R FAQ

Frequently Asked Questions on R
Version 3.1.2015-03-04

Kurt Hornik

Table of Contents

1 Introduction 1

1.1 Legalese 1
1.2 Obtaining this document 1
1.3 Citing this document 1
1.4 Notation 1
1.5 Feedback 1

2 R Basics 3

2.1 What is R? 3
2.2 What machines does R run on? 3
2.3 What is the current version of R? 4
2.4 How can R be obtained? 4
2.5 How can R be installed? 4
2.5.1 How can R be installed (Unix-like) 4
2.5.2 How can R be installed (Windows) 5
2.5.3 How can R be installed (Mac) 5
2.6 Are there Unix-like binaries for R? 5
2.7 What documentation exists for R? 6
2.8 Citing R 8
2.9 What mailing lists exist for R? 8
2.10 What is CRAN? 9
2.11 Can I use R for commercial purposes? 10
2.12 Why is R named R? 10
2.13 What is the R Foundation? 10
2.14 What is R-Forge? 10

3 R and S 12

3.1 What is S? 12
3.2 What is S-PLUS? 12
3.3 What are the differences between R and S? 12
3.3.1 Lexical scoping 13
3.3.2 Models 16
3.3.3 Others 16
3.4 Is there anything R can do that S-PLUS cannot? 18
3.5 What is R-plus? 19

4 R Web Interfaces 20

5 R Add-On Packages ... 22

5.1 Which add-on packages exist for R? ... 22
5.1.1 Add-on packages in R ... 22
5.1.2 Add-on packages from CRAN ... 22
5.1.3 Add-on packages from Omegahat ... 23
5.1.4 Add-on packages from Bioconductor ... 24
5.1.5 Other add-on packages ... 24
5.2 How can add-on packages be installed? ... 24
5.3 How can add-on packages be used? ... 24
5.4 How can add-on packages be removed? ... 25
5.5 How can I create an R package? ... 26
5.6 How can I contribute to R? ... 26

6 R and Emacs ... 27

6.1 Is there Emacs support for R? ... 27
6.2 Should I run R from within Emacs? ... 27
6.3 Debugging R from within Emacs ... 28

7 R Miscellanea ... 29

7.1 How can I set components of a list to NULL? ... 29
7.2 How can I save my workspace? ... 29
7.3 How can I clean up my workspace? ... 29
7.4 How can I get eval() and D() to work? ... 29
7.5 Why do my matrices lose dimensions? ... 30
7.6 How does autoloading work? ... 30
7.7 How should I set options? ... 30
7.8 How do file names work in Windows? ... 31
7.9 Why does plotting give a color allocation error? ... 31
7.10 How do I convert factors to numeric? ... 31
7.11 Are Trellis displays implemented in R? ... 31
7.12 What are the enclosing and parent environments? ... 32
7.13 How can I substitute into a plot label? ... 32
7.14 What are valid names? ... 33
7.15 Are GAMs implemented in R? ... 33
7.16 Why is the output not printed when I source() a file? ... 33
7.17 Why does outer() behave strangely with my function? ... 34
7.18 Why does the output from anova() depend on the order of factors in the model? ... 34
7.19 How do I produce PNG graphics in batch mode? ... 35
7.20 How can I get command line editing to work? ... 35
7.21 How can I turn a string into a variable? ... 35
7.22 Why do lattice/trellis graphics not work? ... 36
7.23 How can I sort the rows of a data frame? ... 36
7.24 Why does the help.start() search engine not work? ... 36
7.25 Why did my .Rprofile stop working when I updated R? ... 36
7.26 Where have all the methods gone? ... 37
7.27 How can I create rotated axis labels? ... 37

7.28 Why is read.table() so inefficient? 37
7.29 What is the difference between package and library? 37
7.30 I installed a package but the functions are not there 38
7.31 Why doesn't R think these numbers are equal? 38
7.32 How can I capture or ignore errors in a long simulation? 38
7.33 Why are powers of negative numbers wrong? 39
7.34 How can I save the result of each iteration in a loop into a separate file? 39
7.35 Why are *p*-values not displayed when using lmer()? 39
7.36 Why are there unwanted borders, lines or grid-like artifacts when viewing a plot saved to a PS or PDF file? 39
7.37 Why does backslash behave strangely inside strings? 40
7.38 How can I put error bars or confidence bands on my plot? 41
7.39 How do I create a plot with two y-axes? 41
7.40 How do I access the source code for a function? 41
7.41 Why does summary() report strange results for the R^2 estimate when I fit a linear model with no intercept? 42
7.42 Why is R apparently not releasing memory? 42

8 R Programming 44

8.1 How should I write summary methods? 44
8.2 How can I debug dynamically loaded code? 44
8.3 How can I inspect R objects when debugging? 44
8.4 How can I change compilation flags? 44
8.5 How can I debug S4 methods? 44

9 R Bugs 45

9.1 What is a bug? 45
9.2 How to report a bug 45

10 Acknowledgments 47

1 Introduction

This document contains answers to some of the most frequently asked questions about R.

1.1 Legalese

This document is copyright © 1998–2015 by Kurt Hornik.

This document is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2, or (at your option) any later version.

This document is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

Copies of the GNU General Public License versions are available at

http://www.R-project.org/Licenses/

1.2 Obtaining this document

The latest version of this document is always available from

http://CRAN.R-project.org/doc/FAQ/

From there, you can obtain versions converted to plain ASCII text, GNU info, HTML, PDF, as well as the Texinfo source used for creating all these formats using the GNU Texinfo system.

You can also obtain the R FAQ from the `doc/FAQ` subdirectory of a CRAN site (see Section 2.10 [What is CRAN?], page 9).

1.3 Citing this document

In publications, please refer to this FAQ as Hornik (2015), "The R FAQ", and give the above, *official* URL:

```
@Misc{,
  author        = {Kurt Hornik},
  title         = {{R} {FAQ}},
  year          = {2014},
  url           = {http://CRAN.R-project.org/doc/FAQ/R-FAQ.html}
}
```

1.4 Notation

Everything should be pretty standard. '`R>`' is used for the R prompt, and a '`$`' for the shell prompt (where applicable).

1.5 Feedback

Feedback via email to Kurt.Hornik@R-project.org is of course most welcome.

In particular, note that I do not have access to Windows or Mac systems. Features specific to the Windows and OS X ports of R are described in the "R for Windows FAQ"

and the "R for Mac OS X FAQ". If you have information on Mac or Windows systems that you think should be added to this document, please let me know.

2 R Basics

2.1 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.

The design of R has been heavily influenced by two existing languages: Becker, Chambers & Wilks' S (see Section 3.1 [What is S?], page 12) and Sussman's Scheme. Whereas the resulting language is very similar in appearance to S, the underlying implementation and semantics are derived from Scheme. See Section 3.3 [What are the differences between R and S?], page 12, for further details.

The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. Most of the user-visible functions in R are written in R. It is possible for the user to interface to procedures written in the C, C++, or FORTRAN languages for efficiency. The R distribution contains functionality for a large number of statistical procedures. Among these are: linear and generalized linear models, nonlinear regression models, time series analysis, classical parametric and nonparametric tests, clustering and smoothing. There is also a large set of functions which provide a flexible graphical environment for creating various kinds of data presentations. Additional modules ("add-on packages") are available for a variety of specific purposes (see Chapter 5 [R Add-On Packages], page 22).

R was initially written by Ross Ihaka and Robert Gentleman at the Department of Statistics of the University of Auckland in Auckland, New Zealand. In addition, a large group of individuals has contributed to R by sending code and bug reports.

Since mid-1997 there has been a core group (the "R Core Team") who can modify the R source code archive. The group currently consists of Doug Bates, John Chambers, Peter Dalgaard, Seth Falcon, Robert Gentleman, Kurt Hornik, Stefano Iacus, Ross Ihaka, Friedrich Leisch, Uwe Ligges, Thomas Lumley, Martin Maechler, Duncan Murdoch, Paul Murrell, Martyn Plummer, Brian Ripley, Deepayan Sarkar, Duncan Temple Lang, Luke Tierney, and Simon Urbanek.

R has a home page at `http://www.R-project.org/`. It is free software distributed under a GNU-style copyleft, and an official part of the GNU project ("GNU S").

2.2 What machines does R run on?

R is being developed for the Unix-like, Windows and Mac families of operating systems. Support for Mac OS Classic ended with R 1.7.1.

The current version of R will configure and build under a number of common Unix-like (e.g., `http://en.wikipedia.org/wiki/Unix-like`) platforms including *cpu*-linux-gnu for the i386, amd64, alpha, arm/armel, hppa, ia64, m68k, mips/mipsel, powerpc, s390 and sparc CPUs (e.g., `http://buildd.debian.org/build.php?&pkg=r-base`), i386-hurd-gnu, *cpu*-kfreebsd-gnu for i386 and amd64, powerpc-apple-darwin, mips-sgi-irix, i386-freebsd, rs6000-ibm-aix, and sparc-sun-solaris.

If you know about other platforms, please drop us a note.

2.3 What is the current version of R?

The current released version is 3.1.3. Based on this 'major.minor.patchlevel' numbering scheme, there are two development versions of R, a patched version of the current release ('r-patched') and one working towards the next minor or eventually major ('r-devel') releases of R, respectively. Version r-patched is for bug fixes mostly. New features are typically introduced in r-devel.

2.4 How can R be obtained?

Sources, binaries and documentation for R can be obtained via CRAN, the "Comprehensive R Archive Network" (see Section 2.10 [What is CRAN?], page 9).

Sources are also available via `https://svn.R-project.org/R/`, the R Subversion repository, but currently not via anonymous rsync (nor CVS).

Tarballs with daily snapshots of the r-devel and r-patched development versions of R can be found at `ftp://ftp.stat.math.ethz.ch/Software/R/`.

2.5 How can R be installed?

2.5.1 How can R be installed (Unix-like)

If R is already installed, it can be started by typing *`R`* at the shell prompt (of course, provided that the executable is in your path).

If binaries are available for your platform (see Section 2.6 [Are there Unix-like binaries for R?], page 5), you can use these, following the instructions that come with them.

Otherwise, you can compile and install R yourself, which can be done very easily under a number of common Unix-like platforms (see Section 2.2 [What machines does R run on?], page 3). The file `INSTALL` that comes with the R distribution contains a brief introduction, and the "R Installation and Administration" guide (see Section 2.7 [What documentation exists for R?], page 6) has full details.

Note that you need a FORTRAN compiler or perhaps `f2c` in addition to a C compiler to build R.

In the simplest case, untar the R source code, change to the directory thus created, and issue the following commands (at the shell prompt):

```
$ ./configure
$ make
```

If these commands execute successfully, the R binary and a shell script front-end called `R` are created and copied to the `bin` directory. You can copy the script to a place where users can invoke it, for example to `/usr/local/bin`. In addition, plain text help pages as well as HTML and LaTeX versions of the documentation are built.

Use *`make dvi`* to create DVI versions of the R manuals, such as `refman.dvi` (an R object reference index) and `R-exts.dvi`, the "R Extension Writers Guide", in the `doc/manual` subdirectory. These files can be previewed and printed using standard programs such as `xdvi` and `dvips`. You can also use *`make pdf`* to build PDF (Portable Document Format) version of the manuals, and view these using e.g. Acrobat. Manuals written in the GNU Texinfo system can also be converted to info files suitable for reading online with Emacs

or stand-alone GNU Info; use *make info* to create these versions (note that this requires Makeinfo version 4.5).

Finally, use *make check* to find out whether your R system works correctly.

You can also perform a "system-wide" installation using *make install*. By default, this will install to the following directories:

`${prefix}/bin`
: the front-end shell script

`${prefix}/man/man1`
: the man page

`${prefix}/lib/R`
: all the rest (libraries, on-line help system, . . .). This is the "R Home Directory" (`R_HOME`) of the installed system.

In the above, `prefix` is determined during configuration (typically `/usr/local`) and can be set by running `configure` with the option

```
$ ./configure --prefix=/where/you/want/R/to/go
```

(E.g., the R executable will then be installed into `/where/you/want/R/to/go/bin`.)

To install DVI, info and PDF versions of the manuals, use *make install-dvi*, *make install-info* and *make install-pdf*, respectively.

