The Statistical Significance of R

Premal P. Vora [1]

[1]Penn State Harrisburg
School of Business Administration
Middletown, PA 17057.
fpv at psu.edu

October 19, 2009

What is R?

- R is a system for statistical computation and graphics.
- It consists of a language plus a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in script files.
- Released under GNU GPL.
- Officially a part of GNU.
- OS: Linux, Unix, Windows and Mac
- CPUs: i386, alpha, arm, hppa, ia64, m68k, mips/mipsel, powerpc, s390, x86_64, powerpc-apple-darwin, mips-sgi-irix, i386-freebsd, rs6000-ibm-aix, and sparc-sun-solaris.

Why should anyone care?

- “I keep saying that the sexy job in the next 10 years will be statiticians” said Hal Varian, chief economist at Google. “And I’m not kidding.”
- From a NY Times article published on August 6, 2009. Available at http://www.nytimes.com/2009/08/06/technology/06stats.html
- Drowning in data.
- Many, many closed commercial statistics packages available but not clear whether there is one winner.
- R has been widely and enthusiastically embraced in academia and in industry.
- R is open source, fast, solid, extensible.

Origins of R

- Created by Ross Ihaka and Robert Gentleman both at the University of Auckland (New Zealand).
- Follows the language definition of S as much as possible.
- S created at Bell Labs (Becker, Chambers, Wilks).
- R has lexical scoping – S does not.
- A lot like Scheme (Steele and Sussman) under the hood.
- Now developed by the R Development Core Team.
- Current stable release 2.9.2 (released on 2009-08-04).
- Available at http://www.r-project.org

Fundamental purpose of R

- A software tool for making inferences from data.
- Syntax and structure of language allows researcher to focus on asking the right questions and on coaxing answers from the data.
- Answers are trustworthy.

Philosophy of R

- A language for expressions and for assignments.
- Expressions are evaluated and the result is immediately displayed.
- Assignments also evaluate an expression, but the result is assigned to an object and not printed.
- All assigned namespaces are held in memory.
- A facility to save and to load assigned namespaces is available.

Simple numerical computation

- Operators: +, -, *, /, ^, ...
- Logical operators: ==, >, >=, !=, &, |, ...
- Hundreds of mathematical, statistical, and other functions: sqrt, log, log10, cos, tan, sum, min, max, mean, median, sort, ...
- Functions operate on numbers and a variety of data objects.

(Data) objects

- Data objects: scalars, vectors, matrices, lists, dataframes.
- Objects can contain numbers, strings, logical quantities, or other objects.
- All elements in vectors and matrices must be of the same "mode" (R converts non-conforming elements on the fly when necessary).
- A list is a flexible data object that can contain other data objects, each of a different mode.
- A dataframe is like a matrix but each vector in that matrix is of a particular "mode". Allows a collection of data of different modes to be treated as one object.

Visualization/Graphics

- Base R automatically loads packages for creating visual displays of data.
- Very strong in this area.
- Graphics are customizable.
- Many open-source add-on packages for graphics are available.
- You can always write your own using the graphics primitives in R if dissatisfied with what's available.

Extension

- Language for creating your own functions.
- Collect a group of functions into a package and share it with others if you like.
- Get feedback from the user base.

Live Demo

- Simple numerical computations.
- Math functions.
- Logical comparisons.
- Vectors, vector arithmetic.
- Reading a dataframe.
- Making tabular summaries of data.
- Visualization of x,y data.
- Running linear regression: $y = a + bx + e$.

Help Sources

- Built-in help: questionmarkfunction
- Built-in help: help.search("subject")
- Built-in help: example(functionname)
- Online help: At the project website...several well-written manuals
- Online help: Several mailing lists including R-help
- Some mailing lists are devoted to specific special-interest groups such as R-SIG-Finance, R-SIG-ecology, etc.
- Support from third-party commercial firms is also available (for a fee).

Strengths

- Fast, reliable, powerful, flexible, solid alternative to commercial packages.
- Well-known in academia and in industry.
- Many packages (2,000?) for different tasks available.
- Active development, eager support.
- Availability of support from third-party commercial firms makes it viable for proprietary use.
- Open source, GPL License.

Weaknesses (All personal observations)

- Memory management restricts data set size to size of(RAM + swap space - space occupied by other processes).
- Learning curve for basic to intermediate usage is relatively flat, but thereafter steep.