R and S

(Hadley Wickham, Rice University)

Abstract

R is a programming language and an environment for interactive data analysis. The first version of R was released in 1993, growing out of the S language developed at Bell Labs in the late 70's. The main strength of R is its rich ecosystem, which includes commercial vendors, thousands of add-on packages, mechanisms for getting help and many venues for scientific publishing.

1 Introduction

R is a programming language and interactive environment for data analysis and statistical computing. The development of R was guided by the principles of exploratory data analysis, with the driving goal to make it easy to ask and answer questions of data. R has an estimated two million users (Vance, 2009; Morgan, 2010) and has had a dramatic impact on the practice of data analysis and statistics, spanning every field of scientific endeavour.

R is an open-source implementation of the S programming language. S was originally developed by John Chambers, Rick Becker and Doug Dunn, Paul Tukey and Graham Wilkinson at AT&T Bell Labs in the late 70's (Becker, 1994). Somewhat later, spurred by the lack of affordable statistical software for teaching and some percevied flaws in S, Ross Ihaka and Robert Gentleman created an open source implementation. This programming language looked much like S, but behaved much like scheme, and was called R (Ihaka and Gentleman, 1996; Ihaka, 1996). The first alpha version of R was released in 1993, the first open source version in June 1995, version 1.0.0 in February 2000, version 2.0.0 in October 2004, and most recently version 2.13.0 in April 2011.

Rather discussing the language itself, I want to focus on a particular strength of R: its ecosystem. Commercial vendors provide paid support and enhanced functionality. R packages provide contributed functionality not provided in base R. A rich support system provides help when you are stuck, and R publications help learn R and keep up to date. Each of these topics is discussed in more depth below.

2 Commercial vendors

There are two main commercial vendors: TIBCO which sells S-Plus and Revolution Analytics which sells Revolution R. S-plus is not R, but was spun off as the commercial implementation of S. The language has changed hands a number of times: from Statistical Sciences to MathSoft to Insightful to TIBCO. Both products offer similar advantages over free R: commercial support, a user-friendly graphical user interface (still in development for Revolution R), and much better support for big data.

Recently, other major vendors of statistical software have started to provide interfaces to R. These vendors includes SAS[1], SPSS[2] and statistica[3].

3 R packages

An R package is a way of bundling R code and documentation for distribution to a wider audience. Packages can also contain data, demos, vignettes (long-form documentation and guides), compiled code (typically C or Fortran, but Java and others are possible) and automated tests. R packages can be used internally or distributed to the world through CRAN, the Common R Archive Network. To be published on CRAN, each

[1] `http://support.sas.com/rnd/app/studio/Rinterface2.html`

[2] `http://www-01.ibm.com/software/analytics/spss/products/statistics/developer/`

[3] `http://www.statsoft.com/solutions/r-language-platform/`

package must go through a stringent set of quality checks (`R CMD check`) on all major platforms: Windows, Mac and Linux (Theußl et al., 2011).

There are over 4,300 published R packages available on CRAN, so typically the challenge is not finding a package that does what you want, but choosing between multiple options. CRAN task views, `http://cran.r-project.org/web/views`, have been developed to help make sense of the huge number of packages. For example, the CRAN task view for environmetrics, maintained by Gavin Simpson, (`http://cran.r-project.org/web/views/Environmetrics.html`), is broken down into categories from relatively specialised topics like soil science (6 packages) to more general topics, like ordination (9 packages).

4 Getting help

Currently, the main venue for R support is the R-help mailing list, `https://stat.ethz.ch/mailman/listinfo/r-help`. This mailing list is high-volume, with approximately 10,000 members and over 100 messages per day (as of January 2010). The tone of R-help is somewhat irascible, and fools are no tolerated gladly, but much excellent advice is available when a clear question and a reproducible example is provided. There are other special interest mailing lists such as R-sig-mixed-models and R-sig-ecology; these tend to be easier to keep up with as they have fewer subscribers and many fewer messages.

Outside of R-help, community question and answer sites are becoming an increasingly popular and useful way of getting help. These include stackoverflow, `http://stackoverflow.com/tags/R`, oriented towards programming questions, and crossvalidated, `http://stats.stackexchange.com/tags/r`, oriented towards statistical questions. These sites offer a sophisticated reputation system designed to encourage informative responses and help good questions and answers float to the top.

5 Publications

While R-related papers are published in a wide range of statistics and applied journals, there are two open-access peer reviewed journals that feature many R related articles: the R journal and the Journal of Statistical software. The R Journal, `http://journal.r-project.org/`, focusses on short to medium length articles of interest to users and developers of R. The journal started in May 2009, growing out of unreviewed R News, which was founded in 2001. The Journal of Statistical Software (JSS), `http://www.jstatsoft.org/`, publishes articles on statistical software generally, but includes many articles on R packages. The papers tend to be much longer than the R Journal (20-40 pages compared to 4-5), and go in the statistical and software underpinning in much greater depth. JSS frequently publishes special issues on a topic area, such as Volume 22, `http://www.jstatsoft.org/v22/`, "Ecology and Ecological Modelling in R".

References

Rick Becker. A brief history of S. Technical Report 11, AT&T Bell Laboratories, 1994.

Ross Ihaka. R: Past and future history. In *Proceedings of Interface 96*, 1996. URL `http://www.stat.auckland.ac.nz/~ihaka/downloads/Interface98.pdf`.

Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

Timothy Prickett Morgan. Open source R in commercial revolution. *The Register*, 2010. URL `http://www.theregister.co.uk/2010/05/06/revolution_commercial_r/`.

Stefan Theußl, Uwe Ligges, and Kurt Hornik. Prospects and challenges in R package development. *Computational Statistics*, 26:395–404, 2011. ISSN 0943-4062. doi: 10.1007/s00180-010-0205-5.

Ashlee Vance. Data Analysts Captivated by R's Power. *NY Times*, 2009. URL http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?_r=1.