2.5.2 How can R be installed (Windows)

The `bin/windows` directory of a CRAN site contains binaries for a base distribution and add-on packages from CRAN to run on Windows XP and later (including 64-bit versions of Windows) on ix86 and x86_64 chips. The Windows version of R was created by Robert Gentleman and Guido Masarotto, and is now being developed and maintained by Duncan Murdoch and Brian D. Ripley.

The same directory has links to snapshots of the r-patched and r-devel versions of R.

See the "R for Windows FAQ" for more details.

2.5.3 How can R be installed (Mac)

The `bin/macosx` directory of a CRAN site contains a standard Apple installer package to run on OS X 10.6 ('Snow Leopard') and later. Once downloaded and executed, the installer will install the current release of R and an R.app OS X GUI. This port of R for OS X is maintained by Simon Urbanek (and previously by Stefano Iacus). The "R for Mac OS X FAQ has more details.

Snapshots of the r-patched and r-devel versions of R are available as Apple installer packages at `http://r.research.att.com`.

2.6 Are there Unix-like binaries for R?

The `bin/linux` directory of a CRAN site contains the following packages.

	CPU	Versions	Provider
Debian	i386/amd64	squeeze/wheezy	Johannes Ranke
	armel	wheezy	Johannes Ranke
Ubuntu	i386/amd64	lucid/precise/trusty	Michael Rutter

Debian packages, maintained by Dirk Eddelbuettel, have long been part of the Debian distribution, and can be accessed through APT, the Debian package maintenance tool. Use e.g. `apt-get install r-base r-recommended` to install the R environment and recommended packages. If you also want to build R packages from source, also run `apt-get install r-base-dev` to obtain the additional tools required for this. So-called "backports" of the current R packages for at least the *stable* distribution of Debian are provided by Johannes Ranke, and available from CRAN. See http://CRAN.R-project.org/bin/linux/debian/README for details on R Debian packages and installing the backports, which should also be suitable for other Debian derivatives. Native backports for Ubuntu are provided by Michael Rutter.

R binaries for Fedora, maintained by Tom "Spot" Callaway, are provided as part of the Fedora distribution and can be accessed through `yum`, the RPM installer/updater. Note that the "Software" application (gnome-software), which is the default GUI for software installation in Fedora 20, cannot be used to install R. It is therefore recommended to use the yum command line tool. The Fedora R RPM is a "meta-package" which installs all the user and developer components of R (available separately as `R-core` and `R-devel`), as well as `R-java`, which ensures that R is configured for use with Java. The R RPM also installs the standalone R math library (`libRmath` and `libRmath-devel`), although this is not necessary to use R. When a new version of R is released, there may be a delay of up to 2 weeks until the Fedora RPM becomes publicly available, as it must pass through the statutory Fedora review process. RPMs for a selection of R packages are also provided by Fedora. The Extra Packages for Enterprise Linux (EPEL) project (http://fedoraproject.org/wiki/EPEL) provides ports of the Fedora RPMs for RedHat Enterprise Linux and compatible distributions (e.g., Centos, Scientific Linux, Oracle Linux).

See http://CRAN.R-project.org/bin/linux/suse/README.html for information about RPMs for openSUSE.

No other binary distributions are currently publically available via CRAN.

2.7 What documentation exists for R?

Online documentation for most of the functions and variables in R exists, and can be printed on-screen by typing *`help(name)`* (or *`?name`*) at the R prompt, where *name* is the name of the topic help is sought for. (In the case of unary and binary operators and control-flow special forms, the name may need to be be quoted.)

This documentation can also be made available as one reference manual for on-line reading in HTML and PDF formats, and as hardcopy via LaTeX, see Section 2.5 [How can R be installed?], page 4. An up-to-date HTML version is always available for web browsing at http://stat.ethz.ch/R-manual/.

Printed copies of the R reference manual for some version(s) are available from Network Theory Ltd, at http://www.network-theory.co.uk/R/base/. For each set of manuals

sold, the publisher donates USD 10 to the R Foundation (see Section 2.13 [What is the R Foundation?], page 10).

The R distribution also comes with the following manuals.

- "An Introduction to R" (`R-intro`) includes information on data types, programming elements, statistical modeling and graphics. This document is based on the "Notes on S-PLUS" by Bill Venables and David Smith.
- "Writing R Extensions" (`R-exts`) currently describes the process of creating R add-on packages, writing R documentation, R's system and foreign language interfaces, and the R API.
- "R Data Import/Export" (`R-data`) is a guide to importing and exporting data to and from R.
- "The R Language Definition" (`R-lang`), a first version of the "Kernighan & Ritchie of R", explains evaluation, parsing, object oriented programming, computing on the language, and so forth.
- "R Installation and Administration" (`R-admin`).
- "R Internals" (`R-ints`) is a guide to R's internal structures. (Added in R 2.4.0.)

An annotated bibliography (BibTEX format) of R-related publications can be found at

`http://www.R-project.org/doc/bib/R.bib`

Books on R by R Core Team members include

John M. Chambers (2008), "Software for Data Analysis: Programming with R". Springer, New York, ISBN 978-0-387-75935-7, `http://statweb.stanford.edu/~jmc4/Rbook/`.

Peter Dalgaard (2008), "Introductory Statistics with R", 2nd edition. Springer, ISBN 978-0-387-79053-4, `http://www.biostat.ku.dk/~pd/ISwR.html`.

Robert Gentleman (2008), "R Programming for Bioinformatics". Chapman & Hall/CRC, Boca Raton, FL, ISBN 978-1-420-06367-7, `http://www.bioconductor.org/pub/RBioinf/`.

Stefano M. Iacus (2008), "Simulation and Inference for Stochastic Differential Equations: With R Examples". Springer, New York, ISBN 978-0-387-75838-1.

Deepayan Sarkar (2007), "Lattice: Multivariate Data Visualization with R". Springer, New York, ISBN 978-0-387-75968-5.

W. John Braun and Duncan J. Murdoch (2007), "A First Course in Statistical Programming with R". Cambridge University Press, Cambridge, ISBN 978-0521872652.

P. Murrell (2005), "R Graphics", Chapman & Hall/CRC, ISBN: 1-584-88486-X, `http://www.stat.auckland.ac.nz/~paul/RGraphics/rgraphics.html`.

William N. Venables and Brian D. Ripley (2002), "Modern Applied Statistics with S" (4th edition). Springer, ISBN 0-387-95457-0, `http://www.stats.ox.ac.uk/pub/MASS4/`.

Jose C. Pinheiro and Douglas M. Bates (2000), "Mixed-Effects Models in S and S-Plus". Springer, ISBN 0-387-98957-0.

Last, but not least, Ross' and Robert's experience in designing and implementing R is described in Ihaka & Gentleman (1996), "R: A Language for Data Analysis and Graphics", *Journal of Computational and Graphical Statistics*, **5**, 299–314.

2.8 Citing R

To cite R in publications, use

```
@Manual{,
  title        = {R: A Language and Environment for Statistical
                  Computing},
  author       = {{R Core Team}},
  organization = {R Foundation for Statistical Computing},
  address      = {Vienna, Austria},
  year         = 2014,
  url          = {http://www.R-project.org}
}
```

Citation strings (or BibTEX entries) for R and R packages can also be obtained by `citation()`.

2.9 What mailing lists exist for R?

Thanks to Martin Maechler, there are four mailing lists devoted to R.

`R-announce`
: A moderated list for major announcements about the development of R and the availability of new code.

`R-packages`
: A moderated list for announcements on the availability of new or enhanced contributed packages.

`R-help`
: The 'main' R mailing list, for discussion about problems and solutions using R, announcements (not covered by 'R-announce' and 'R-packages') about the development of R and the availability of new code.

`R-devel`
: This list is for questions and discussion about code development in R.

Please read the posting guide *before* sending anything to any mailing list.

Note in particular that R-help is intended to be comprehensible to people who want to use R to solve problems but who are not necessarily interested in or knowledgeable about programming. Questions likely to prompt discussion unintelligible to non-programmers (e.g., questions involving C or C++) should go to R-devel.

Convenient access to information on these lists, subscription, and archives is provided by the web interface at `http://stat.ethz.ch/mailman/listinfo/`. One can also subscribe (or unsubscribe) via email, e.g. to R-help by sending '`subscribe`' (or '`unsubscribe`') in the *body* of the message (not in the subject!) to `R-help-request@lists.R-project.org`.

Send email to `R-help@lists.R-project.org` to send a message to everyone on the R-help mailing list. Subscription and posting to the other lists is done analogously, with '`R-help`' replaced by '`R-announce`', '`R-packages`', and '`R-devel`', respectively. Note that

the R-announce and R-packages lists are gatewayed into R-help. Hence, you should subscribe to either of them only in case you are not subscribed to R-help.

It is recommended that you send mail to R-help rather than only to the R Core developers (who are also subscribed to the list, of course). This may save them precious time they can use for constantly improving R, and will typically also result in much quicker feedback for yourself.

Of course, in the case of bug reports it would be very helpful to have code which reliably reproduces the problem. Also, make sure that you include information on the system and version of R being used. See Chapter 9 [R Bugs], page 45 for more details.

See `http://www.R-project.org/mail.html` for more information on the R mailing lists.

The R Core Team can be reached at `R-core@lists.R-project.org` for comments and reports.

Many of the R project's mailing lists are also available via Gmane, from which they can be read with a web browser, using an NNTP news reader, or via RSS feeds. See `http://dir.gmane.org/index.php?prefix=gmane.comp.lang.r.` for the available mailing lists, and `http://www.gmane.org/rss.php` for details on RSS feeds.

2.10 What is CRAN?

The "Comprehensive R Archive Network" (CRAN) is a collection of sites which carry identical material, consisting of the R distribution(s), the contributed extensions, documentation for R, and binaries.

The CRAN master site at WU (Wirtschaftsuniversität Wien) in Austria can be found at the URL

`http://CRAN.R-project.org/`

Daily mirrors are available at URLs including

`http://cran.at.R-project.org/`	(Wirtschaftsuniversität Wien, Austria)
`http://cran.au.R-project.org/`	(University of Melbourne, Australia)
`http://cran.br.R-project.org/`	(Universidade Federal do Paraná, Brazil)
`http://cran.ch.R-project.org/`	(ETH Zürich, Switzerland)
`http://cran.dk.R-project.org/`	(dotsrc.org, Aalborg, Denmark)
`http://cran.es.R-project.org/`	(Spanish National Research Network, Madrid, Spain)
`http://cran.pt.R-project.org/`	(Universidade do Porto, Portugal)
`http://cran.uk.R-project.org/`	(U of Bristol, United Kingdom)

See `http://CRAN.R-project.org/mirrors.html` for a complete list of mirrors. Please use the CRAN site closest to you to reduce network load.

From CRAN, you can obtain the latest official release of R, daily snapshots of R (copies of the current source trees), as gzipped and bzipped tar files, a wealth of additional contributed code, as well as prebuilt binaries for various operating systems (Linux, Mac OS Classic, OS

X, and MS Windows). CRAN also provides access to documentation on R, existing mailing lists and the R Bug Tracking system.

Please always use the URL of the master site when referring to CRAN.

2.11 Can I use R for commercial purposes?

R is released under the GNU General Public License (GPL). If you have any questions regarding the legality of using R in any particular situation you should bring it up with your legal counsel. We are in no position to offer legal advice.

It is the opinion of the R Core Team that one can use R for commercial purposes (e.g., in business or in consulting). The GPL, like all Open Source licenses, permits all and any use of the package. It only restricts distribution of R or of other programs containing code from R. This is made clear in clause 6 ("No Discrimination Against Fields of Endeavor") of the Open Source Definition:

> The license must not restrict anyone from making use of the program in a specific field of endeavor. For example, it may not restrict the program from being used in a business, or from being used for genetic research.

It is also explicitly stated in clause 0 of the GPL, which says in part

> Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running the Program is not restricted, and the output from the Program is covered only if its contents constitute a work based on the Program.

Most add-on packages, including all recommended ones, also explicitly allow commercial use in this way. A few packages are restricted to "non-commercial use"; you should contact the author to clarify whether these may be used or seek the advice of your legal counsel.

None of the discussion in this section constitutes legal advice. The R Core Team does not provide legal advice under any circumstances.

