Introduction to R: Part V

Annex

Alexandre Perera i Lluna[1,2]

[1]Centre de Recerca en Enginyeria Biomèdica (CREB)
Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial (ESAII)
Universitat Politècnica de Catalunya
mailto:Alexandre.Perera@upc.edu

[2]Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN)

Jan 2011 / Introduction to R
Universitat Rovira i Virgili

Contents I

R and Calc

R and Calc is an extension to *Calc* application, part of the OO.o suite (OpenOffice.org):

- Allows to see R objects from the OO.o menu
- R should be started with `Rserve` running:

```
library(Rserve)
Rserve()
```

- in OO.o, Use the menu bar *R Add-on* > *Rdump()*
- It will open a dialog, input

```
rnorm(10)
cor.test(c(1,2,3,4), c(1,2,2,3))
```

- A new sheet will open with the result of the link to R
- `http://wiki.services.openoffice.org/wiki/R_and_Calc`

What is latex?

LaTeX (written as LaTeX in plain text, pronounced as `/l\'atej/`) is:

- Is a document markup language
- Is document preparation system for the TeX typesetting program
- Originally written in 1984 by Leslie Lamport at SRI International
- Only to the language in which documents are written, not to the text editor itself.

Widely used in :

- Academia
- Thesis
- Technical books

What is latex? II

Why we should all use LaTeX?

1. The typographic quality is comparable to scientific publishing editorials
2. Is *content* oriented as opposite to *WYSIWYG*. Splits presentation from content
3. LaTeX is open source distributed under LPPL.

Latex example

```
\documentclass[12pt]{article}
\title{\LaTeX}
\date{}
\begin{document}
  \maketitle
  \LaTeX{} is a document preparation system for the \TeX{}
  typesetting program. It offers programmable desktop publishing
  features and extensive facilities for automating most aspects of
  typesetting and desktop publishing, including numbering and
  cross-referencing, tables and figures, page layout, bibliographies,
  and much more. \LaTeX{} was originally written in 1984 by Leslie
  Lamport and has become the dominant method for using \TeX; few
  people write in plain \TeX{} anymore. The current version is
  \LaTeXe.
  \newline
   % This is a comment, it is not shown in the final output.
   % The following shows a little of the typesetting power of LaTeX
  \begin{eqnarray}
    E &=& mc^2                                  \\
    m &=& \frac{m_0}{\sqrt{1-\frac{v^2}{c^2}}}
  \end{eqnarray}
\end{document}
```

Latex example II

LaTeX

LaTeX is a document preparation system for the TeX typesetting program. It offers programmable desktop publishing features and extensive facilities for automating most aspects of typesetting and desktop publishing, including numbering and cross-referencing, tables and figures, page layout, bibliographies, and much more. LaTeX was originally written in 1984 by Leslie Lamport and has become the dominant method for using TeX; few people write in plain TeX anymore. The current version is LaTeX 2_ε.

$$E = mc^2 \tag{1}$$

$$m = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{2}$$

Sweave

In LaTeX code...

```
<<ejemploSweave>>= x <- 10
 y <- 20
 x * y
@
```

Output...

```
> x <- 10
> y <- 20
> x * y

[1] 200
```

Sweave

In LaTeX code

```
<<ejemploSweave>>=

library(iris)
head(iris[,c(1,2,5)])
summary(iris[,c(1,2,5)])

@
```

In LaTeX code

```
<<ejplot,echo=FALSE,fig=TRUE>>=

plot(iris,col=iris$Species)

@
```

Sweave II

Output...

