

## R

## Open Source Statistical Analysis

By Frank LoPresti, with Robert Norman & Marc Scott

*frank.lopresti@nyu.edu, robert.norman@nyu.edu, marc.scott@nyu.edu*

**R** is a statistics package that, like SAS or SPSS, can be used for statistical analysis or statistics courses at any level. It is also much more: it is free and open source!<sup>1</sup> Though my introduction to open source software took place almost two decades ago, the initiative has taken on a new meaning for me since I've started looking into R. Part of my excitement stems from R's interaction with the Web; R grew up on the Internet and has been designed to easily interact with the global alphabet soup of statistics available via the Web.

### A SAMPLE R SESSION

Our sample session will involve a look at a set of data on the weights and miles-per-gallon of various types of automobiles.

1. In the RGui window, open the "Packages" menu and select "Set CRAN Mirror." This is the Comprehensive R Archive Network (CRAN) site that the session will access for downloads and help. (I usually pick the University of North Carolina at Chapel Hill.)
2. Open the "Packages" menu again and select "Load Package." For this example, let's say we plan to import an SPSS dataset. Since we'll need more than the "Base"

package for this session, select the "Foreign" package.

3. The "R Console" window shows the commands that you generated by using the "RGui" pull-down menus in steps 1 and 2.
4. Enter the following commands in the "R Console" window at the prompt:

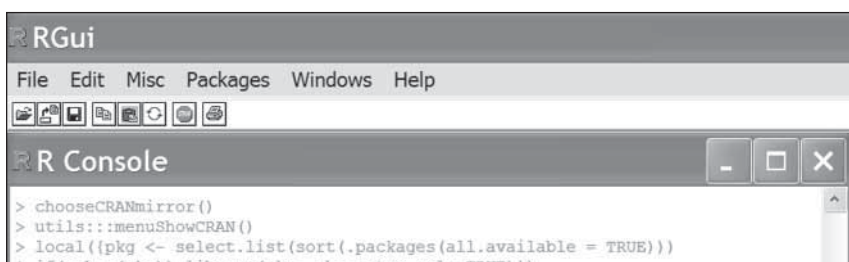
```
> cars<-read.spss("c:/Program
Files/SPSS/cars.sav")
> cars
> summary(cars$MPG)
> plot(cars$MPG,
cars$WEIGHT)
```

The "read.spss" command imports an SPSS file with all its labels and other metadata and assigns it to the

object "cars." The second command ("cars") allows us to see the names of the variables in the dataset "cars" on the screen. The third and fourth commands (more accurately, functions, since R is object oriented), "summary" and "plot," result in the simple statistics shown in figures 2 and 3. Notice how we use "cars\$MPG," since the variable "MPG" is a column of the data set "cars." Also, notice that R is case sensitive.

### R ARCHIVES

CRAN offers a window into the world of R. The CRAN page (<http://cran.r-project.org>) provides a variety of links. The "Mirrors" link points



**Fig. 1. The R Graphical User Interface (GUI) console at startup. The "RGui" frame holds the "R Console" frame, which is where we enter commands.**

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
9.00	17.50	23.00	23.51	29.00	46.60	8.00

**Fig. 2. Miles per gallon (MPG) obtained by a sample of automobiles. These simple statistics are the result of the "summary(cars\$MPG)" command.**

1. Open source software, as defined by the Open Source Initiative ([www.opensource.org/docs/definition.php](http://www.opensource.org/docs/definition.php)), must comply with certain criteria; notably, the software must be distributed free of charge and the source code must remain public for study and adaptation to other open source uses. If the open source code is used in the development of other software, that software cannot become proprietary.

## Why I Use R

By Marc Scott

The statistical languages S-PLUS and R grew out of the exploratory data analysis community that John Tukey\* and others built. At this point, nearly any statistical technique that you might want to employ is available for R as a downloadable library, making R not only a good environment for visualizing your data (it certainly is this), but also a great tool for estimating and evaluating simple to very complex models.