2.12 Why is R named R?

The name is partly based on the (first) names of the first two R authors (Robert Gentleman and Ross Ihaka), and partly a play on the name of the Bell Labs language 'S' (see Section 3.1 [What is S?], page 12).

2.13 What is the R Foundation?

The R Foundation is a not for profit organization working in the public interest. It was founded by the members of the R Core Team in order to provide support for the R project and other innovations in statistical computing, provide a reference point for individuals, institutions or commercial enterprises that want to support or interact with the R development community, and to hold and administer the copyright of R software and documentation. See `http://www.R-project.org/foundation/` for more information.

2.14 What is R-Forge?

R-Forge (`http://R-Forge.R-project.org/`) offers a central platform for the development of R packages, R-related software and further projects. It is based on GForge offering easy

access to the best in SVN, daily built and checked packages, mailing lists, bug tracking, message boards/forums, site hosting, permanent file archival, full backups, and total web-based administration. For more information, see the R-Forge web page and Stefan Theußl and Achim Zeileis (2009), "Collaborative software development using R-Forge", *The R Journal*, **1**(1):9–14.

3 R and S

3.1 What is S?

S is a very high level language and an environment for data analysis and graphics. In 1998, the Association for Computing Machinery (ACM) presented its Software System Award to John M. Chambers, the principal designer of S, for

> the S system, which has forever altered the way people analyze, visualize, and manipulate data . . .
>
> S is an elegant, widely accepted, and enduring software system, with conceptual integrity, thanks to the insight, taste, and effort of John Chambers.

The evolution of the S language is characterized by four books by John Chambers and coauthors, which are also the primary references for S.

- Richard A. Becker and John M. Chambers (1984), "S. An Interactive Environment for Data Analysis and Graphics," Monterey: Wadsworth and Brooks/Cole.

  This is also referred to as the "*Brown Book*", and of historical interest only.
- Richard A. Becker, John M. Chambers and Allan R. Wilks (1988), "The New S Language," London: Chapman & Hall.

  This book is often called the "*Blue Book*", and introduced what is now known as S version 2.
- John M. Chambers and Trevor J. Hastie (1992), "Statistical Models in S," London: Chapman & Hall.

  This is also called the "*White Book*", and introduced S version 3, which added structures to facilitate statistical modeling in S.
- John M. Chambers (1998), "Programming with Data," New York: Springer, ISBN 0-387-98503-4 (`http://cm.bell-labs.com/cm/ms/departments/sia/Sbook/`).

  This "*Green Book*" describes version 4 of S, a major revision of S designed by John Chambers to improve its usefulness at every stage of the programming process.

See `http://cm.bell-labs.com/cm/ms/departments/sia/S/history.html` for further information on "Stages in the Evolution of S".

There is a huge amount of user-contributed code for S, available at the S Repository at CMU.

3.2 What is S-PLUS?

S-PLUS is a value-added version of S currently sold by TIBCO Software Inc as 'TIBCO Spotfire S+'. See `http://en.wikipedia.org/wiki/S-PLUS` for more information.

3.3 What are the differences between R and S?

We can regard S as a language with three current implementations or "engines", the "old S engine" (S version 3; S-PLUS 3.x and 4.x), the "new S engine" (S version 4; S-PLUS 5.x and above), and R. Given this understanding, asking for "the differences between R and S" really amounts to asking for the specifics of the R implementation of the S language, i.e., the difference between the R and S *engines*.

For the remainder of this section, "S" refers to the S engines and not the S language.

3.3.1 Lexical scoping

Contrary to other implementations of the S language, R has adopted an evaluation model in which nested function definitions are lexically scoped. This is analogous to the evaluation model in Scheme.

This difference becomes manifest when *free* variables occur in a function. Free variables are those which are neither formal parameters (occurring in the argument list of the function) nor local variables (created by assigning to them in the body of the function). In S, the values of free variables are determined by a set of global variables (similar to C, there is only local and global scope). In R, they are determined by the environment in which the function was created.

Consider the following function:

```
cube <- function(n) {
  sq <- function() n * n
  n * sq()
}
```

Under S, `sq()` does not "know" about the variable `n` unless it is defined globally:

```
S> cube(2)
Error in sq():  Object "n" not found
Dumped
S> n <- 3
S> cube(2)
[1] 18
```

In R, the "environment" created when `cube()` was invoked is also looked in:

```
R> cube(2)
[1] 8
```

As a more "interesting" real-world problem, suppose you want to write a function which returns the density function of the r-th order statistic from a sample of size n from a (continuous) distribution. For simplicity, we shall use both the cdf and pdf of the distribution as explicit arguments. (Example compiled from various postings by Luke Tierney.)

The S-PLUS documentation for `call()` basically suggests the following:

```
dorder <- function(n, r, pfun, dfun) {
  f <- function(x) NULL
  con <- round(exp(lgamma(n + 1) - lgamma(r) - lgamma(n - r + 1)))
  PF <- call(substitute(pfun), as.name("x"))
  DF <- call(substitute(dfun), as.name("x"))
  f[[length(f)]] <-
    call("*", con,
         call("*", call("^", PF, r - 1),
              call("*", call("^", call("-", 1, PF), n - r),
                   DF)))
  f
}
```

Rather tricky, isn't it? The code uses the fact that in S, functions are just lists of special mode with the function body as the last argument, and hence does not work in R (one could make the idea work, though).

A version which makes heavy use of `substitute()` and seems to work under both S and R is

```
dorder <- function(n, r, pfun, dfun) {
  con <- round(exp(lgamma(n + 1) - lgamma(r) - lgamma(n - r + 1)))
  eval(substitute(function(x) K * PF(x)^a * (1 - PF(x))^b * DF(x),
                  list(PF = substitute(pfun), DF = substitute(dfun),
                       a = r - 1, b = n - r, K = con)))
}
```

(the `eval()` is not needed in S).

However, in R there is a much easier solution:

```
dorder <- function(n, r, pfun, dfun) {
  con <- round(exp(lgamma(n + 1) - lgamma(r) - lgamma(n - r + 1)))
  function(x) {
    con * pfun(x)^(r - 1) * (1 - pfun(x))^(n - r) * dfun(x)
  }
}
```

This seems to be the "natural" implementation, and it works because the free variables in the returned function can be looked up in the defining environment (this is lexical scope).

Note that what you really need is the function *closure*, i.e., the body along with all variable bindings needed for evaluating it. Since in the above version, the free variables in the value function are not modified, you can actually use it in S as well if you abstract out the closure operation into a function `MC()` (for "make closure"):

```
dorder <- function(n, r, pfun, dfun) {
  con <- round(exp(lgamma(n + 1) - lgamma(r) - lgamma(n - r + 1)))
  MC(function(x) {
       con * pfun(x)^(r - 1) * (1 - pfun(x))^(n - r) * dfun(x)
     },
     list(con = con, pfun = pfun, dfun = dfun, r = r, n = n))
}
```

Given the appropriate definitions of the closure operator, this works in both R and S, and is much "cleaner" than a substitute/eval solution (or one which overrules the default scoping rules by using explicit access to evaluation frames, as is of course possible in both R and S).

For R, `MC()` simply is

```
MC <- function(f, env) f
```

(lexical scope!), a version for S is

```
MC <- function(f, env = NULL) {
  env <- as.list(env)
  if (mode(f) != "function")
    stop(paste("not a function:", f))
  if (length(env) > 0 && any(names(env) == ""))
    stop(paste("not all arguments are named:", env))
  fargs <- if(length(f) > 1) f[1:(length(f) - 1)] else NULL
  fargs <- c(fargs, env)
  if (any(duplicated(names(fargs))))
    stop(paste("duplicated arguments:", paste(names(fargs)),
         collapse = ", "))
  fbody <- f[length(f)]
  cf <- c(fargs, fbody)
  mode(cf) <- "function"
  return(cf)
}
```

Similarly, most optimization (or zero-finding) routines need some arguments to be optimized over and have other parameters that depend on the data but are fixed with respect to optimization. With R scoping rules, this is a trivial problem; simply make up the function with the required definitions in the same environment and scoping takes care of it. With S, one solution is to add an extra parameter to the function and to the optimizer to pass in these extras, which however can only work if the optimizer supports this.

Nested lexically scoped functions allow using function closures and maintaining local state. A simple example (taken from Abelson and Sussman) is obtained by typing `demo("scoping")` at the R prompt. Further information is provided in the standard R reference "R: A Language for Data Analysis and Graphics" (see Section 2.7 [What documentation exists for R?], page 6) and in Robert Gentleman and Ross Ihaka (2000), "Lexical Scope and Statistical Computing", *Journal of Computational and Graphical Statistics*, **9**, 491–508.

Nested lexically scoped functions also imply a further major difference. Whereas S stores all objects as separate files in a directory somewhere (usually `.Data` under the current directory), R does not. All objects in R are stored internally. When R is started up it grabs a piece of memory and uses it to store the objects. R performs its own memory management of this piece of memory, growing and shrinking its size as needed. Having everything in memory is necessary because it is not really possible to externally maintain all relevant "environments" of symbol/value pairs. This difference also seems to make R *faster* than S.

The down side is that if R crashes you will lose all the work for the current session. Saving and restoring the memory "images" (the functions and data stored in R's internal memory at any time) can be a bit slow, especially if they are big. In S this does not happen, because everything is saved in disk files and if you crash nothing is likely to happen to them. (In fact, one might conjecture that the S developers felt that the price of changing their approach to persistent storage just to accommodate lexical scope was far too expensive.) Hence, when doing important work, you might consider saving often (see Section 7.2 [How can I save my workspace?], page 29) to safeguard against possible crashes. Other possibilities are logging your sessions, or have your R commands stored in text files which can be read in using `source()`.

> **Note:** If you run R from within Emacs (see Chapter 6 [R and Emacs], page 27), you can save the contents of the interaction buffer to a file and conveniently manipulate it using `ess-transcript-mode`, as well as save source copies of all functions and data used.

3.3.2 Models

There are some differences in the modeling code, such as

- Whereas in S, you would use `lm(y ~ x^3)` to regress `y` on `x^3`, in R, you have to insulate powers of numeric vectors (using `I()`), i.e., you have to use `lm(y ~ I(x^3))`.
- The glm family objects are implemented differently in R and S. The same functionality is available but the components have different names.
- Option `na.action` is set to `"na.omit"` by default in R, but not set in S.
- Terms objects are stored differently. In S a terms object is an expression with attributes, in R it is a formula with attributes. The attributes have the same names but are mostly stored differently.
- Finally, in R `y ~ x + 0` is an alternative to `y ~ x - 1` for specifying a model with no intercept. Models with no parameters at all can be specified by `y ~ 0`.

3.3.3 Others

Apart from lexical scoping and its implications, R follows the S language definition in the Blue and White Books as much as possible, and hence really is an "implementation" of S. There are some intentional differences where the behavior of S is considered "not clean". In general, the rationale is that R should help you detect programming errors, while at the same time being as compatible as possible with S.

Some known differences are the following.

- In R, if `x` is a list, then `x[i] <- NULL` and `x[[i]] <- NULL` remove the specified elements from `x`. The first of these is incompatible with S, where it is a no-op. (Note that you can set elements to `NULL` using `x[i] <- list(NULL)`.)
- In S, the functions named `.First` and `.Last` in the `.Data` directory can be used for customizing, as they are executed at the very beginning and end of a session, respectively.

  In R, the startup mechanism is as follows. Unless `--no-environ` was given on the command line, R searches for site and user files to process for setting environment variables. Then, R searches for a site-wide startup profile unless the command line option `--no-site-file` was given. This code is loaded in package **base**. Then, unless `--no-init-file` was given, R searches for a user profile file, and sources it into the user workspace. It then loads a saved image of the user workspace from `.RData` in case there is one (unless `--no-restore-data` or `--no-restore` were specified). Next, a function `.First()` is run if found on the search path. Finally, function `.First.sys` in the **base** package is run. When terminating an R session, by default a function `.Last` is run if found on the search path, followed by `.Last.sys`. If needed, the functions `.First()` and `.Last()` should be defined in the appropriate startup profiles. See the help pages for `.First` and `.Last` for more details.
- In R, `T` and `F` are just variables being set to `TRUE` and `FALSE`, respectively, but are not reserved words as in S and hence can be overwritten by the user. (This helps e.g. when

you have factors with levels `"T"` or `"F"`.) Hence, when writing code you should always use `TRUE` and `FALSE`.

