()

20 6 2 8 9 8 12 15 10 10

community =
data.frame(comm.labels,accidents,vehicles,police)
```

community

community.altone $=$ edit(as.data.frame(NULL))

community.altone

fix(community)

```
community.alttwo =
read.table('G:/r_class_material/communitydata.txt', header =
TRUE)

community.alttwo
```

attach(community)

summary(community)

community.model = lm(accidents $\sim$ vehicles + police)

summary(community.model)

plot(community.model)

Accessing and Using Data in Libraries

Assuming that a package has been "installed":

install.packages('car')

library(car)

data(Prestige)

attach(Prestige)

objects()

search()

dim(Prestige)

mean(income)

summary(type)

summary(education)

plot(income,prestige)

abline(lm(prestige $\sim$ income))

title('Scatterplot of Prestige vs. Income')

identify(income,prestige,row.names(Prestige))

User Defined Functions

Functions are the core of R. They take inputs and produce outputs.

hello = function(x)

cat("hello",x,"\n")

hello("and goodbye")

```
absolute = function(x)

if (x < 0)

return (-x)

else

return(x)


absolute(-5)
```

```
factorial = function(x)

ret = 1

for(i in 1:x)

ret = ret*i


return(ret)

factorial(5)
```

```
factorial.alt = function(x)

i = 1

f = 1

while (i <= x)

f = f*i

i = i + 1


f

factorial.alt(5)
```

ourhistogram = function(x,breaks = "Scott", col =
"purple",...)

hist(x,breaks = breaks, probability = TRUE, col = col, ...)

x = rnorm(200)

ourhistogram(x,xlab = "histogram of x", col = "green")

R Commander

R Commander is a Basic Statistics GUI for R, written by John Fox from McMasters (Sociology)

install.packages("Rcmdr",dependencies = True)

library(Rcmdr)

RWinEdt

RWinEdt is a package that provides a plug-in for using WinEdt as an editor for R. This gives syntax highlighting, indenting, and so forth.

Remember the shareware license information on WinEdit that we have in our Latex presentation. The Psychology Department has a site license.