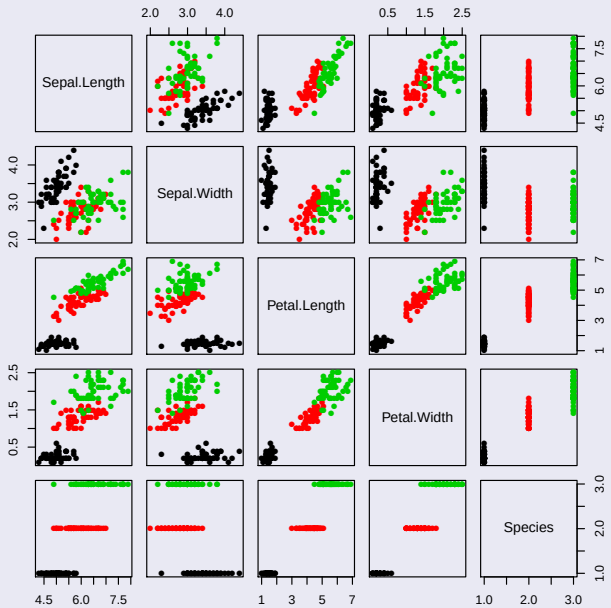
```
> data(iris)
> head(iris[, c(1, 2, 5)])

  Sepal.Length Sepal.Width Species
1          5.1         3.5  setosa
2          4.9         3.0  setosa
3          4.7         3.2  setosa
4          4.6         3.1  setosa
5          5.0         3.6  setosa
6          5.4         3.9  setosa

> summary(iris[, c(1, 2, 5)])

  Sepal.Length    Sepal.Width          Species
 Min.   :4.300   Min.   :2.000   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   versicolor:50
 Median :5.800   Median :3.000   virginica :50
 Mean   :5.843   Mean   :3.057
 3rd Qu.:6.400   3rd Qu.:3.300
 Max.   :7.900   Max.   :4.400
```

Output...



Parallel Computing?

Massive and distributed computing

1. Large number of parameters
2. Large number of variables
3. Large number of cases

MPI

- *MPI is a library specification for message-passing, proposed as a standard by a broadly based committee of vendors, implementors, and users.*
- Standard MPI available from `http://www-unix.mcs.anl.gov/mpi/standard.html`
- *MPI was designed for high performance on both massively parallel machines and on workstation clusters.*
- Extensive use: `http://www-unix.mcs.anl.gov/mpi/libraries.html`

R, Snow, Rmpi

It is possibel to go for "easy" parallel computing with R:

1. SNOW (simple network of worksations) is a easy parallel computing framework, under the frame of *lam-mpi*, via `Rmpi`
2. Snow exists thanks to Sigal Blay (`http://www.sfu.ca/~sblay/`).

Start up an R cluster is easy:

```
# carregar les llibreries de paralelitzacip
library(snow,rmpi)

# Arrancar 8 procesos de R
cl <- makeCluster(8, "MPI")
```

To stop the cluster

```
stopcluster(cl)
```

Rsnow, example

```
> cl <- makeCluster(8, type = "MPI")
> x <- list(alpha = 1:10, beta = exp(-3:3))
> x

$alpha
 [1]  1  2  3  4  5  6  7  8  9 10

$beta
[1]  0.04978707  0.13533528  0.36787944  1.00000000
[5]  2.71828183  7.38905610  20.08553692

> parLapply(cl, x, quantile)

$alpha
   0%   25%   50%   75%  100%
 1.00  3.25  5.50  7.75 10.00

$beta
         0%          25%          50%          75%         100%
 0.04978707  0.25160736  1.00000000  5.05366896 20.08553692

stopcluster(cl)
```

Bibliografy I

Jim Albert.

Bayesian Computation with R.
Springer, New York, 2007.
ISBN 978-0-387-71384-7.

Richard A. Becker, John M. Chambers, and Allan R. Wilks.

The New S Language.
Chapman & Hall, London, 1988.

Andreas Behr.

Einführung in die Statistik mit R.
WiSo Kurzlehrbücher. Vahlen, München, 2005.
ISBN 3-8006-3219-5, in German.

Roger S. Bivand, Edzer J. Pebesma, and Virgilio Gómez-Rubio.

Applied Spatial Data Analysis with R.
Springer, New York, 2008.
ISBN 978-0-387-78170-.

Philip J. Boland.

Statistical and Probabilistic Methods in Actuarial Science.
Chapman & Hall/CRC, Boca Raton, FL, 2007.
ISBN 1-584-88695-1.

W. John Braun and Duncan J. Murdoch.

A First Course in Statistical Programming with R.
Cambridge University Press, Cambridge, 2007.
ISBN 978-0521872652.


Bibliografy II

John M. Chambers.

Programming with Data.
Springer, New York, 1998.
ISBN 0-387-98503-4.

John M. Chambers.