- In R, `dyn.load()` can only load *shared objects*, as created for example by *`R CMD SHLIB`*.
- In R, `attach()` currently only works for lists and data frames, but not for directories. (In fact, `attach()` also works for R data files created with `save()`, which is analogous to attaching directories in S.) Also, you cannot attach at position 1.
- Categories do not exist in R, and never will as they are deprecated now in S. Use factors instead.
- In R, `For()` loops are not necessary and hence not supported.
- In R, `assign()` uses the argument `envir=` rather than `where=` as in S.
- The random number generators are different, and the seeds have different length.
- R passes integer objects to C as `int *` rather than `long *` as in S.
- R has no single precision storage mode. However, as of version 0.65.1, there is a single precision interface to C/FORTRAN subroutines.
- By default, `ls()` returns the names of the objects in the current (under R) and global (under S) environment, respectively. For example, given

  ```
  x <- 1; fun <- function() {y <- 1; ls()}
  ```

  then `fun()` returns `"y"` in R and `"x"` (together with the rest of the global environment) in S.
- R allows for zero-extent matrices (and arrays, i.e., some elements of the `dim` attribute vector can be 0). This has been determined a useful feature as it helps reducing the need for special-case tests for empty subsets. For example, if `x` is a matrix, `x[, FALSE]` is not `NULL` but a "matrix" with 0 columns. Hence, such objects need to be tested for by checking whether their `length()` is zero (which works in both R and S), and not using `is.null()`.
- Named vectors are considered vectors in R but not in S (e.g., `is.vector(c(a = 1:3))` returns `FALSE` in S and `TRUE` in R).
- Data frames are not considered as matrices in R (i.e., if `DF` is a data frame, then `is.matrix(DF)` returns `FALSE` in R and `TRUE` in S).
- R by default uses treatment contrasts in the unordered case, whereas S uses the Helmert ones. This is a deliberate difference reflecting the opinion that treatment contrasts are more natural.
- In R, the argument of a replacement function which corresponds to the right hand side must be named '`value`'. E.g., `f(a) <- b` is evaluated as `a <- "f<-"(a, value = b)`. S always takes the last argument, irrespective of its name.
- In S, `substitute()` searches for names for substitution in the given expression in three places: the actual and the default arguments of the matching call, and the local frame (in that order). R looks in the local frame only, with the special rule to use a "promise" if a variable is not evaluated. Since the local frame is initialized with the actual arguments or the default expressions, this is usually equivalent to S, until assignment takes place.
- In S, the index variable in a `for()` loop is local to the inside of the loop. In R it is local to the environment where the `for()` statement is executed.

- In S, `tapply(simplify=TRUE)` returns a vector where R returns a one-dimensional array (which can have named dimnames).
- In S(-PLUS) the C locale is used, whereas in R the current operating system locale is used for determining which characters are alphanumeric and how they are sorted. This affects the set of valid names for R objects (for example accented chars may be allowed in R) and ordering in sorts and comparisons (such as whether `"aA" < "Bb"` is true or false). From version 1.2.0 the locale can be (re-)set in R by the `Sys.setlocale()` function.
- In S, `missing(`*arg*`)` remains `TRUE` if *arg* is subsequently modified; in R it doesn't.
- From R version 1.3.0, `data.frame` strips `I()` when creating (column) names.
- In R, the string `"NA"` is not treated as a missing value in a character variable. Use `as.character(NA)` to create a missing character value.
- R disallows repeated formal arguments in function calls.
- In S, `dump()`, `dput()` and `deparse()` are essentially different interfaces to the same code. In R from version 2.0.0, this is only true if the same `control` argument is used, but by default it is not. By default `dump()` tries to write code that will evaluate to reproduce the object, whereas `dput()` and `deparse()` default to options for producing deparsed code that is readable.
- In R, indexing a vector, matrix, array or data frame with `[` using a character vector index looks only for exact matches (whereas `[[` and `$` allow partial matches). In S, `[` allows partial matches.
- S has a two-argument version of `atan` and no `atan2`. A call in S such as `atan(x1, x2)` is equivalent to R's `atan2(x1, x2)`. However, beware of named arguments since S's `atan(x = a, y = b)` is equivalent to R's `atan2(y = a, x = b)` with the meanings of `x` and `y` interchanged. (R used to have undocumented support for a two-argument `atan` with positional arguments, but this has been withdrawn to avoid further confusion.)
- Numeric constants with no fractional and exponent (i.e., only integer) part are taken as integer in S-PLUS 6.x or later, but as double in R.

There are also differences which are not intentional, and result from missing or incorrect code in R. The developers would appreciate hearing about any deficiencies you may find (in a written report fully documenting the difference as you see it). Of course, it would be useful if you were to implement the change yourself and make sure it works.

3.4 Is there anything R can do that S-PLUS cannot?

Since almost anything you can do in R has source code that you could port to S-PLUS with little effort there will never be much you can do in R that you couldn't do in S-PLUS if you wanted to. (Note that using lexical scoping may simplify matters considerably, though.)

R offers several graphics features that S-PLUS does not, such as finer handling of line types, more convenient color handling (via palettes), gamma correction for color, and, most importantly, mathematical annotation in plot texts, via input expressions reminiscent of TEX constructs. See the help page for `plotmath`, which features an impressive on-line example. More details can be found in Paul Murrell and Ross Ihaka (2000), "An Approach to Providing Mathematical Annotation in Plots", *Journal of Computational and Graphical Statistics*, **9**, 582–599.

3.5 What is R-plus?

For a very long time, there was no such thing.

XLSolutions Corporation is currently beta testing a commercially supported version of R named R+ (read R plus).

REvolution Computing has released REvolution R, an enterprise-class statistical analysis system based on R, suitable for deployment in professional, commercial and regulated environments.

See also `http://en.wikipedia.org/wiki/R_programming_language#Commercialized_versions_of_R` for pointers to commercialized versions of R.

4 R Web Interfaces

Rweb is developed and maintained by Jeff Banfield. The Rweb Home Page provides access to all three versions of Rweb—a simple text entry form that returns output and graphs, a more sophisticated JavaScript version that provides a multiple window environment, and a set of point and click modules that are useful for introductory statistics courses and require no knowledge of the R language. All of the Rweb versions can analyze Web accessible datasets if a URL is provided.

The paper "Rweb: Web-based Statistical Analysis", providing a detailed explanation of the different versions of Rweb and an overview of how Rweb works, was published in the Journal of Statistical Software (http://www.jstatsoft.org/v04/i01/).

Ulf Bartel has developed **R-Online**, a simple on-line programming environment for R which intends to make the first steps in statistical programming with R (especially with time series) as easy as possible. There is no need for a local installation since the only requirement for the user is a JavaScript capable browser. See http://www.osvisions.com/r_online/ for more information.

Rcgi is a CGI WWW interface to R by MJ Ray. It had the ability to use "embedded code": you could mix user input and code, allowing the HTML author to do anything from load in data sets to enter most of the commands for users without writing CGI scripts. Graphical output was possible in PostScript or GIF formats and the executed code was presented to the user for revision. However, it is not clear if the project is still active. Currently, a modified version of **Rcgi** by Mai Zhou (actually, two versions: one with (bitmap) graphics and one without) as well as the original code are available from http://www.ms.uky.edu/~statweb/.

There are many additional examples of web interfaces to R which basically allow to submit R code to a remote server, see for example the collection of links available from http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/StatCompCourse.

David Firth has written **CGIwithR**, an R add-on package available from CRAN. It provides some simple extensions to R to facilitate running R scripts through the CGI interface to a web server, and allows submission of data using both GET and POST methods. It is easily installed using Apache under Linux and in principle should run on any platform that supports R and a web server provided that the installer has the necessary security permissions. David's paper "CGIwithR: Facilities for Processing Web Forms Using R" was published in the Journal of Statistical Software (http://www.jstatsoft.org/v08/i10/). The package is now maintained by Duncan Temple Lang and has a web page at http://www.omegahat.org/CGIwithR/.

Jeff Horner is working on the R/Apache Integration Project which embeds the R interpreter inside Apache 2 (and beyond). A tutorial and presentation are available from the project web page at http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/RApacheProject.

Rserve is a project actively developed by Simon Urbanek. It implements a TCP/IP server which allows other programs to use facilities of R. Clients are available from the web site for Java and C++ (and could be written for other languages that support TCP/IP sockets).

Two projects use PHP to provide a web interface to R. R_PHP_Online by Steve Chen (though it is unclear if this project is still active) is somewhat similar to the above Rcgi and Rweb. R-php is actively developed by Alfredo Pontillo and Angelo Mineo and provides both a web interface to R and a set of pre-specified analyses that need no R code input.

webbioc is "an integrated web interface for doing microarray analysis using several of the Bioconductor packages" and is designed to be installed at local sites as a shared computing resource.

Rwui is a web application to create user-friendly web interfaces for R scripts. All code for the web interface is created automatically. There is no need for the user to do any extra scripting or learn any new scripting techniques.

The **R.rsp** package by Henrik Bengtsson introduces "R Server Pages". Analogous to Java Server Pages, an R server page is typically HTML with embedded R code that gets evaluated when the page is requested. The package includes an internal cross-platform HTTP server implemented in Tcl, so provides a good framework for including web-based user interfaces in packages. The approach is similar to the use of the **brew** package with Rapache with the advantage of cross-platform support and easy installation.

The **Rook** package by Jeffrey Horner provides a web server interface borrowing heavily from Ruby's Rack project.

Finally, Concerto is a user friendly open-source Web Interface to R developed at the Psychometrics Centre of Cambridge University. It was designed as an online platform to design and run Computerized Adaptive Tests, but can be also used as a general-purpose R Web Interface. It allows R users with no programming or web designing background to quickly develop flexible and powerful online applications, websites, and psychometrics tests. To maximize its reliability, security, and performance, Concerto relies on the popular and reliable open-source elements such as MySQL server (exchange and storage of the data), Rstudio (R code designing and testing, file management), CKEditor (HTML Layer design), and PHP.

See `http://rwiki.sciviews.org/doku.php?id=faq-r#web_interfaces` for additional information.

5 R Add-On Packages

5.1 Which add-on packages exist for R?

5.1.1 Add-on packages in R

The R distribution comes with the following packages:

base Base R functions (and datasets before R 2.0.0).

compiler R byte code compiler (added in R 2.13.0).

datasets Base R datasets (added in R 2.0.0).

grDevices Graphics devices for base and grid graphics (added in R 2.0.0).

graphics R functions for base graphics.

grid A rewrite of the graphics layout capabilities, plus some support for interaction.

methods Formally defined methods and classes for R objects, plus other programming tools, as described in the Green Book.

parallel Support for parallel computation, including by forking and by sockets, and random-number generation (added in R 2.14.0).

splines Regression spline functions and classes.

stats R statistical functions.

stats4 Statistical functions using S4 classes.

tcltk Interface and language bindings to Tcl/Tk GUI elements.

tools Tools for package development and administration.

utils R utility functions.

These "base packages" were substantially reorganized in R 1.9.0. The former **base** was split into the four packages **base**, **graphics**, **stats**, and **utils**. Packages **ctest**, **eda**, **modreg**, **mva**, **nls**, **stepfun** and **ts** were merged into **stats**, package **lqs** returned to the recommended package **MASS**, and package **mle** moved to **stats4**.

5.1.2 Add-on packages from CRAN

The CRAN `src/contrib` area contains a wealth of add-on packages, including the following *recommended* packages which are to be included in all binary distributions of R.

KernSmooth
Functions for kernel smoothing (and density estimation) corresponding to the book "Kernel Smoothing" by M. P. Wand and M. C. Jones, 1995.

MASS Functions and datasets from the main package of Venables and Ripley, "Modern Applied Statistics with S". (Contained in the `VR` bundle for R versions prior to 2.10.0.)

