R-software: A Newer Tool in Epidemiological Data Analysis

Amir Maroof Khan
Department of Community Medicine, University College of Medical Sciences and GTB Hospital, Delhi, India

Background
An analyzing epidemiological data has always been a matter of concern especially for those researchers who have a background of biological sciences and not of mathematics. As the dataset is usually large in epidemiology, calculating even simple statistics like mean or standard deviation is quite cumbersome to be done manually. For many, even finding a statistician becomes difficult in their setting. So many datasets remain unexplored, sometimes forever waiting to be analyzed even by simple exploratory and descriptive data analysis.

Softwares in Data Analysis
With the introduction of softwares for statistical computations, things changed and data analysis came to be thought of something within the realm of possibility by the medical researchers. But for developing countries, the scenario did not change as expected because of the very high cost of the statistical packages.

The World Health Organization and Centers for Disease Control promoted free software known as Epi Info to be used by medical researchers. It was first launched as a Disk Operating System (DOS based) version, which was command driven and difficult to learn by the medical researchers. In 2001, windows-based version, which was menu driven, was launched and it became very popular among the medical researchers. Epi Info is also not suitable for data manipulation for longitudinal studies and its regression analysis facilities cannot cope with repeated measures and multilevel modeling. Also the graphing facilities are limited. Other statistical softwares such as Statistical Package for Social Sciences (SPSS), Stata, etc., are upgrading with newer dimensions in statistical analysis but they are not affordable to most institutions in developing countries.

What is R-software?
R is a relatively new and freely available programing language and software environment for statistical computing and graphics. The name is partly based on the (first) names of the first two R authors (Robert Gentleman and Ross Ihaka), and concept being partly taken from the name of the Bell Labs language 'S'.(1) It compiles and runs on a wide variety of UNIX platforms, Windows, and MacOS.(2) It has almost everything that an epidemiological data analyst needs. R is an environment that can handle several datasets simultaneously. R is also a programming language with an extensive set of built-in functions. One can write their own code to build their own statistical tools. Advanced users can even incorporate functions written in other languages, such as C, C++, and Fortran.(3) R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible. R is available as a Free Software under the terms of the Free Software Foundation’s GNU General Public License in source code form. One of R’s strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.(4)

The R Environment
R is an integrated suite of software facilities for data
manipulation, calculation and graphical display. It includes
• An effective data handling and storage facility,
• A suite of operators for calculations on arrays, in particular matrices,
• A large, coherent, integrated collection of intermediate
tools for data analysis,
• Graphical facilities for data analysis and display
either on-screen or on hardcopy, and
• A well-developed, simple and effective programming
language which includes conditionals, loops, user-
defined recursive functions and input and output
facilities.

The term environment is intended to characterize it as
a fully planned and coherent system, rather than an
incremental accretion of very specific and inflexible
 tools, as is frequently the case with other data analysis
software. R is not a typical statistics system but an
environment within which statistical techniques are
implemented. R can be extended via packages.\(^\text{(4)}\)

What is CRAN?

CRAN stands for Comprehensive R Archive Network.\(^\text{(2)}\)
 It is a network of ftp and web servers around the world
that store identical, up-to-date, versions of code and
documentation for R. One can use the nearest (with
respect to geographical location) CRAN mirror to
minimize network load. Apart from the packages which
automatically come with R; there are more than 2000
packages available at CRAN. So depending on the type
of statistical analytical techniques, one can download
the package required. CRAN does not have Windows
systems and therefore cannot check for viruses. It is
important to use the normal precautions that is taken
while downloading data on our hard disk.\(^\text{(5)}\)

Packages in R

The functions of R and its datasets are stored in
“packages,” whose contents are available only after it
has been downloaded. R is highly extensible through
the use of usersubmitted packages for specific functions
or specific areas of study. There are about 25 packages
supplied with R (called “standard” and “recommended”
packages) and many more are available through
the CRAN family of Internet sites (via http://CRAN.R-
project.org) and elsewhere. It requires some effort to
find which package contains the statistical techniques
that we require. For example, the “survfit” function
from the “survival” package computes the Kaplan-
Meier estimator for truncated and/or censored data
and various confidence intervals and confidence bands
for the Kaplan–Meier estimator are implemented in the
“km.ci” package.

There is an important difference between R and the
other main statistical systems. In R, a statistical analysis
is normally done as a series of steps, with intermediate
results being stored in objects. Thus whereas SAS and SPSS
will give all the details in the output from a regression or
discriminant analysis, R will give the desired and minimal
output and store the results in a fit object for subsequent
interrogation by further R functions.\(^\text{(6)}\)

Epicalc Package

Epicalc, an add-on package of R enables R to deal more
easily with epidemiological data. Epicalc, written by
Virasakdi Chongsuvivatwong of Prince of Songkla
University, Hat Yai, Thailand has been well accepted
by members of the R core-team and the package is
downloadable from CRAN which is mirrored by 69
academic institutes in 29 countries. The main advantage
of using this package is that it gives rise to display which
is more understandable by most epidemiologists. On
one hand, it assists data analysts in data exploration and
management. On the other hand, it has the potential to
help young epidemiologists to learn the key terms and
concepts based on numerical and graphical results of
the analysis. For basic biostatistical and epidemiological
purposes Epicalc package is sufficient to start with and
then to go on for other packages as and when required.

Limitations of R

R is provided with a command line interface (CLI),
which is the preferred user interface for power users
because it allows direct control on calculations and it
is flexible. However, good knowledge of the language
is required. CLI is thus intimidating for beginners. The
learning curve is typically longer than with a graphical
user interface (GUI), although it is recognized that the
effort is profitable and leads to better practice (finer
understanding of the analysis; command easily saved
and replayed).\(^\text{(7)}\) Therefore one has to understand what
one is doing or else giving a certain command will be
nearly impossible. The other limitation is that, being an
open source software, hackers can easily know about
the weaknesses or loopholes of the software more easily
than closed-source software and so it is more prone to
bug attacks.

Conclusions

Being free of cost, it is surely a boon for researchers in
developing countries and resource scarce institutions
The quality of this software in terms of handling large
datasets, having hundreds of functions with ever
increasing number of add on packages and the neat
outputs is also an advantage. As R is command driven,
learning R will by default make the user to attempt to
understand what is going on in the analysis and thus
learn the details of biostatistics and epidemiology. The steep learning of R is a serious disadvantage which if eased by the introduction of menu driven R can make it more popular among the non-mathematicians dealing with epidemiological data.

References

1. Frequently asked questions on R. Kurt Hornik. Available from: http://cran.r-project.org/doc/FAQ/R-FAQ.html#Why-is-R-names-R_003f [Last cited on 2010 June 8].
2. The R Project for statistical computing. Available from: http://www.r-project.org/ [Last cited on 2010 June 8].
3. R software introduction for stat 571. Available from: http://www.stat.wisc.edu/~yandell/st571/R/ [Last cited on 2010 June 8].
4. What is R. Available from: http://www.r-project.org/about.html [Last cited on 2010 June 9].
5. R for Windows. Available from: http://cran.stat.ucla.edu/bin/windows/ [Last cited on 2010 June 9]. [Last accessed on 2004 Apr 4].
6. An introduction to R. Available from: http://cran.r-project.org/doc/manuals/R-intro.html#Making-data-frames [Last cited on 2010 June 25].
7. R GUI projects. Available from: http://www.sciviews.org/_rgui/. [Last cited on 2010 Jul 26].