Software for Data Analysis: Programming with R.
Springer, New York, 2007.
ISBN 978-0-387-75935-7.

John M. Chambers and Trevor J. Hastie.

Statistical Models in S.
Chapman & Hall, London, 1992.

Julien Claude.

Morphometrics with R.
Springer, New York, 2008.
ISBN 978-0-387-77789-4.

Dianne Cook and Deborah F. Swayne.

Interactive and Dynamic Graphics for Data Analysis.
Springer, New York, 2007.
ISBN 978-0-387-71761-6.

Michael J. Crawley.

Statistics: An Introduction using R.
Wiley, 2005.
ISBN 0-470-02297-3.

Bibliografy III

Francisco Cribari-Neto and Spyros G. Zarkos.

R: Yet another econometric programming environment.
Journal of Applied Econometrics, 14:319–329, 1999.

Jonathan D. Cryer and Kung-Sik Chan.

Time Series Analysis With Applications in R.
Springer, New York, 2008.
ISBN 978-0-387-75958-6.

Peter Dalgaard.

Introductory Statistics with R.
Springer, 2002.
ISBN 0-387-95475-9.

Richard C. Deonier, Simon Tavaré, and Michael S. Waterman.

Computational Genome Analysis: An Introduction.
Springer, 2005.
ISBN: 0-387-98785-1.

Peter J. Diggle and Paulo Justiniano Ribeiro.

Model-based Geostatistics.
Springer, 2006.
ISBN 0-387-32907-2.

Dubravko Dolic.

Statistik mit R. Einführung für Wirtschafts- und Sozialwissenschaftler.
R. Oldenbourg, München, Wien, 2004.
ISBN 3-486-27537-2, in German.

Bibliografy IV

Sandrine Dudoit and Mark J. van der Laan.

Multiple Testing Procedures and Applications to Genomics.
Springer Series in Statistics. Springer, 2007.
ISBN: 978-0-387-49316-9.

Stephen P. Ellner.

Review of R, version 1.1.1.
Bulletin of the Ecological Society of America, 82(2):127–128, April 2001.

Brian Everitt and Torsten Hothorn.

A Handbook of Statistical Analyses Using R.
Chapman & Hall/CRC, Boca Raton, FL, 2006.
ISBN 1-584-88539-4.

Brian S. Everitt.

An R and S-Plus Companion to Multivariate Analysis.
Springer, 2005.
ISBN 1-85233-882-2.

Julian J. Faraway.

Linear Models with R.
Chapman & Hall/CRC, Boca Raton, FL, 2004.
ISBN 1-584-88425-8.

Julian J. Faraway.

Extending Linear Models with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.
Chapman & Hall/CRC, Boca Raton, FL, 2006.
ISBN 1-584-88424-X.

Bibliografy V

John Fox.

An R and S-Plus Companion to Applied Regression.
Sage Publications, Thousand Oaks, CA, USA, 2002.
ISBN 0-761-92279-2.

Reinhard Furrer and Diego Kuonen.

GRASS GIS et R: main dans la main dans un monde libre.
Flash Informatique Spécial Été, pages 51–56, sep 2001.

Robert Gentleman.

Bioinformatics with R.
Chapman & Hall/CRC, Boca Raton, FL, 2008.
ISBN 1-420-06367-7.

Robert Gentleman, Vince Carey, Wolfgang Huber, Rafael Irizarry, and Sandrine Dudoit, editors.

Bioinformatics and Computational Biology Solutions Using R and Bioconductor.
Statistics for Biology and Health. Springer, 2005.
ISBN: 0-387-25146-4.

Robert Gentleman and Ross Ihaka.

Lexical scope and statistical computing.
Journal of Computational and Graphical Statistics, 9:491–508, 2000.

Michael Greenacre.

Correspondence Analysis in Practice, Second Edition.
Chapman & Hall/CRC, Boca Raton, FL, 2007.
ISBN 1-584-88616-1.

Bibliografy VI

Frank E. Harrell.

Regression Modeling Strategies, with Applications to Linear Models, Survival Analysis and Logistic Regression.
Springer, 2001.
ISBN 0-387-95232-2.

Richard M. Heiberger and Burt Holland.

