A Practical Introduction to the R Programming Language

Irucka Embry, EIT
&
Jennifer Murphy

Wednesday, 21 August 2013

Last Revision 30 September 2015

Outline

- R Resources page
- GNU and the FLOSS community
- Why use R
- Reasons to NOT use Microsoft Excel
- Introduction to R and computer programming
- Applications of R
- Useful R Resources

R Resources pages

- http://www.ecoccs.com/RandUSGS.html Created by Irucka Embry for the USGS (wealth of useful links)
- http://www.ecoccs.com/tsuresearch.html Created by Irucka Embry for TSU (wealth of useful links)

GNU and the R community

- GNU [“GNU's Not Unix!”] http://www.gnu.org/
- R community http://www.r-project.org/
- “R is a language and environment for statistical computing and graphics. It is a GNU project (http://www.gnu.org/) which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.
- R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

GNU and the R community 2

- One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.
- R is available as Free Software under the terms of the Free Software Foundation's (http://www.fsf.org/) GNU General Public License (http://www.r-project.org/COPYING) in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and GNU/Linux), Windows and MacOS." (Source: http://www.r-project.org/about.html)

Why use R

- Free (freedom and price)
- Derived from S/S-PLUS and thus built on a powerful programming language
- Publication quality graphics
- Large global community of helpful users
- There are over 7,000 ways to extend base R through packages (30 September 2015)
- Easy to create your own functions and/or packages [collection of function(s)]

Reasons to NOT use Microsoft Excel

- Easy to mess up your data and not know how you did it (incorrect keyboard entry, for example)
- No log or record of the work/tasks you've preformed on your data (unlike writing a script in R)
- Can be cumbersome to run analysis on subsets of data, especially if the subsets are not uniform
- Output of analysis from Excel can also be cumbersome (the results from the histogram analysis, for example)
- Some of the statistical procedures are not well documented (difficult to determine how something is actually being calculated)

Reasons to NOT use Microsoft Excel 2

- Lacks publication quality graphics
- Can not non-destructively alter your data
- http://www.ecoccs.com/RandUSGS.html#excel_nomore
  Why you should ditch Microsoft Excel
- http://blog.revolutionanalytics.com/2013/04/more-reasons-not-to-use-excel-for-modeling.html
  More reasons not to use Excel for modeling
- http://en.wikibooks.org/wiki/Statistics:Numerical_Methods/Numerics_in_Excel
  Statistics/Numerical Methods/Numerics in Excel
- https://answers.microsoft.com/en-us/office/forum/officeversion_other-excel/excel-r-squared-is-incorrect/1dd48555-5f9e-41db-a0d9-b42d95f85499
  Excel R squared is Incorrect - Microsoft Community (2011)

Introduction to R

The R environment

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either on-screen or on hardcopy, and
- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

- The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

Introduction to R 2

- R, like S, is designed around a true computer language, and it allows users to add additional functionality by defining new functions. Much of the system is itself written in the R dialect of S, which makes it easy for users to follow the algorithmic choices made. For computationally-intensive tasks, C, C++ and Fortran code can be linked and called at run time. Advanced users can write C code to manipulate R objects directly.
- Many users think of R as a statistics system. We prefer to think of it of an environment within which statistical techniques are implemented. R can be extended (easily) via packages. There are about eight packages supplied with the R distribution and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics.
- R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hardcopy. (source: http://www.r-project.org/about.html)

Introduction to R 3

- To start R in a *nix based environment
  - Type R from the command line interface (CLI)
  - Click on the icon for a R GUI [ex. RStudio, etc.]
- To start R in a Microsoft Windows
  - Type R from the cmd.exe prompt
  - Click on the icon for a R GUI [ex. R 32-bit or R 64-bit, RStudio, etc.]

Introduction to R 4

- R is ***case sensitive*** (this is not a native Windows application)
- getwd()
  - This command provides the current working directory (where R will save all data)
- setwd("filepath_of_working_directory")
  - This command allows the user to set the current working directory (where R will save all data)
- library()
  - This command allows the user to see a list of the installed R packages

Introduction to R 5

- help(topic)
  - Help for the “topic”
  - help(plot) provides help on the plot function within the R base package graphics
- ?topic
  - Help for the “topic”
  - ?plot provides help on the plot function within the R base package graphics
- help.start()
  - Starts the HTML version of help

Introduction to R 6

- apropos("topic")
  - The names of all objects in the search list that match the regular expression “topic”
  - apropos("plot")