Matrix A Matrix package. (Recommended for R 2.9.0 or later.)

boot Functions and datasets for bootstrapping from the book "Bootstrap Methods and Their Applications" by A. C. Davison and D. V. Hinkley, 1997, Cambridge University Press.

class Functions for classification (*k*-nearest neighbor and LVQ). (Contained in the `VR` bundle for R versions prior to 2.10.0.)

cluster Functions for cluster analysis.

codetools Code analysis tools. (Recommended for R 2.5.0 or later.)

foreign Functions for reading and writing data stored by statistical software like Minitab, S, SAS, SPSS, Stata, Systat, etc.

lattice Lattice graphics, an implementation of Trellis Graphics functions.

mgcv Routines for GAMs and other generalized ridge regression problems with multiple smoothing parameter selection by GCV or UBRE.

nlme Fit and compare Gaussian linear and nonlinear mixed-effects models.

nnet Software for single hidden layer perceptrons ("feed-forward neural networks"), and for multinomial log-linear models. (Contained in the `VR` bundle for R versions prior to 2.10.0.)

rpart Recursive PARTitioning and regression trees.

spatial Functions for kriging and point pattern analysis from "Modern Applied Statistics with S" by W. Venables and B. Ripley. (Contained in the `VR` bundle for R versions prior to 2.10.0.)

survival Functions for survival analysis, including penalized likelihood.

See the CRAN contributed packages page for more information.

Many of these packages are categorized into CRAN Task Views, allowing to browse packages by topic and providing tools to automatically install all packages for special areas of interest.

Some CRAN packages that do not build out of the box on Windows, require additional software, or are shipping third party libraries for Windows cannot be made available on CRAN in form of a Windows binary packages. Nevertheless, some of these packages are available at the "CRAN extras" repository at `http://www.stats.ox.ac.uk/pub/RWin/` kindly provided by Brian D. Ripley. Note that this repository is a default repository for recent versions of R for Windows.

5.1.3 Add-on packages from Omegahat

The Omega Project for Statistical Computing provides a variety of open-source software for statistical applications, with special emphasis on web-based software, Java, the Java virtual machine, and distributed computing. A CRAN style R package repository is available via `http://www.omegahat.org/R/`. See `http://www.omegahat.org/` for information on most R packages available from the Omega project.

5.1.4 Add-on packages from Bioconductor

Bioconductor is an open source and open development software project for the analysis and comprehension of genomic data. Most Bioconductor components are distributed as R add-on packages. Initially most of the Bioconductor software packages focused primarily on DNA microarray data analysis. As the project has matured, the functional scope of the software packages broadened to include the analysis of all types of genomic data, such as SAGE, sequence, or SNP data. In addition, there are metadata (annotation, CDF and probe) and experiment data packages. See `http://www.bioconductor.org/download/` for available packages and a complete taxonomy via BioC Views.

5.1.5 Other add-on packages

Many more packages are available from places other than the three default repositories discussed above (CRAN, Bioconductor and Omegahat). In particular, R-Forge provides a CRAN style repository at `http://R-Forge.R-project.org/`.

More code has been posted to the R-help mailing list, and can be obtained from the mailing list archive.

5.2 How can add-on packages be installed?

(Unix-like only.) The add-on packages on CRAN come as gzipped tar files named *`pkg_version`*`.tar.gz`, which may in fact be "bundles" containing more than one package. Let *path* be the path to such a package file. Provided that `tar` and `gzip` are available on your system, type

```
$ R CMD INSTALL path/pkg_version.tar.gz
```

at the shell prompt to install to the library tree rooted at the first directory in your library search path (see the help page for `.libPaths()` for details on how the search path is determined).

To install to another tree (e.g., your private one), use

```
$ R CMD INSTALL -l lib path/pkg_version.tar.gz
```

where *lib* gives the path to the library tree to install to.

Even more conveniently, you can install and automatically update packages from within R if you have access to repositories such as CRAN. See the help page for `available.packages()` for more information.

5.3 How can add-on packages be used?

To find out which additional packages are available on your system, type

```
library()
```

at the R prompt.

This produces something like

```
Packages in '/home/me/lib/R':

mystuff       My own R functions, nicely packaged but not documented

Packages in '/usr/local/lib/R/library':

KernSmooth    Functions for kernel smoothing for Wand & Jones (1995)
MASS          Main Package of Venables and Ripley's MASS
Matrix        Sparse and Dense Matrix Classes and Methods
base          The R Base package
boot          Bootstrap R (S-Plus) Functions (Canty)
class         Functions for Classification
cluster       Functions for clustering (by Rousseeuw et al.)
codetools     Code Analysis Tools for R
datasets      The R Datasets Package
foreign       Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat,
              dBase, ...
grDevices     The R Graphics Devices and Support for Colours and Fonts
graphics      The R Graphics Package
grid          The Grid Graphics Package
lattice       Lattice Graphics
methods       Formal Methods and Classes
mgcv          GAMs with GCV/AIC/REML smoothness estimation and GAMMs
              by PQL
nlme          Linear and Nonlinear Mixed Effects Models
nnet          Feed-forward Neural Networks and Multinomial Log-Linear
              Models
rpart         Recursive Partitioning
spatial       Functions for Kriging and Point Pattern Analysis
splines       Regression Spline Functions and Classes
stats         The R Stats Package
stats4        Statistical functions using S4 Classes
survival      Survival analysis, including penalised likelihood
tcltk         Tcl/Tk Interface
tools         Tools for Package Development
utils         The R Utils Package
```

You can "load" the installed package *pkg* by

```
library(pkg)
```

You can then find out which functions it provides by typing one of

```
library(help = pkg)
help(package = pkg)
```

You can unload the loaded package *pkg* by

```
detach("package:pkg", unload = TRUE)
```

(where `unload = TRUE` is needed only for packages with a namespace, see `?unload`).

5.4 How can add-on packages be removed?

Use

```
$ R CMD REMOVE pkg_1 ... pkg_n
```

to remove the packages *pkg_1*, ..., *pkg_n* from the library tree rooted at the first directory given in `R_LIBS` if this is set and non-null, and from the default library otherwise. (Versions of R prior to 1.3.0 removed from the default library by default.)

To remove from library *lib*, do

```
$ R CMD REMOVE -l lib pkg_1 ... pkg_n
```

5.5 How can I create an R package?

A package consists of a subdirectory containing a file `DESCRIPTION` and the subdirectories `R`, `data`, `demo`, `exec`, `inst`, `man`, `po`, `src`, and `tests` (some of which can be missing). The package subdirectory may also contain files `INDEX`, `NAMESPACE`, `configure`, `cleanup`, `LICENSE`, `LICENCE`, `COPYING` and `NEWS`.

See Section "Creating R packages" in *Writing R Extensions*, for details.

R version 1.3.0 has added the function `package.skeleton()` which will set up directories, save data and code, and create skeleton help files for a set of R functions and datasets.

See Section 2.10 [What is CRAN?], page 9, for information on uploading a package to CRAN.

5.6 How can I contribute to R?

R is in active development and there is always a risk of bugs creeping in. Also, the developers do not have access to all possible machines capable of running R. So, simply using it and communicating problems is certainly of great value.

The R Developer Page acts as an intermediate repository for more or less finalized ideas and plans for the R statistical system. It contains (pointers to) TODO lists, RFCs, various other writeups, ideas lists, and SVN miscellanea.

6 R and Emacs

6.1 Is there Emacs support for R?

There is an Emacs package called ESS ("Emacs Speaks Statistics") which provides a standard interface between statistical programs and statistical processes. It is intended to provide assistance for interactive statistical programming and data analysis. Languages supported include: S dialects (R, S 3/4, and S-PLUS 3.x/4.x/5.x/6.x/7.x), LispStat dialects (XLispStat, ViSta), SAS, Stata, and BUGS.

ESS grew out of the need for bug fixes and extensions to S-mode 4.8 (which was a GNU Emacs interface to S/S-PLUS version 3 only). The current set of developers desired support for XEmacs, R, S4, and MS Windows. In addition, with new modes being developed for R, Stata, and SAS, it was felt that a unifying interface and framework for the user interface would benefit both the user and the developer, by helping both groups conform to standard Emacs usage. The end result is an increase in efficiency for statistical programming and data analysis, over the usual tools.

R support contains code for editing R source code (syntactic indentation and highlighting of source code, partial evaluations of code, loading and error-checking of code, and source code revision maintenance) and documentation (syntactic indentation and highlighting of source code, sending examples to running ESS process, and previewing), interacting with an inferior R process from within Emacs (command-line editing, searchable command history, command-line completion of R object and file names, quick access to object and search lists, transcript recording, and an interface to the help system), and transcript manipulation (recording and saving transcript files, manipulating and editing saved transcripts, and re-evaluating commands from transcript files).

The latest stable version of ESS is available via CRAN or the ESS web page. The HTML version of the documentation can be found at `http://stat.ethz.ch/ESS/`.

ESS comes with detailed installation instructions.

For help with ESS, send email to `ESS-help@stat.math.ethz.ch`.

Please send bug reports and suggestions on ESS to `ESS-bugs@stat.math.ethz.ch`. The easiest way to do this from is within Emacs by typing *`M-x ess-submit-bug-report`* or using the [ESS] or [iESS] pulldown menus.

6.2 Should I run R from within Emacs?

Yes, *definitely*. Inferior R mode provides a readline/history mechanism, object name completion, and syntax-based highlighting of the interaction buffer using Font Lock mode, as well as a very convenient interface to the R help system.

Of course, it also integrates nicely with the mechanisms for editing R source using Emacs. One can write code in one Emacs buffer and send whole or parts of it for execution to R; this is helpful for both data analysis and programming. One can also seamlessly integrate with a revision control system, in order to maintain a log of changes in your programs and data, as well as to allow for the retrieval of past versions of the code.

In addition, it allows you to keep a record of your session, which can also be used for error recovery through the use of the transcript mode.

To specify command line arguments for the inferior R process, use *C-u M-x R* for starting R.

6.3 Debugging R from within Emacs

To debug R "from within Emacs", there are several possibilities. To use the Emacs GUD (Grand Unified Debugger) library with the recommended debugger GDB, type *M-x gdb* and give the path to the R *binary* as argument. At the `gdb` prompt, set `R_HOME` and other environment variables as needed (using e.g. *set env R_HOME /path/to/R/*, but see also below), and start the binary with the desired arguments (e.g., *run --quiet*).

If you have ESS, you can do *C-u M-x R RET - d SPC g d b RET* to start an inferior R process with arguments `-d gdb`.

A third option is to start an inferior R process via ESS (*M-x R*) and then start GUD (*M-x gdb*) giving the R binary (using its full path name) as the program to debug. Use the program `ps` to find the process number of the currently running R process then use the `attach` command in gdb to attach it to that process. One advantage of this method is that you have separate `*R*` and `*gud-gdb*` windows. Within the `*R*` window you have all the ESS facilities, such as object-name completion, that we know and love.

When using GUD mode for debugging from within Emacs, you may find it most convenient to use the directory with your code in it as the current working directory and then make a symbolic link from that directory to the R binary. That way `.gdbinit` can stay in the directory with the code and be used to set up the environment and the search paths for the source, e.g. as follows:

```
set env R_HOME /opt/R
set env R_PAPERSIZE letter
set env R_PRINTCMD lpr
dir /opt/R/src/appl
dir /opt/R/src/main
dir /opt/R/src/nmath
dir /opt/R/src/unix
```

7 R Miscellanea

7.1 How can I set components of a list to NULL?

You can use

```
x[i] <- list(NULL)
```

to set component `i` of the list `x` to `NULL`, similarly for named components. Do not set `x[i]` or `x[[i]]` to `NULL`, because this will remove the corresponding component from the list.

For dropping the row names of a matrix `x`, it may be easier to use `rownames(x) <- NULL`, similarly for column names.

7.2 How can I save my workspace?

`save.image()` saves the objects in the user's `.GlobalEnv` to the file `.RData` in the R startup directory. (This is also what happens after *`q("yes")`*.) Using `save.image(`*`file`*`)` one can save the image under a different name.

7.3 How can I clean up my workspace?

To remove all objects in the currently active environment (typically `.GlobalEnv`), you can do

```
rm(list = ls(all = TRUE))
```

(Without `all = TRUE`, only the objects with names not starting with a '`.`' are removed.)

7.4 How can I get eval() and D() to work?

Strange things will happen if you use `eval(print(x), envir = e)` or `D(x^2, "x")`. The first one will either tell you that `"x"` is not found, or print the value of the wrong `x`. The other one will likely return zero if `x` exists, and an error otherwise.

