Biostat 140.776 - Basic Information on R

Ingo Ruczinski

The information in the following slides was partially gathered from the websites of John Chambers and Robert Gentleman, and the CRAN website at http://www.r-project.org/

Some S facts

- S is a language and system for organizing, visualizing, and analyzing data.
- S has been a project of statistics research at Bell Labs since 1976.
- The language has evolved through several major versions to become the most widely used environment for research in data analysis and statistics.
- In 1998, S became the first statistical system to receive the Software System Award, the top software award from the ACM.
The stages in the evolution of S

<table>
<thead>
<tr>
<th>Version</th>
<th>Period</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version 4</td>
<td>1998-</td>
<td>Formal class/method model (Roger will talk about this in detail), interfaces, the green book.</td>
</tr>
</tbody>
</table>

Visit http://cm.bell-labs.com/cm/ms/departments/sia/S/ for more information.

Products and projects based on S

- The S-Plus language is based on the S software from Bell Labs. S-Plus products are distributed by the Insightful Corporation, which has an exclusive license to distribute software based on Bell Labs’ S.

- The R language is an open-source system distributed under the GPL license. It is a separate project based on the S language, with a number of differences to Splus.

- Omegahat is also an open-source project for development of (“next-generation”) software for statistical applications, with emphasis on web-based software.

- Bioconductor is an open source software project for the analysis and comprehension of genomic data.
Some R facts

• R is an environment for data analysis and visualization.

• R is an open source implementation of the S language (S-Plus is a commercial implementation of the S language).

• The current version of R (as of November 4, 2003) is 1.8.0.

• The R Core group consists of Doug Bates, John Chambers, Peter Dalgaard, Robert Gentleman, Kurt Hornik, Stefano Iacus, Ross Ihaka, Friedrich Leisch, Thomas Lumley, Martin Maechler, Guido Masarotto, Paul Murrell, Brian Ripley, Duncan Temple Lang, and Luke Tierney.

• If you use R extensively, be a good citizen and join the R Foundation for Statistical Computing! http://www.r-project.org/

The R history

1991 Ross Ihaka and Robert Gentleman begin work on a project that will ultimately become R.
1992 Design and implementation of pre-R.
1993 The first announcement of R.
1995 R available by ftp under the GPL.
1996 A mailing list is started and maintained by Martin Maechler at ETH.
1997 The R core group is formed.
1999 DSC meeting in Vienna, the first time many R core members meet.
2000 R 1.0.0 is released.
2003 R 1.8.0 is the current release.
The R license

- R is both open source and open development.
- You can look at the source code and you can propose changes.
- R is not in the public domain.
- You are given a license to run the software (currently GPL).

The R software

- R has many of the features of S but with slightly different semantics (based on Scheme) and memory management.
- R is mainly written in C.
- R is available for many platforms:
  - Unix of many flavors, including Linux, Solaris, FreeBSD, AIX.
  - Windows 95 and later.
  - MacOS X.
- Both binaries and source code are available from www.r-project.org.
- R “talks” to databases, programming languages, and other statistical packages.
- R should be source code compatible with most of the Splus code written.

Roger will later talk about some internal issues like environments, lexical scope, . . .
The R package system

- Packages are self-contained units of code with documentation.

- There are automatic testing features built in.

- All functions must have examples and the examples must run.

- The packages are simple to obtain, figure out, and update (commands like `install.packages()`, `example()`, `update.packages()`, etc. help).

Rafa will later show you how to write your own packages!