Introduction to R 7

- x <- c(2:4) # this is the variable named x
  - "#" for comments in R
  - "<-" is similar to = in other languages
    - "<-" is left assignment
    - "=" is left assignment (**NOT** recommended)
  - ":" for sequence with a step size of 1
- x # in R
- [1] 2 3 4 # in R
  - [1] is row number 1

Introduction to R 8: Some Common Mathematical Operations (Morrell 9-10)

Mathematical Operation	R expression
x - y	x - y
x + y	x + y
xy	x * y
x/y (fraction)	x/y
x^y	x^y
e^x	exp(x)
$\log_{10}(x)$	log10(x)
ln(x)	log(x)
$\log_2(x)$	log2(x)
cos(x)	cos(x) [radians]
sin(x)	sin(x) [radians]
sqrt(x)	sqrt(x)

Introduction to R 9: Example expressions (Morrell 10)

formula	R expression	Computed Value
$5^2 + 4^2$	5^2 + 4^2	41
$(5 + 4)^2$	(5 + 4)^2	81
2+3/4-5 (fraction)	(2+3)/(4-5)	-5
$\log_{10}(100)$	log10(100)	2
ln(4 * (2 + 3))	log(4*(2+3))	2.995732
cos(30°)	cos(30*pi/180)	0.8660254
sin(30°)	sin(30*pi/180)	0.5

Introduction to R 10: Other useful expressions cont

examplex <- c(-1.6, -1.5, -1.4, 1.4, 1.5, 1.6)

- ceiling(examplex)
  - [1] -1 -1 -1 2 2 2
    - ceiling takes a single numeric argument x and returns a numeric vector containing the smallest integers not less than the corresponding elements of x.
- floor(examplex)
  - [1] -2 -2 -2 1 1 1
    - floor takes a single numeric argument x and returns a numeric vector containing the largest integers not greater than the corresponding elements of x.

Introduction to R 11: Other useful expressions cont

examplex <- c(-1.6, -1.5, -1.4, 1.4, 1.5, 1.6)

- trunc(examplex)
  - [1] -1 -1 -1 1 1 1
    - trunc takes a single numeric argument x and returns a numeric vector containing the integers formed by truncating the values in x toward 0.
- round(examplex, digits = 0)
  - [1] -2 -2 -1 1 2 2
    - round rounds the values in its first argument to the specified number of decimal places (default 0).
- signif(examplex, digits = 1)
  - [1] -2 -2 -1 1 2 2
    - signif rounds the values in its first argument to the specified number of significant digits.