Statistical Analysis and Data Display: An Intermediate Course with Examples in S-Plus, R, and SAS.
Springer Texts in Statistics. Springer, 2004.
ISBN 0-387-40270-5.

Kurt Hornik and Friedrich Leisch, editors.

Proceedings of the 2nd International Workshop on Distributed Statistical Computing (DSC 2001), Technische Universität Wien, Vienna, Austria, 2001.
ISSN 1609-395X.

Sylvie Huet, Annie Bouvier, Marie-Anne Gruet, and Emmanuel Jolivet.

Statistical Tools for Nonlinear Regression.
Springer, New York, 2003.
ISBN 0-387-40081-8.

Stefano Iacus and Guido Masarotto.

Laboratorio di statistica con R.
McGraw-Hill, Milano, 2003.
ISBN 88-386-6084-0.

Stefano M. Iacus.

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

Bibliografy VII

Ross Ihaka and Robert Gentleman.

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

Jana Jureckova and Jan Picek.

Robust Statistical Methods with R.
Chapman & Hall/CRC, Boca Raton, FL, 2006.
ISBN 1-584-88454-1.

Luke Keele.

Semiparametric Regression for the Social Sciences.
Wiley, Chichester, UK, 2008.
ISBN 978-0470319918.

Christian Kleiber and Achim Zeileis.

Applied Econometrics with R.
Springer, New York, 2008.
ISBN 978-0-387-77316-2.

Diego Kuonen.

Introduction au data mining avec R : vers la reconquête du ‘knowledge discovery in databases’ par les statisticiens.
Bulletin of the Swiss Statistical Society, 40:3–7, 2001.

Diego Kuonen and Valerie Chavez.

R - un exemple du succès des modèles libres.
Flash Informatique, 2:3–7, 2001.

Diego Kuonen and Reinhard Furrer.

Data mining avec R dans un monde libre.
Flash Informatique Spécial Été, pages 45–50, sep 2001.

Bibliografy VIII

Nhu D. Le and James V. Zidek.

Statistical Analysis of Environmental Space-Time Processes.
Springer, 2006.
ISBN 0-387-26209-1.

Uwe Ligges.

Programmieren mit R.
Springer-Verlag, Heidelberg, 2005.
ISBN 3-540-20727-9, in German.

Uwe Ligges.

Programmieren mit R.
Springer-Verlag, Heidelberg, 2nd edition, 2007.
ISBN 3-540-36332-7, in German.

Manuel Castejón Limas, Joaquín Ordieres Meré, Fco. Javier de Cos Juez, and Fco. Javier Martínez de Pisón Ascacibar.

Control de Calidad. Metodologia para el analisis previo a la modelización de datos en procesos industriales. Fundamentos teóricos y aplicaciones con R.
Servicio de Publicaciones de la Universidad de La Rioja, 2001.
ISBN 84-95301-48-2.

Scott M. Lynch.

Introduction to Applied Bayesian Statistics and Estimation for Social Scientists.
Springer, New York, 2007.
ISBN 978-0-387-71264-2.

John Maindonald and John Braun.

Data Analysis and Graphics Using R.
Cambridge University Press, Cambridge, 2003.
ISBN 0-521-81336-0.

Bibliografy IX

Jean-Michel Marin and Christian P. Robert.

Bayesian Core: A Practical Approach to Computational Bayesian Statistics.
Springer, New York, 2007.
ISBN 978-0-387-38979-0.

S. Mase, T. Kamakura, M. Jimbo, and K. Kanefuji.

Introduction to Data Science for engineers— Data analysis using free statistical software R (in Japanese).
Suuri-Kogaku-sha, Tokyo, April 2004.
ISBN 4901683128.

Paul Murrell.

R Graphics.
Chapman & Hall/CRC, Boca Raton, FL, 2005.
ISBN 1-584-88486-X.

Paul Murrell and Ross Ihaka.

An approach to providing mathematical annotation in plots.
Journal of Computational and Graphical Statistics, 9:582–599, 2000.

Fionn Murtagh.

Correspondence Analysis and Data Coding with JAVA and R.
Chapman & Hall/CRC, Boca Raton, FL, 2005.
ISBN 1-584-88528-9.