This is because in both cases, the first argument is evaluated in the calling environment first. The result (which should be an object of mode `"expression"` or `"call"`) is then evaluated or differentiated. What you (most likely) really want is obtained by "quoting" the first argument upon surrounding it with `expression()`. For example,

```
R> D(expression(x^2), "x")
2 * x
```

Although this behavior may initially seem to be rather strange, it is perfectly logical. The "intuitive" behavior could easily be implemented, but problems would arise whenever the expression is contained in a variable, passed as a parameter, or is the result of a function call. Consider for instance the semantics in cases like

```
D2 <- function(e, n) D(D(e, n), n)
```

or

```
g <- function(y) eval(substitute(y), sys.frame(sys.parent(n = 2)))
g(a * b)
```

See the help page for `deriv()` for more examples.

7.5 Why do my matrices lose dimensions?

When a matrix with a single row or column is created by a subscripting operation, e.g., `row <- mat[2, ]`, it is by default turned into a vector. In a similar way if an array with dimension, say, 2 x 3 x 1 x 4 is created by subscripting it will be coerced into a 2 x 3 x 4 array, losing the unnecessary dimension. After much discussion this has been determined to be a *feature*.

To prevent this happening, add the option `drop = FALSE` to the subscripting. For example,

```
rowmatrix <- mat[2, , drop = FALSE]  # creates a row matrix
colmatrix <- mat[, 2, drop = FALSE]  # creates a column matrix
a <- b[1, 1, 1, drop = FALSE]        # creates a 1 x 1 x 1 array
```

The `drop = FALSE` option should be used defensively when programming. For example, the statement

```
somerows <- mat[index, ]
```

will return a vector rather than a matrix if `index` happens to have length 1, causing errors later in the code. It should probably be rewritten as

```
somerows <- mat[index, , drop = FALSE]
```

7.6 How does autoloading work?

R has a special environment called `.AutoloadEnv`. Using *`autoload(name, pkg)`*, where *name* and *pkg* are strings giving the names of an object and the package containing it, stores some information in this environment. When R tries to evaluate *name*, it loads the corresponding package *pkg* and reevaluates *name* in the new package's environment.

Using this mechanism makes R behave as if the package was loaded, but does not occupy memory (yet).

See the help page for `autoload()` for a very nice example.

7.7 How should I set options?

The function `options()` allows setting and examining a variety of global "options" which affect the way in which R computes and displays its results. The variable `.Options` holds the current values of these options, but should never directly be assigned to unless you want to drive yourself crazy—simply pretend that it is a "read-only" variable.

For example, given

```
test1 <- function(x = pi, dig = 3) {
  oo <- options(digits = dig); on.exit(options(oo));
  cat(.Options$digits, x, "\n")
}
test2 <- function(x = pi, dig = 3) {
  .Options$digits <- dig
  cat(.Options$digits, x, "\n")
}
```

we obtain:

```
R> test1()
3 3.14
R> test2()
3 3.141593
```

What is really used is the *global* value of `.Options`, and using *`options(OPT = VAL)`* correctly updates it. Local copies of `.Options`, either in `.GlobalEnv` or in a function environment (frame), are just silently disregarded.

7.8 How do file names work in Windows?

As R uses C-style string handling, '\' is treated as an escape character, so that for example one can enter a newline as '`\n`'. When you really need a '\', you have to escape it with another '\'.

Thus, in filenames use something like `"c:\\data\\money.dat"`. You can also replace '\' by '/' (`"c:/data/money.dat"`).

7.9 Why does plotting give a color allocation error?

On an X11 device, plotting sometimes, e.g., when running `demo("image")`, results in "Error: color allocation error". This is an X problem, and only indirectly related to R. It occurs when applications started prior to R have used all the available colors. (How many colors are available depends on the X configuration; sometimes only 256 colors can be used.)

One application which is notorious for "eating" colors is Netscape. If the problem occurs when Netscape is running, try (re)starting it with either the `-no-install` (to use the default colormap) or the `-install` (to install a private colormap) option.

You could also set the `colortype` of `X11()` to `"pseudo.cube"` rather than the default `"pseudo"`. See the help page for `X11()` for more information.

7.10 How do I convert factors to numeric?

It may happen that when reading numeric data into R (usually, when reading in a file), they come in as factors. If `f` is such a factor object, you can use

```
as.numeric(as.character(f))
```

to get the numbers back. More efficient, but harder to remember, is

```
as.numeric(levels(f))[as.integer(f)]
```

In any case, do not call `as.numeric()` or their likes directly for the task at hand (as `as.numeric()` or `unclass()` give the internal codes).

7.11 Are Trellis displays implemented in R?

The recommended package **lattice** (which is based on base package **grid**) provides graphical functionality that is compatible with most Trellis commands.

You could also look at `coplot()` and `dotchart()` which might do at least some of what you want. Note also that the R version of `pairs()` is fairly general and provides most of the functionality of `splom()`, and that R's default plot method has an argument `asp` allowing to specify (and fix against device resizing) the aspect ratio of the plot.

(Because the word "Trellis" has been claimed as a trademark we do not use it in R. The name "lattice" has been chosen for the R equivalent.)

7.12 What are the enclosing and parent environments?

Inside a function you may want to access variables in two additional environments: the one that the function was defined in ("enclosing"), and the one it was invoked in ("parent").

If you create a function at the command line or load it in a package its enclosing environment is the global workspace. If you define a function `f()` inside another function `g()` its enclosing environment is the environment inside `g()`. The enclosing environment for a function is fixed when the function is created. You can find out the enclosing environment for a function `f()` using `environment(f)`.

The "parent" environment, on the other hand, is defined when you invoke a function. If you invoke `lm()` at the command line its parent environment is the global workspace, if you invoke it inside a function `f()` then its parent environment is the environment inside `f()`. You can find out the parent environment for an invocation of a function by using `parent.frame()` or `sys.frame(sys.parent())`.

So for most user-visible functions the enclosing environment will be the global workspace, since that is where most functions are defined. The parent environment will be wherever the function happens to be called from. If a function `f()` is defined inside another function `g()` it will probably be used inside `g()` as well, so its parent environment and enclosing environment will probably be the same.

Parent environments are important because things like model formulas need to be evaluated in the environment the function was called from, since that's where all the variables will be available. This relies on the parent environment being potentially different with each invocation.

Enclosing environments are important because a function can use variables in the enclosing environment to share information with other functions or with other invocations of itself (see the section on lexical scoping). This relies on the enclosing environment being the same each time the function is invoked. (In C this would be done with static variables.)

Scoping *is* hard. Looking at examples helps. It is particularly instructive to look at examples that work differently in R and S and try to see why they differ. One way to describe the scoping differences between R and S is to say that in S the enclosing environment is *always* the global workspace, but in R the enclosing environment is wherever the function was created.

7.13 How can I substitute into a plot label?

Often, it is desired to use the value of an R object in a plot label, e.g., a title. This is easily accomplished using `paste()` if the label is a simple character string, but not always obvious in case the label is an expression (for refined mathematical annotation). In such a case, either use `parse()` on your pasted character string or use `substitute()` on an expression. For example, if `ahat` is an estimator of your parameter a of interest, use

```
title(substitute(hat(a) == ahat, list(ahat = ahat)))
```

(note that it is '`==`' and not '`=`'). Sometimes `bquote()` gives a more compact form, e.g.,

```
title(bquote(hat(a) = .(ahat)))
```

where subexpressions enclosed in '`.()`' are replaced by their values.

There are more examples in the mailing list archives.

7.14 What are valid names?

When creating data frames using `data.frame()` or `read.table()`, R by default ensures that the variable names are syntactically valid. (The argument `check.names` to these functions controls whether variable names are checked and adjusted by `make.names()` if needed.)

To understand what names are "valid", one needs to take into account that the term "name" is used in several different (but related) ways in the language:

1. A *syntactic name* is a string the parser interprets as this type of expression. It consists of letters, numbers, and the dot and (for versions of R at least 1.9.0) underscore characters, and starts with either a letter or a dot not followed by a number. Reserved words are not syntactic names.
2. An *object name* is a string associated with an object that is assigned in an expression either by having the object name on the left of an assignment operation or as an argument to the `assign()` function. It is usually a syntactic name as well, but can be any non-empty string if it is quoted (and it is always quoted in the call to `assign()`).
3. An *argument name* is what appears to the left of the equals sign when supplying an argument in a function call (for example, `f(trim=.5)`). Argument names are also usually syntactic names, but again can be anything if they are quoted.
4. An *element name* is a string that identifies a piece of an object (a component of a list, for example.) When it is used on the right of the '`$`' operator, it must be a syntactic name, or quoted. Otherwise, element names can be any strings. (When an object is used as a database, as in a call to `eval()` or `attach()`, the element names become object names.)
5. Finally, a *file name* is a string identifying a file in the operating system for reading, writing, etc. It really has nothing much to do with names in the language, but it is traditional to call these strings file "names".

7.15 Are GAMs implemented in R?

Package **gam** from CRAN implements all the Generalized Additive Models (GAM) functionality as described in the GAM chapter of the White Book. In particular, it implements backfitting with both local regression and smoothing splines, and is extendable. There is a `gam()` function for GAMs in package **mgcv**, but it is not an exact clone of what is described in the White Book (no `lo()` for example). Package **gss** can fit spline-based GAMs too. And if you can accept regression splines you can use `glm()`. For Gaussian GAMs you can use `bruto()` from package **mda**.

7.16 Why is the output not printed when I source() a file?

Most R commands do not generate any output. The command

```
1+1
```

computes the value 2 and returns it; the command

```
summary(glm(y~x+z, family=binomial))
```

fits a logistic regression model, computes some summary information and returns an object of class `"summary.glm"` (see Section 8.1 [How should I write summary methods?], page 44).

If you type '`1+1`' or '`summary(glm(y~x+z, family=binomial))`' at the command line the returned value is automatically printed (unless it is `invisible()`), but in other circumstances, such as in a `source()`d file or inside a function it isn't printed unless you specifically print it.

To print the value use

```
print(1+1)
```

or

```
print(summary(glm(y~x+z, family=binomial)))
```

instead, or use `source(`*`file`*`, echo=TRUE)`.

7.17 Why does outer() behave strangely with my function?

As the help for `outer()` indicates, it does not work on arbitrary functions the way the `apply()` family does. It requires functions that are vectorized to work elementwise on arrays. As you can see by looking at the code, `outer(x, y, FUN)` creates two large vectors containing every possible combination of elements of `x` and `y` and then passes this to `FUN` all at once. Your function probably cannot handle two large vectors as parameters.

If you have a function that cannot handle two vectors but can handle two scalars, then you can still use `outer()` but you will need to wrap your function up first, to simulate vectorized behavior. Suppose your function is

```
foo <- function(x, y, happy) {
  stopifnot(length(x) == 1, length(y) == 1) # scalars only!
  (x + y) * happy
}
```

If you define the general function

```
wrapper <- function(x, y, my.fun, ...) {
  sapply(seq_along(x), FUN = function(i) my.fun(x[i], y[i], ...))
}
```

then you can use `outer()` by writing, e.g.,

```
outer(1:4, 1:2, FUN = wrapper, my.fun = foo, happy = 10)
```

Scalar functions can also be vectorized using `Vectorize()`.

7.18 Why does the output from anova() depend on the order of factors in the model?

In a model such as `~A+B+A:B`, R will report the difference in sums of squares between the models `~1`, `~A`, `~A+B` and `~A+B+A:B`. If the model were `~B+A+A:B`, R would report differences between `~1`, `~B`, `~A+B`, and `~A+B+A:B` . In the first case the sum of squares for `A` is comparing `~1` and `~A`, in the second case it is comparing `~B` and `~B+A`. In a non-orthogonal design (i.e., most unbalanced designs) these comparisons are (conceptually and numerically) different.

Some packages report instead the sums of squares based on comparing the full model to the models with each factor removed one at a time (the famous 'Type III sums of squares'

from SAS, for example). These do not depend on the order of factors in the model. The question of which set of sums of squares is the Right Thing provokes low-level holy wars on R-help from time to time.

