R: Modern Tool for Scientific Computing

Marcin Kozak *, Manjit S. Kang **

* Department of Biometry, Warsaw Agricultural University, Nowoursynowska 159, 02-787 Warsaw, Poland; E-mail: ***m.kozak@omega.sggw.waw.pl***; (corresponding author)

***** Vice Chancellor, Punjab Agricultural University, Thapar Hall, Ludhiana 141 004, India

Abstract: The R language and environment, free software, has a wide range of features that make it useful in statistical computing. These features may help the R language become a main tool for statisticians and those needing to analyze their research data. In this paper, we point the readers' attention to this software and discuss some of the advantages and disadvantages of R. [Nature and Science. 2007;5(2):41-43] (ISSN: 1545-0740).

Key words: data analysis, graphics, programming, software, statistical computing, statistics

Introduction

Nowadays, the necessity of using statistical software in research is undeniable. Statistical packages help analyze research data, teach statistical and non-statistical students, save time spent on statistical analysis, and even enable analyses that were not possible earlier (Ramasubramanian, 2002). They did improve quality in teaching statistics: non-statistical students may now find it easier to understand statistics and interpret results; only knowing how to calculate tedious formulae hardly helps them gain a better understanding of a problem.

There are many statistical packages for a researcher to choose from. What criteria should one take into account? First of all, of course, is the demand for analyses; then, functionality and ease of application/use must be considered. Acceptance of software by researchers throughout the world also must be taken into account. Finally, software cost may also be an important consideration; can the scientist, especially in Third World countries, afford it?

The R language and environment (R Development Core Team, 2006) shows promise of becoming a widely used computational tool in various disciplines. There are many features and issues that should make R software of choice in many statistical applications. This paper discusses these issues and shows why R might be an important analytical tool for researchers from both developed and developing countries, irrespective of their field of research.

Features, advantages and disadvantages of R

The R language, which is based on S language (Becker et al., 1988), is a system of statistical computation and graphics. It provides many useful features, including a wide range of statistical and mathematical functions, models and methods, a programming language, a matrix language, a user-friendly interface, high level graphics, and the like. Various features of R make it useful in classical data analyses as well as in advanced studies that require sophisticated computational tools.

Developing software. In the 10 years of its existence, R has become a tool for mathematicians, statisticians, engineers, biologists, psychologists, and other scientists. Since its inception, a huge team has worked on its development. Besides the basic R software, which is developed by a small core group, there are almost a thousand contributed add-on CRAN packages developed by numerous scientists who are not directly connected with the core R group. The add-on packages constitute a powerful tool for statistical analyses in an extremely wide range of areas of science. While other statistical programs, even the most expensive and popular ones, are rather slow in implementing novelties, R contains a lot of state-of-the-art methods, quite often implemented by their developers (e.g., Mclust and Mclust2, the packages written by Fraley and Raftery (2006), the acknowledged experts in model-based clustering).

Ease of use. At a cursory glance, R does not seem to be very user-friendly. Its environment requires the user to learn how to use its procedures, even for simple analyses. However, for those who may not want to learn R as a programming language and use sometimes quite complex R functions, there is a user-friendly interface in Rcmdr (R commander) package (Fox, 2006). This interface facilitates data handling and gives an opportunity to perform basic statistical analyses, such as, among others, basic parametric and non-parametric hypotheses, analysis of variance, regression analysis, general linear models, or some multivariate analyses (like principal component, factor, and cluster analyses). The user may also make quite interesting, useful graphs. Using Rcmdr seems to be as easy as most of the other statistical packages. Moreover, the R language provides advanced techniques to develop self-implemented procedures, which are necessary in analyses of sophisticated statistical problems and simulation studies.

Books with R. Because R has gained increasing attention in recent years, those who develop statistical methods and those who teach statistics also have started paying attention to this software. As a result, quite a few statistical books base their examples on R. The reader is encouraged to visit the R official web page, http://www.r-project.org, to see a long list of books concerned with R and S-plus. Examples are books by Dalgaard (2002), Faraway (2004), Fox (2002), and Jureckova and Picek (2006). The books from the list given by the R team deal with a wide range of statistical methods, from the introductory statistics and graphics to various advanced methodologies, such as regression; linear, generalized and mixed models; multivariate analysis; survival data; time series; engineering issues; environmental studies; geostatistics; phylogenetics and evolution; and bioinformatics and genomics. Of course, these topics have been discussed in the books mentioned, but R offers a much broader spectrum of possible analyses. This implies that R is not a local, unimportant tool for advanced programmers, but that it is directed to researchers who need to apply statistics to various kinds of data.

Free Software. That R is freeware may not be an important advantage for scientists from developed countries. However, those from developing countries may find it to be an important option when choosing software for their purposes. Of course, it should not be the only argument; linking it with what we have already said about R goes beyond the price issue.

Conclusion

As scientists' awareness of necessity of using statistics continues to increase, statistical packages will become more and more popular; soon, using a calculator for statistics (which is still sometimes used) will become obsolete. Among many statistical programmes, the R language and environment has a significant chance of becoming software of choice for researchers representing various scientific disciplines. It has already become a well-known package for statistical and mathematical computing, and in the near future researchers themselves should be able to use R as a tool for statistical analyses.

Correspondence to:
Marcin Kozak
Department of Biometry
Warsaw Agricultural University
Nowoursynowska 159, 02-787
Warsaw, Poland
E-mail: ***m.kozak@omega.sggw.waw.pl***

References

1. Ramasubramanian, V. Impact of statistical software packages on scientific research and statistical education. Current Science 2002;83(6):678.
2. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org;2006.
3. Becker, R.A., Chambers, J.M., Wilks, A.R. The new S language. Pacific Grove, Ca.: Wadsworth & Brooks;1988.
4. Fraley, C. and A.E. Raftery. MCLUST: Model-Based Clustering/Normal Mixture Modeling. R package version 3.0-0;2006.
5. Fox, J., with contributions from M. Ash, T. Boye, S. Calza, A. Chang, P. Grosjean, R. Heiberger, G. J. Kerns, R. Lancelot, M. Lesnoff, S. Messad, M. Maechler, D. Putler, M. Ristic and P. Wolf. Rcmdr: R

Commander. R package version 1.2-1. http://www.r-project.org, http://socserv.socsci.mcmaster.ca/jfox/Misc/Rcmdr/; 2006.
6. Dalgaard, P. Introductory Statistics with R. Springer;2002.
7. Faraway, J.J. Linear Models with R. Chapman & Hall/CRC, Boca Raton, FL;2004.
8. Fox, J. An R and S-Plus Companion to Applied Regression. Sage Publications, Thousand Oaks, CA, USA;2002.
9. Jureckova, J., and Picek, J. Robust Statistical Methods with R. Chapman & Hall/CRC, Boca Raton, FL;2006.