Introduction to R 12

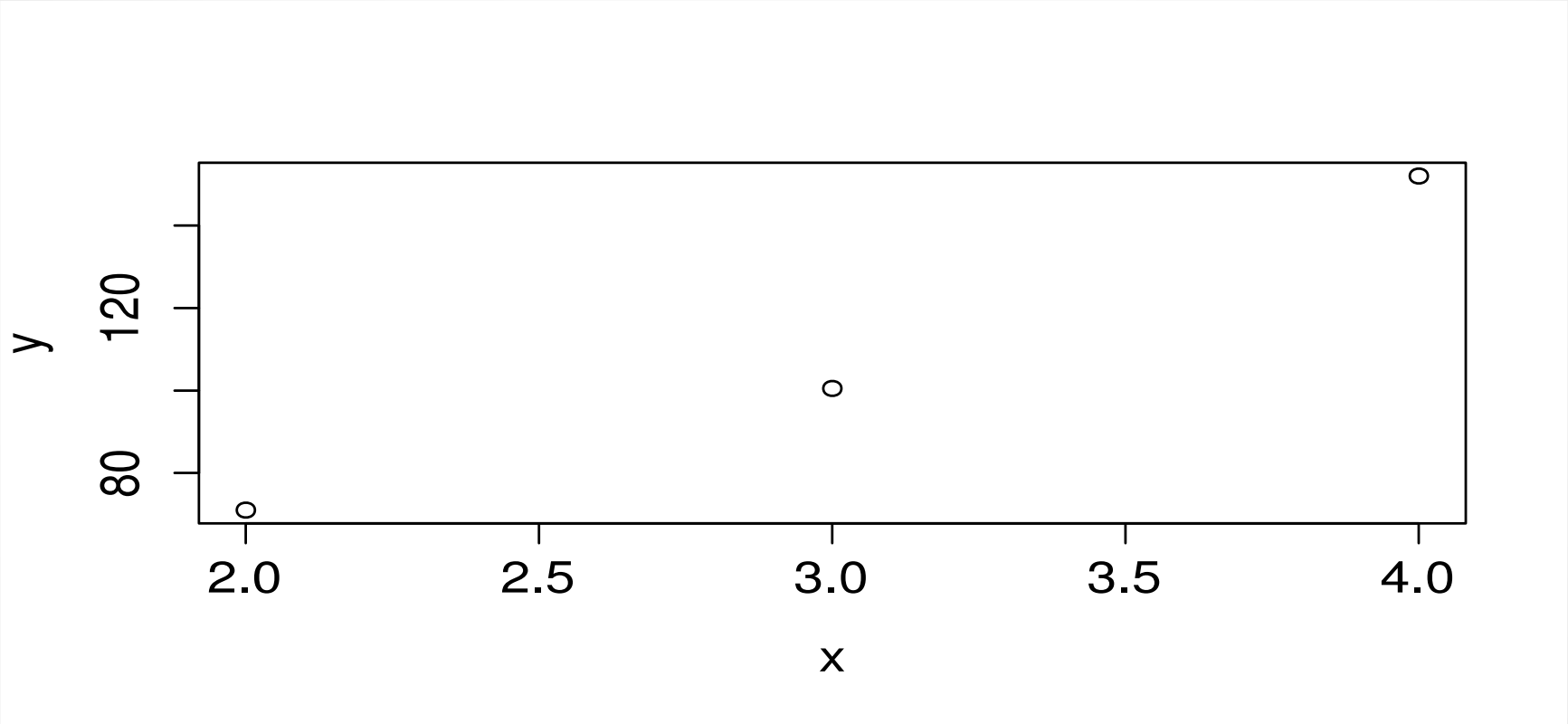
- summary(x)
  - Generic function to give a “summary” of x, often a statistical one
    - x <- c(2:4) # in R
    - summary(x) # in R
      - Min. 1st Qu. Median Mean 3rd Qu. Max.
      - 2.0 2.5 3.0 3.0 3.5 4.0
    - Type library(help = "stats")
      - Shows the complete list of statistical functions

Introduction to R 13

- str(x)
  - Displays the internal structure of an R object
  - str(x) # in R
    - int [1:3] 2 3 4
      - int is for integer
- y <- x^3 + 2*x^2 + 0.5*x + 54 # still using the x from before stored in R
  - str(y) # in R
    - num [1:3] 71 100 152
      - num is for numeric (includes decimals)

Introduction to R 14

- plot(x, y, xlab = "x", ylab = "y")
  - generates a plot of x versus y



Applications of R

- This will be performed within the R environment in RStudio
- The data used in this tutorial will be posted online at http://www.ecoccs.com/RandUSGS.html#tut & http://www.ecoccs.com/tsuresearch.html#tut along with this tutorial

Resources used

- http://www.r-project.org/ R Project for Statistical Computing
- http://cran.r-project.org/doc/contrib/Baggott-refcard-v2.pdf [R Reference Card 2.0 by Matt Baggott]
- Chapra, Steven C., *Applied Numerical Methods with MATLAB for Engineers and Scientists*, 2nd Edition, Boston, Massachusetts: McGraw-Hill, 2008, p. 22, 24-26, 29-34, 36.
- The MathWorks, Inc, MATLAB R2013b Documentation, “2-D line plot - MATLAB plot”, <http://www.mathworks.com/help/matlab/ref/plot.html >, Accessed: 13 October 2013.
- Morrell, D. Freshman Engineering Problem Solving with MATLAB, Connexions Web site. http://cnx.org/content/col10325/1.18/ , Apr 23, 2007, p. 9-11.

Useful R Resources

- http://www.ecoccs.com/RandUSGS.html [R Resources provided by Irucka Embry]
- http://www.ecoccs.com/tsuresearch.html [Research Resources for Tennessee State University (TSU) students and faculty provided by Irucka Embry]
- http://users.monash.edu.au/~murray/stats/Rmanual.pdf [R and S-Plus: Basic Instructions by Murray Logan, July 25, 2005]
- http://www.stat.berkeley.edu/~spector/Rcourse.pdf [Introduction to R (presentation) by Phil Spector, Department of Statistics, University of California]
- http://www.stat.berkeley.edu/users/spector/R.pdf [An Introduction to R by Phil Spector, Department of Statistics, University of California, September 24, 2004]
- http://pairach.com/2012/02/26/r-tutorials-from-universities-around-the-world/ [R-Uni: (A List of Free R Tutorials and Resources in University webpages) by Pairach on February 26, 2012]