There is no need to be agitated about the particular sums of squares that R reports. You can compute your favorite sums of squares quite easily. Any two models can be compared with `anova(model1, model2)`, and `drop1(model1)` will show the sums of squares resulting from dropping single terms.

7.19 How do I produce PNG graphics in batch mode?

Under a Unix-like, if your installation supports the `type="cairo"` option to the `png()` device there should be no problems, and the default settings should just work. This option is not available for versions of R prior to 2.7.0, or without support for cairo. From R 2.7.0 `png()` by default uses the Quartz device on OS X, and that too works in batch mode.

Earlier versions of the `png()` device used the X11 driver, which is a problem in batch mode or for remote operation. If you have Ghostscript you can use `bitmap()`, which produces a PostScript or PDF file then converts it to any bitmap format supported by Ghostscript. On some installations this produces ugly output, on others it is perfectly satisfactory. Many systems now come with Xvfb from X.Org (possibly as an optional install), which is an X11 server that does not require a screen; and there is the **GDD** package from CRAN, which produces PNG, JPEG and GIF bitmaps without X11.

7.20 How can I get command line editing to work?

The Unix-like command-line interface to R can only provide the inbuilt command line editor which allows recall, editing and re-submission of prior commands provided that the GNU readline library is available at the time R is configured for compilation. Note that the 'development' version of readline including the appropriate headers is needed: users of Linux binary distributions will need to install packages such as `libreadline-dev` (Debian) or `readline-devel` (Red Hat).

7.21 How can I turn a string into a variable?

If you have

```
varname <- c("a", "b", "d")
```

you can do

```
get(varname[1]) + 2
```

for

```
a + 2
```

or

```
assign(varname[1], 2 + 2)
```

for

```
a <- 2 + 2
```

or

```
eval(substitute(lm(y ~ x + variable),
                list(variable = as.name(varname[1]))))
```

for

```
lm(y ~ x + a)
```

At least in the first two cases it is often easier to just use a list, and then you can easily index it by name

```
vars <- list(a = 1:10, b = rnorm(100), d = LETTERS)
vars[["a"]]
```

without any of this messing about.

7.22 Why do lattice/trellis graphics not work?

The most likely reason is that you forgot to tell R to display the graph. Lattice functions such as `xyplot()` create a graph object, but do not display it (the same is true of **ggplot2** graphics, and Trellis graphics in S-PLUS). The `print()` method for the graph object produces the actual display. When you use these functions interactively at the command line, the result is automatically printed, but in `source()` or inside your own functions you will need an explicit `print()` statement.

7.23 How can I sort the rows of a data frame?

To sort the rows within a data frame, with respect to the values in one or more of the columns, simply use `order()` (e.g., `DF[order(DF$a, DF[["b"]]), ]` to sort the data frame `DF` on columns named `a` and `b`).

7.24 Why does the help.start() search engine not work?

The browser-based search engine in `help.start()` utilizes a Java applet. In order for this to function properly, a compatible version of Java must installed on your system and linked to your browser, and both Java *and* JavaScript need to be enabled in your browser.

There have been a number of compatibility issues with versions of Java and of browsers. See Section "Enabling search in HTML help" in *R Installation and Administration*, for further details.

7.25 Why did my .Rprofile stop working when I updated R?

Did you read the `NEWS` file? For functions that are not in the **base** package you need to specify the correct package namespace, since the code will be run *before* the packages are loaded. E.g.,

```
ps.options(horizontal = FALSE)
help.start()
```

needs to be

```
grDevices::ps.options(horizontal = FALSE)
utils::help.start()
```

(`graphics::ps.options(horizontal = FALSE)` in R 1.9.x).

7.26 Where have all the methods gone?

Many functions, particularly S3 methods, are now hidden in namespaces. This has the advantage that they cannot be called inadvertently with arguments of the wrong class, but it makes them harder to view.

To see the code for an S3 method (e.g., `[.terms`) use

```
getS3method("[", "terms")
```

To see the code for an unexported function `foo()` in the namespace of package `"bar"` use `bar:::foo`. Don't use these constructions to call unexported functions in your own code—they are probably unexported for a reason and may change without warning.

7.27 How can I create rotated axis labels?

To rotate axis labels (using base graphics), you need to use `text()`, rather than `mtext()`, as the latter does not support `par("srt")`.

```
## Increase bottom margin to make room for rotated labels
par(mar = c(7, 4, 4, 2) + 0.1)
## Create plot with no x axis and no x axis label
plot(1 : 8, xaxt = "n",  xlab = "")
## Set up x axis with tick marks alone
axis(1, labels = FALSE)
## Create some text labels
labels <- paste("Label", 1:8, sep = " ")
## Plot x axis labels at default tick marks
text(1:8, par("usr")[3] - 0.25, srt = 45, adj = 1,
     labels = labels, xpd = TRUE)
## Plot x axis label at line 6 (of 7)
mtext(1, text = "X Axis Label", line = 6)
```

When plotting the x axis labels, we use `srt = 45` for text rotation angle, `adj = 1` to place the right end of text at the tick marks, and `xpd = TRUE` to allow for text outside the plot region. You can adjust the value of the `0.25` offset as required to move the axis labels up or down relative to the x axis. See `?par` for more information.

Also see Figure 1 and associated code in Paul Murrell (2003), "Integrating grid Graphics Output with Base Graphics Output", *R News*, **3/2**, 7–12.

7.28 Why is read.table() so inefficient?

By default, `read.table()` needs to read in everything as character data, and then try to figure out which variables to convert to numerics or factors. For a large data set, this takes considerable amounts of time and memory. Performance can substantially be improved by using the `colClasses` argument to specify the classes to be assumed for the columns of the table.

7.29 What is the difference between package and library?

A *package* is a standardized collection of material extending R, e.g. providing code, data, or documentation. A *library* is a place (directory) where R knows to find packages it can use

(i.e., which were *installed*). R is told to use a package (to "load" it and add it to the search path) via calls to the function `library`. I.e., `library()` is employed to load a package from libraries containing packages.

See Chapter 5 [R Add-On Packages], page 22, for more details. See also Uwe Ligges (2003), "R Help Desk: Package Management", *R News*, **3/3**, 37–39.

7.30 I installed a package but the functions are not there

To actually *use* the package, it needs to be *loaded* using `library()`.

See Chapter 5 [R Add-On Packages], page 22 and Section 7.29 [What is the difference between package and library?], page 37 for more information.

7.31 Why doesn't R think these numbers are equal?

The only numbers that can be represented exactly in R's numeric type are integers and fractions whose denominator is a power of 2. Other numbers have to be rounded to (typically) 53 binary digits accuracy. As a result, two floating point numbers will not reliably be equal unless they have been computed by the same algorithm, and not always even then. For example

```
R> a <- sqrt(2)
R> a * a == 2
[1] FALSE
R> a * a - 2
[1] 4.440892e-16
```

The function `all.equal()` compares two objects using a numeric tolerance of `.Machine$double.eps ^ 0.5`. If you want much greater accuracy than this you will need to consider error propagation carefully.

For more information, see e.g. David Goldberg (1991), "What Every Computer Scientist Should Know About Floating-Point Arithmetic", *ACM Computing Surveys*, **23/1**, 5–48, also available via `http://www.validlab.com/goldberg/paper.pdf`.

To quote from "The Elements of Programming Style" by Kernighan and Plauger:

> *10.0 times 0.1 is hardly ever 1.0.*

7.32 How can I capture or ignore errors in a long simulation?

Use `try()`, which returns an object of class `"try-error"` instead of an error, or preferably `tryCatch()`, where the return value can be configured more flexibly. For example

```
beta[i,] <- tryCatch(coef(lm(formula, data)),
                     error = function(e) rep(NaN, 4))
```

would return the coefficients if the `lm()` call succeeded and would return `c(NaN, NaN, NaN, NaN)` if it failed (presumably there are supposed to be 4 coefficients in this example).

7.33 Why are powers of negative numbers wrong?

You are probably seeing something like

```
R> -2^2
[1] -4
```

and misunderstanding the precedence rules for expressions in R. Write

```
R> (-2)^2
[1] 4
```

to get the square of -2.

The precedence rules are documented in `?Syntax`, and to see how R interprets an expression you can look at the parse tree

```
R> as.list(quote(-2^2))
[[1]]
`-`

[[2]]
2^2
```

7.34 How can I save the result of each iteration in a loop into a separate file?

One way is to use `paste()` (or `sprintf()`) to concatenate a stem filename and the iteration number while `file.path()` constructs the path. For example, to save results into files `result1.rda`, ..., `result100.rda` in the subdirectory `Results` of the current working directory, one can use

```
for(i in 1:100) {
  ## Calculations constructing "some_object" ...
  fp <- file.path("Results", paste("result", i, ".rda", sep = ""))
  save(list = "some_object", file = fp)
}
```

7.35 Why are p-values not displayed when using lmer()?

Doug Bates has kindly provided an extensive response in a post to the r-help list, which can be reviewed at https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html.

7.36 Why are there unwanted borders, lines or grid-like artifacts when viewing a plot saved to a PS or PDF file?

This can occur when using functions such as `polygon()`, `filled.contour()`, `image()` or other functions which may call these internally. In the case of `polygon()`, you may observe unwanted borders between the polygons even when setting the `border` argument to `NA` or `"transparent"`.

The source of the problem is the PS/PDF viewer when the plot is anti-aliased. The details for the solution will be different depending upon the viewer used, the operating system and may change over time. For some common viewers, consider the following:

Acrobat Reader (cross platform)
: There are options in Preferences to enable/disable text smoothing, image smoothing and line art smoothing. Disable line art smoothing.

Preview (OS X)
: There is an option in Preferences to enable/disable anti-aliasing of text and line art. Disable this option.

GSview (cross platform)
: There are settings for Text Alpha and Graphics Alpha. Change Graphics Alpha from 4 bits to 1 bit to disable graphic anti-aliasing.

gv (Unix-like X)
: There is an option to enable/disable anti-aliasing. Disable this option.

Evince (Linux/GNOME)
: There is not an option to disable anti-aliasing in this viewer.

Okular (Linux/KDE)
: There is not an option in the GUI to enable/disable anti-aliasing. From a console command line, use:

```
$ kwriteconfig --file okularpartrc --group 'Dlg Performance' \
               --key GraphicsAntialias Disabled
```

Then restart Okular. Change the final word to '`Enabled`' to restore the original setting.

7.37 Why does backslash behave strangely inside strings?

This question most often comes up in relation to file names (see Section 7.8 [How do file names work in Windows?], page 31) but it also happens that people complain that they cannot seem to put a single '\' character into a text string unless it happens to be followed by certain other characters.

To understand this, you have to distinguish between character strings and *representations* of character strings. Mostly, the representation in R is just the string with a single or double quote at either end, but there are strings that cannot be represented that way, e.g., strings that themselves contain the quote character. So

```
> str <- "This \"text\" is quoted"
> str
[1] "This \"text\" is quoted"
> cat(str, "\n")
This "text" is quoted
```

The *escape sequences* '`\"`' and '`\n`' represent a double quote and the newline character respectively. Printing text strings, using `print()` or by typing the name at the prompt will use the escape sequences too, but the `cat()` function will display the string as-is. Notice that '`"\n"`' is a one-character string, not two; the backslash is not actually in the string, it is just generated in the printed representation.

```
> nchar("\n")
[1] 1
> substring("\n", 1, 1)
```

```
[1] "\n"
```

So how do you put a backslash in a string? For this, you have to escape the escape character. I.e., you have to double the backslash. as in

```
> cat("\\n", "\n")
\n
```

Some functions, particularly those involving regular expression matching, themselves use metacharacters, which may need to be escaped by the backslash mechanism. In those cases you may need a *quadruple* backslash to represent a single literal one.

In versions of R up to 2.4.1 an unknown escape sequence like '`\p`' was quietly interpreted as just '`p`'. Current versions of R emit a warning.

7.38 How can I put error bars or confidence bands on my plot?

Some functions will display a particular kind of plot with error bars, such as the `bar.err()` function in the **agricolae** package, the `plotCI()` function in the **gplots** package, the `plotCI()` and `brkdn.plot()` functions in the **plotrix** package and the `error.bars()`, `error.crosses()` and `error.bars.by()` functions in the **psych** package. Within these types of functions, some will accept the measures of dispersion (e.g., `plotCI`), some will calculate the dispersion measures from the raw values (`bar.err`, `brkdn.plot`), and some will do both (`error.bars`). Still other functions will just display error bars, like the dispersion function in the **plotrix** package. Most of the above functions use the `arrows()` function in the base **graphics** package to draw the error bars.