As a reasonably high-level programming language, R makes it easy to run regressions, plot data, and so on, but one could also use SAS or SPSS for that. Perhaps one way to understand the difference is that in SPSS, if you want to fit a loess (local polynomial) smooth to an X-Y plot, you have to double-click on the simple plot, click on a data point, and pull down a menu item “add (loess) regression line” to the plot. In R, you simply add lines to the first plot based on a call to the function `loess`; once you know how to add lines, you can add any type of line, and shade or color the line in unique ways.

In SPSS, you must rely on the software vendor to provide the functionality you need, but R is extensible—you can build routines of your own—and this is what really sets the program apart. SAS has an IML and macro facility that could be used in this manner, and STATA has a decent programming language as well, but in R, one is simultaneously writing functions and using them. Even the data are just another object that you can manipulate in your working environment. An additional reason I like R is that there is a large user community and a well-maintained archive of FAQs and other documentation that is readily available.

*Marc Scott is an Assistant Professor in the Department of the Humanities and Social Sciences in the Professions in NYU's Steinhardt School of Education.*

---

\* John Tukey was a statistician who made significant contributions to statistics and science during his career at Bell Laboratories and Princeton, and as a consultant to government and industry. (<http://cm.bell-labs.com/cm/ms/departments/sia/tukey/index.html>)

to other websites that duplicate the CRAN archives; it is generally a good idea to select a mirror at a relatively nearby location. Other links include “Task Views” (by types of statistical application), “Search,” and “Packages,” as well as “Manuals,” “FAQs,” and “Contributed” (documentation). You can connect to these archives from within an R session or simply by pointing your web browser to the above link.

The Task Views section offers the following types of analysis: “Bayesian,” “Cluster,” “Econometrics,” “Environmetrics,” “Finance,” “Genetics,” “gRaphical models,” “Machine & Statistical Learning,” “Multivariate,” “SocialSciences,” and “Spatial Analysis.” The views provide an indexed digest of R libraries gathered together in bundles and packages relating to a field of interest, with an introduction by the view’s maintainer. Each view reveals the statistical tools desktop of the main-

tainer, along with his or her research community’s related work (including, for example, those attributed to scientists working at prestigious research organizations, such as the National Institutes of Health, under National Cancer Institute grants).

### WHY USE R?

I’ve encountered a wide range of opinions about who should use R and how easy it is to learn. On the positive side, it is free and almost every student and researcher already owns a computer that can run R. It has an edit window, so we can view our datasets (for those of us who need the SPSS-like reassurance during a session that we really have data). It is also more user-friendly than Mathematics libraries and FORTRAN or C++. If you use recursion, you’ll like R. If you do quantitative computation daily, you may want to look into using R. On the other hand, it’s not for everyone;

I wouldn’t, for example, ask an introductory statistics class to use R.

### READY TO GET STARTED?

Take the following steps, and you’ll be off and running with this exciting open source program:

1. Carefully read “An Introduction to R” and “R Data Import/Export” from CRAN (<http://cran.r-project.org>; both are available on the “Manuals” page). Explore what else is available on the CRAN website while you’re there.
2. Download R from CRAN and install it on your PC. Look it over and get a feel for it, particularly for the pull-down menus in the RGui interface. Note that since R is object oriented, when you enter the name of any object, variable, or function, R will echo the contents of that object.
3. Look at other R users’ code. For example, I found a paper titled “A