Useful R Resources 2

- http://cran.r-project.org/other-docs.html [R Contributed Documentation]
- http://www.r-bloggers.com/free-r-book-collection/ [Free R Book Collection By Richard O. Legendi, September 6, 2011]
- http://stackoverflow.com/questions/tagged/r [R Tagged Questions: Stack Overflow]
- http://www.personality-project.org/r/r.commands.html [A short list of the most useful R commands by William Revelle, Department of Psychology, Northwestern University. A summary of the most important commands with minimal examples. See the relevant part of the guide for better examples. For all of these commands, using the help(function) or ? function is the most useful source of information. Unfortunately, knowing what to ask for help about is the hardest problem.]
- http://www.gardenersown.co.uk/Education/Lectures/R/ [Using R for statistical analyses by Dr. Mark Gardener. This page is intended to be a help in getting to grips with the powerful statistical program called R. It is not intended as a course in statistics. If you have an analysis to perform I hope that you will be able to find the commands you need here and copy/paste them into R to get going. On this page learn how to create data files, read them into R and generally get ready to perform analyses. Also find out about getting further help and documentation.]

Useful R Resources 3

- http://ww2.coastal.edu/kingw/statistics/R-tutorials/index.html [R Tutorials by William B. King, Ph.D., Coastal Carolina University]
- http://rtutorialseries.blogspot.com/ [R Tutorial Series By John M Quick: The R Tutorial Series provides a collection of user-friendly guides to researchers, students, and others who want to learn how to use R for their statistical analyses.]
- http://www.r-bloggers.com/computerworlds-beginners-guide-to-r/ [Computerworld's Beginners Guide to R By David Smith, June 17, 2013]
- http://pj.freefaculty.org/R/Rtips.html [Rtips. Revival 2012! by Paul E. Johnson]
- http://science.nature.nps.gov/im/datamgmt/statistics/r/index.cfm [Using R Statistical and Graphics Tools for Natural Resource Stewardship Science]
- http://math.illinoisstate.edu/dhkim/Rstuff/Rtutor.html [Statistical Computing with R: A tutorial by Dong-Yun Kim]

Useful R Resources 4

- http://cran.r-project.org/doc/contrib/Burns-unwilling_S.pdf [The R language - a short companion: This companion is essentially based on the documents “An Introduction to R” and “R language definition” both version 1.7.1, available on the R website http://www.r-project.org/ . Graphical and statistical functionalities are not considered. Version 1.2. Marc Vandemeulebroecke, July 14th, 2003]
- http://www.itc.nl/~rossiter/teach/R/RIntro_ov.pdf [Introduction to the R Project for Statistical Computing by D G Rossiter, University of Twente, August 10, 2010]
- http://cran.r-project.org/doc/contrib/Baggott-refcard-v2.pdf [R Reference Card 2.0 by Matt Baggott]
- http://www.ats.ucla.edu/stat/r/ [Resources to help you learn and use R: UCLA: Statistical Consulting Group]
- http://www.ats.ucla.edu/stat/r/library/ [R Library by Matt Baggott]

Useful R Resources 5

- http://www.ats.ucla.edu/stat/r/library/advanced_function_r.htm [R Library: Advanced function by Matt Baggott]
- http://phoxis.org/2013/05/04/get-list-of-installed-packages-and-their-details-in-r/ [Get list of installed packages and their details in R: Phoxis]
- http://www.statmethods.net/interface/packages.html"> [Quick-R: R Packages]
- http://data.princeton.edu/R/ [Introducing R by German Rodriguez, Office of Population Research, Princeton University]
- http://data.princeton.edu/R/readingData.html [Reading and Examining Data by German Rodriguez, Office of Population Research, Princeton University]
- http://msenux.redwoods.edu/math/R/dataframe.php [Data Frames in R: Department of Mathematics, College of the Redwoods]
- http://www.r-statistics.com/tag/data-frame/ [Data Frame: R-statistics blog]
- http://nsaunders.wordpress.com/2010/08/20/a-brief-introduction-to-apply-in-r/ [A brief introduction to “apply” in R: What You’re Doing Is Rather Desperate: Notes from the life of a bioinformatician]

Closing

- Thank you for your time