G. P. Nason.

Wavelet Methods in Statistics with R.
Springer, New York, 2008.
ISBN 978-0-387-75960-9.

Bibliografy X

Deborah Nolan and Terry Speed.

Stat Labs: Mathematical Statistics Through Applications.
Springer Texts in Statistics. Springer, 2000.
ISBN 0-387-98974-9.

Emmanuel Paradis.

Analysis of Phylogenetics and Evolution with R.
Use R. Springer, New York, 2006.
ISBN 0-387-32914-5.

Giovanni Parmigiani, Elizabeth S. Garrett, Rafael A. Irizarry, and Scott L. Zeger.

The Analysis of Gene Expression Data.
Springer, New York, 2003.
ISBN 0-387-95577-1.

Roger D. Peng and Francesca Dominici.

Statistical Methods for Environmental Epidemiology with R: A Case Study in Air Pollution and Health.
Springer, New York, 2008.
ISBN 978-0-387-78166-2.

Bernhard Pfaff.

Analysis of Integrated and Cointegrated Time Series with R.
Use R. Springer, 2006.
ISBN 0-387-98784-3.

Bernhard Pfaff.

Analysis of Integrated and Cointegrated Time Series with R, Second Edition.
Springer, New York, 2008.
ISBN 978-0-387-75966-1.

Bibliografy XI

Jose C. Pinheiro and Douglas M. Bates.

Mixed-Effects Models in S and S-Plus.
Springer, 2000.
ISBN 0-387-98957-0.

Clemens Reimann, Peter Filzmoser, Robert Garrett, and Rudolf Dutter.

Statistical Data Analysis Explained: Applied Environmental Statistics with R.
Wiley, Chichester, UK, 2008.
ISBN: 978-0-470-98581-6.

Paulo J. Ribeiro, Jr. and Patrick E. Brown.

Some words on the R project.
The ISBA Bulletin, 8(1):12–16, March 2001.

Vito Ricci.

R : un ambiente opensource per l'analisi statistica dei dati.
Economia e Commercio, 1:69–82, 2004.

Vito Ricci.

Rappresentazione analitica delle distribuzioni statistiche con R (prima parte).
Economia e Commercio, 1/2:47–60, 2005.

Brian D. Ripley.

The R project in statistical computing.
MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network., 1(1):23–25, February 2001.

Maria L. Rizzo.

Statistical Computing with R.
Chapman & Hall/CRC, Boca Raton, FL, 2008.
ISBN 1-584-88545-9.

Bibliografy XII

Lothar Sachs and Jürgen Hedderich.

Angewandte Statistik. Methodensammlung mit R.
Springer, Berlin, Heidelberg, 12th (completely revised) edition, 2006.
ISBN 978-3-540-32160-6.

Deepayan Sarkar.

Lattice Multivariate Data Visualization with R.
Springer, New York, 2007.
ISBN 978-0-387-75968-5.

Robert H. Shumway and David S. Stoffer.

Time Series Analysis and Its Applications With R Examples.
Springer, New York, 2006.
ISBN 978-0-387-29317-2.

David Siegmund and Benjamin Yakir.

The Statistics of Gene Mapping.
Springer, New York, 2007.
ISBN 978-0-387-49684-9.

Phil Spector.

Data Manipulation with R.
Springer, New York, 2008.
ISBN 978-0-387-74730-9.

Terry M. Therneau and Patricia M. Grambsch.

Modeling Survival Data: Extending the Cox Model.
Statistics for Biology and Health. Springer, 2000.
ISBN: 0-387-98784-3.

Bibliografy XIII

William N. Venables and Brian D. Ripley.

S Programming.
Springer, New York, 2000.
ISBN 0-387-98966-8.

William N. Venables and Brian D. Ripley.

Modern Applied Statistics with S. Fourth Edition.
Springer, New York, 2002.
ISBN 0-387-95457-0.

John Verzani.

Using R for Introductory Statistics.
Chapman & Hall/CRC, Boca Raton, FL, 2005.
ISBN 1-584-88450-9.

Simon N. Wood.

Generalized Additive Models: An Introduction with R.
Chapman & Hall/CRC, Boca Raton, FL, 2006.
ISBN 1-584-88474-6.

End Part V