The above functions all use the base graphics system. The grid and lattice graphics systems also have specific functions for displaying error bars, e.g., the `grid.arrow()` function in the **grid** package, and the `geom_errorbar()`, `geom_errorbarh()`, `geom_pointrange()`, `geom_linerange()`, `geom_crossbar()` and `geom_ribbon()` functions in the **ggplot2** package. In the lattice system, error bars can be displayed with `Dotplot()` or `xYplot()` in the **Hmisc** package and `segplot()` in the **latticeExtra** package.

7.39 How do I create a plot with two y-axes?

Creating a graph with two y-axes, i.e., with two sorts of data that are scaled to the same vertical size and showing separate vertical axes on the left and right sides of the plot that reflect the original scales of the data, is possible in R but is not recommended. The basic approach for constructing such graphs is to use `par(new=TRUE)` (see `?par`); functions `twoord.plot()` (in the **plotrix** package) and `doubleYScale()` (in the **latticeExtra** package) automate the process somewhat. See `http://rwiki.sciviews.org/doku.php?id=tips:graphics-base:2yaxes` for more information, including strong arguments against this sort of graph.

7.40 How do I access the source code for a function?

In most cases, typing the name of the function will print its source code. However, code is sometimes hidden in a namespace, or compiled. For a complete overview on how to access source code, see Uwe Ligges (2006), "Help Desk: Accessing the sources", *R News*, **6/4**, 43–45 (`http://CRAN.R-project.org/doc/Rnews/Rnews_2006-4.pdf`).

7.41 Why does summary() report strange results for the R^2 estimate when I fit a linear model with no intercept?

As described in `?summary.lm`, when the intercept is zero (e.g., from `y ~ x - 1` or `y ~ x + 0`), `summary.lm()` uses the formula $R^2 = 1 - \sum_i R_i^2 / \sum_i y_i^2$ which is different from the usual $R^2 = 1 - \sum R_i^2 / \sum_i (y_i - \text{mean}(y))^2$. There are several reasons for this:

- Otherwise the R^2 could be negative (because the model with zero intercept can fit *worse* than the constant-mean model it is implicitly compared to).
- If you set the slope to zero in the model with a line through the origin you get fitted values y*=0
- The model with constant, non-zero mean is not nested in the model with a line through the origin.

All these come down to saying that if you know *a priori* that $E[Y] = 0$ when $x = 0$ then the 'null' model that you should compare to the fitted line, the model where x doesn't explain any of the variance, is the model where $E[Y] = 0$ everywhere. (If you don't know a priori that $E[Y] = 0$ when $x = 0$, then you probably shouldn't be fitting a line through the origin.)

7.42 Why is R apparently not releasing memory?

This question is often asked in different flavors along the lines of "I have removed objects in R and run `gc()` and yet `ps`/`top` still shows the R process using a lot of memory", often on Linux machines.

This is an artifact of the way the operating system (OS) allocates memory. In general it is common that the OS is not capable of releasing all unused memory. In extreme cases it is possible that even if R frees almost all its memory, the OS can not release any of it due to its design and thus tools such as `ps` or `top` will report substantial amount of resident RAM used by the R process even though R has released all that memory. In general such tools do *not* report the actual memory usage of the process but rather what the OS is reserving for that process.

The short answer is that this is a limitation of the memory allocator in the operating system and there is nothing R can do about it. That space is simply kept by the OS in the hope that R will ask for it later. The following paragraph gives more in-depth answer with technical details on how this happens.

Most systems use two separate ways to allocate memory. For allocation of large chunks they will use `mmap` to map memory into the process address space. Such chunks can be released immediately when they are completely free, because they can reside anywhere in the virtual memory. However, this is a relatively expensive operation and many OSes have a limit on the number of such allocated chunks, so this is only used for allocating large memory regions. For smaller allocations the system can expand the data segment of the process (historically using the `brk` system call), but this whole area is always contiguous. The OS can only move the end of this space, it cannot create any "holes". Since this operation is fairly cheap, it is used for allocations of small pieces of memory. However, the side-effect is that even if there is just one byte that is in use at the end of the data segment, the OS cannot release any memory at all, because it cannot change the address of that byte. This is actually more common than it may seem, because allocating a lot of intermediate objects, then allocating a result object and removing all intermediate objects

is a very common practice. Since the result is allocated at the end it will prevent the OS from releasing any memory used by the intermediate objects. In practice, this is not necessarily a problem, because modern operating systems can page out unused portions of the virtual memory so it does not necessarily reduce the amount of real memory available for other applications. Typically, small objects such as strings or pairlists will be affected by this behavior, whereas large objects such as long vectors will be allocated using `mmap` and thus not affected. On Linux (and possibly other Unix-like systems) it is possible to use the `mallinfo` system call (also see the mallinfo package) to query the allocator about the layout of the allocations, including the actually used memory as well as unused memory that cannot be released.

8 R Programming

8.1 How should I write summary methods?

Suppose you want to provide a summary method for class `"foo"`. Then `summary.foo()` should not print anything, but return an object of class `"summary.foo"`, *and* you should write a method `print.summary.foo()` which nicely prints the summary information and invisibly returns its object. This approach is preferred over having `summary.foo()` print summary information and return something useful, as sometimes you need to grab something computed by `summary()` inside a function or similar. In such cases you don't want anything printed.

8.2 How can I debug dynamically loaded code?

Roughly speaking, you need to start R inside the debugger, load the code, send an interrupt, and then set the required breakpoints.

See Section "Finding entry points in dynamically loaded code" in *Writing R Extensions*.

8.3 How can I inspect R objects when debugging?

The most convenient way is to call `R_PV` from the symbolic debugger.

See Section "Inspecting R objects when debugging" in *Writing R Extensions*.

8.4 How can I change compilation flags?

Suppose you have C code file for dynloading into R, but you want to use `R CMD SHLIB` with compilation flags other than the default ones (which were determined when R was built).

Starting with R 2.1.0, users can provide personal Makevars configuration files in `$HOME/.R` to override the default flags. See Section "Add-on packages" in *R Installation and Administration*.

For earlier versions of R, you could change the file *`R_HOME`*`/etc/Makeconf` to reflect your preferences, or (at least for systems using GNU Make) override them by the environment variable `MAKEFLAGS`. See Section "Creating shared objects" in *Writing R Extensions*.

8.5 How can I debug S4 methods?

Use the `trace()` function with argument `signature=` to add calls to the browser or any other code to the method that will be dispatched for the corresponding signature. See `?trace` for details.

9 R Bugs

9.1 What is a bug?

If R executes an illegal instruction, or dies with an operating system error message that indicates a problem in the program (as opposed to something like "disk full"), then it is certainly a bug. If you call `.C()`, `.Fortran()`, `.External()` or `.Call()` (or `.Internal()`) yourself (or in a function you wrote), you can always crash R by using wrong argument types (modes). This is not a bug.

Taking forever to complete a command can be a bug, but you must make certain that it was really R's fault. Some commands simply take a long time. If the input was such that you *know* it should have been processed quickly, report a bug. If you don't know whether the command should take a long time, find out by looking in the manual or by asking for assistance.

If a command you are familiar with causes an R error message in a case where its usual definition ought to be reasonable, it is probably a bug. If a command does the wrong thing, that is a bug. But be sure you know for certain what it ought to have done. If you aren't familiar with the command, or don't know for certain how the command is supposed to work, then it might actually be working right. For example, people sometimes think there is a bug in R's mathematics because they don't understand how finite-precision arithmetic works. Rather than jumping to conclusions, show the problem to someone who knows for certain. Unexpected results of comparison of decimal numbers, for example `0.28 * 100 != 28` or `0.1 + 0.2 != 0.3`, are not a bug. See Section 7.31 [Why doesn't R think these numbers are equal?], page 38, for more details.

Finally, a command's intended definition may not be best for statistical analysis. This is a very important sort of problem, but it is also a matter of judgment. Also, it is easy to come to such a conclusion out of ignorance of some of the existing features. It is probably best not to complain about such a problem until you have checked the documentation in the usual ways, feel confident that you understand it, and know for certain that what you want is not available. If you are not sure what the command is supposed to do after a careful reading of the manual this indicates a bug in the manual. The manual's job is to make everything clear. It is just as important to report documentation bugs as program bugs. However, we know that the introductory documentation is seriously inadequate, so you don't need to report this.

If the online argument list of a function disagrees with the manual, one of them must be wrong, so report the bug.

9.2 How to report a bug

When you decide that there is a bug, it is important to report it and to report it in a way which is useful. What is most useful is an exact description of what commands you type, starting with the shell command to run R, until the problem happens. Always include the version of R, machine, and operating system that you are using; type `version` in R to print this.

The most important principle in reporting a bug is to report *facts*, not hypotheses or categorizations. It is always easier to report the facts, but people seem to prefer to strain

to posit explanations and report them instead. If the explanations are based on guesses about how R is implemented, they will be useless; others will have to try to figure out what the facts must have been to lead to such speculations. Sometimes this is impossible. But in any case, it is unnecessary work for the ones trying to fix the problem.

For example, suppose that on a data set which you know to be quite large the command

```
R> data.frame(x, y, z, monday, tuesday)
```

never returns. Do not report that `data.frame()` fails for large data sets. Perhaps it fails when a variable name is a day of the week. If this is so then when others got your report they would try out the `data.frame()` command on a large data set, probably with no day of the week variable name, and not see any problem. There is no way in the world that others could guess that they should try a day of the week variable name.

Or perhaps the command fails because the last command you used was a method for `"["()` that had a bug causing R's internal data structures to be corrupted and making the `data.frame()` command fail from then on. This is why others need to know what other commands you have typed (or read from your startup file).

It is very useful to try and find simple examples that produce apparently the same bug, and somewhat useful to find simple examples that might be expected to produce the bug but actually do not. If you want to debug the problem and find exactly what caused it, that is wonderful. You should still report the facts as well as any explanations or solutions. Please include an example that reproduces (e.g., http://en.wikipedia.org/wiki/Reproducibility) the problem, preferably the simplest one you have found.

Invoking R with the `--vanilla` option may help in isolating a bug. This ensures that the site profile and saved data files are not read.

Before you actually submit a bug report, you should check whether the bug has already been reported and/or fixed. First, try the "Show open bugs new-to-old" or the search facility on http://bugs.R-project.org/. Second, consult https://svn.R-project.org/R/trunk/doc/NEWS.Rd, which records changes that will appear in the *next* release of R, including bug fixes that do not appear on the Bug Tracker. Third, if possible try the current r-patched or r-devel version of R. If a bug has already been reported or fixed, please do not submit further bug reports on it. Finally, check carefully whether the bug is with R, or a contributed package. Bug reports on contributed packages should be sent first to the package maintainer, and only submitted to the R-bugs repository by package maintainers, mentioning the package in the subject line.

A bug report can be generated using the function `bug.report()`. For reports on R this will open the Web page at http://bugs.R-project.org/: for a contributed package it will open the package's bug tracker Web page or help you compose an email to the maintainer.

There is a section of the bug repository for suggestions for enhancements for R labelled '`wishlist`'. Suggestions can be submitted in the same ways as bugs, but please ensure that the subject line makes clear that this is for the wishlist and not a bug report, for example by starting with '`Wishlist:`'.

Comments on and suggestions for the Windows port of R should be sent to R-windows@R-project.org.

Corrections to and comments on message translations should be sent to the last translator (listed at the top of the appropriate '`.po`' file) or to the translation team as listed at http://developer.R-project.org/TranslationTeams.html.

10 Acknowledgments

Of course, many many thanks to Robert and Ross for the R system, and to the package writers and porters for adding to it.

Special thanks go to Doug Bates, Peter Dalgaard, Paul Gilbert, Stefano Iacus, Fritz Leisch, Jim Lindsey, Thomas Lumley, Martin Maechler, Brian D. Ripley, Anthony Rossini, and Andreas Weingessel for their comments which helped me improve this FAQ.

More to come soon . . .