## Why I Use R

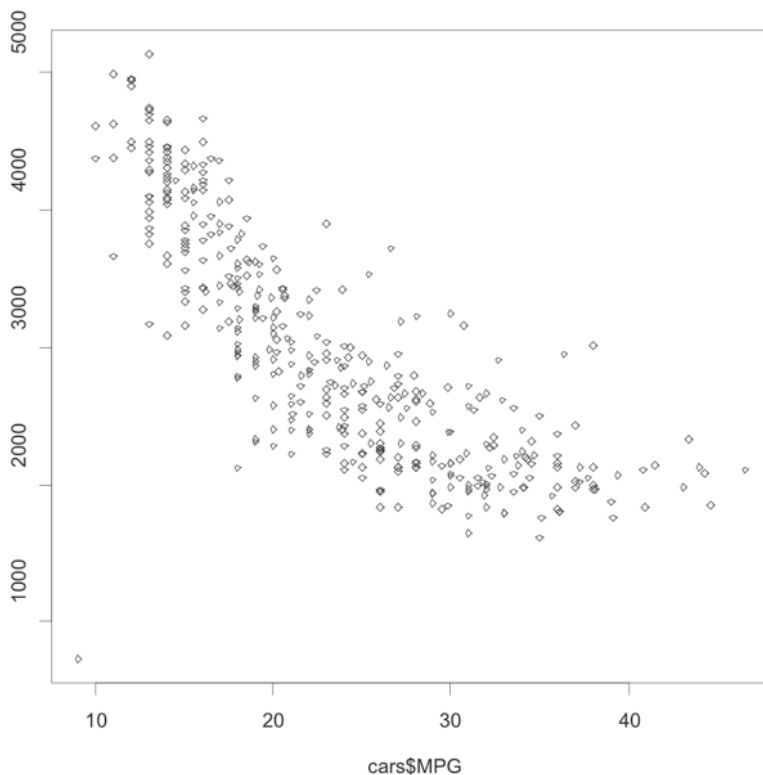
By Robert Norman

The R environment for statistical computing and graphics provides a unique programming environment in which to conduct research into new statistical methods as well as in the application of advanced methods. Its comprehensive programming language (S) has provided a platform for the development of a permutation-based method of evaluating inter-rater agreement under conditions in which standard measures are inappropriate, and for the development of linear mixed models of growth and decline in lung function, incorporating a random, unknown change point between these phases. While these developments would have been possible with other systems, they were much simpler and more rapid in R, due to its comprehensive mathematical and probability capabilities.

In addition to development, R provides a comprehensive collection of over 600 add-on packages that provide well-known, classical statistical methods, as well as those on the cutting edge of research. In a recent application of survival analysis modeling techniques in the description of human sleep continuity, some of the techniques we used were available in a minority of the existing commercial statistics packages. In R, however, all of the models were readily available and easily applied. In addition, S simplified the development of functions necessary to process the data for analysis.

While the R system may have a steeper learning curve than other statistical packages, it rewards its adherents with a power, simplicity, and clarity that is difficult to find elsewhere. It is well worth the effort to learn it.

*Robert Norman, an Adjunct Assistant Professor of Humanities and Social Sciences in the Professions at NYU's Steinhardt School of Education, will be teaching Biostatistics I in fall 2006. His ongoing research interests include measurement of sleep continuity, inter-rater agreement, linear mixed effects models, and mechanisms of chronic hypercapnia.*



**Figure 3. Cars' weights plotted against miles per gallon. Plot created by the command "`plot(cars$MPG, cars$WEIGHT)`". R is known for the high quality of its graphics.**

Computer Evolution in Teaching Undergraduate Time Series" by Erin M. Hodgess of the University of Houston – Downtown ([www.amstat.org/publications/jse/v12n3/hodgess.html](http://www.amstat.org/publications/jse/v12n3/hodgess.html)), *Journal of Statistics Education*, v12n3 (2004). This paper provides R code in an analysis context, with examples of Prof. Hodgess preparing a dataset, investigating it, then moving on to a time series Box-Cox analysis.

4. Search for other online R resources. I found R introductory course notes on the University of Wisconsin website at [www.stat.wisc.edu/courses/st371-ane/R/r.html](http://www.stat.wisc.edu/courses/st371-ane/R/r.html), and a helpful tutorial on the University of Illinois site at [www.econ.uiuc.edu/~econ472/tutorial2.html](http://www.econ.uiuc.edu/~econ472/tutorial2.html).

---

Frank LoPresti is a Senior Faculty Technology Specialist in ITS .edu Services' Faculty Technology Services division.