An Introduction to R

March 31, 2009

Joseph G. Voelkel
Center for Quality and Applied Statistics
Kate Gleason College of Engineering
What is R? (mostly from www.r-project.org)

- Integrated suite of software facilities for data manipulation, calculation and graphical display. It includes
  - Effective **data handling** and storage facility,
  - **Suite of operators** for arrays, lists, and other objects
  - Large, integrated set of **intermediate tools** for data analysis,
  - **Graphical facilities** for analysis & display (computer/hardcopy)
  - Well developed, effective **programming language** (‘S’) which includes conditionals, loops, recursive functions, I/O facilities. (Most of system-supplied functions are written in S.)

- Some Features
  - **Object-oriented**
  - Designed to be run **interactively**
  - **Free**
R is an environment

• “environment” is intended to characterize R as a fully planned and coherent system
• Not an incremental accretion of very specific and inflexible tools, frequently the case with other data analysis software.
• A vehicle for newly developing methods of interactive data analysis.
  - It has developed rapidly, and has been extended by a large collection of packages.
  - However, most programs written in R are essentially ephemeral, written for a single piece of data analysis.
Origins of R

• The design of R has been heavily influenced by two existing languages:
  o S (Becker, Chambers & Wilks)
    ▪ S is a very high level language and an environment for data analysis and graphics.
    ▪ In 1998, the ACM presented its Software System Award to John M. Chambers, the principal designer of S
  o Scheme (Sussman)
    ▪ Dialect of Lisp stressing conceptual elegance and simplicity
    ▪ Much smaller than Common Lisp
• Resulting language is very similar in appearance to S or S-Plus
• Underlying implementation and semantics derived from Scheme
• R (“GNU S”)
• “R”: Robert Gentleman and Ross Ihaka—University of Auckland
R is well-known

- Google
  - Minitab software: 149,000
  - JMP software: 173,000
  - SAS software: 7,220,000
  - Java software: 31,200,000
- R software: 69,700,000
R is well-known

- Google
  - Minitab software: 149,000
  - JMP software: 173,000
  - SAS software: 7,220,000
  - Java software: 31,200,000
  - R software: 69,700,000
  - C software: 285,000,000
R is well-known

- Google
  - Minitab software: 149,000
  - JMP software: 173,000
  - SAS software: 7,220,000
  - Java software: 31,200,000
  - **R software**: 69,700,000
  - C software: 285,000,000

- Linux, Mac OS X, Windows

- De facto standard language for many grad statistics programs
- Many corporations (some paying for “R+”)
- You never know where it might show up ...
Yeah, but **What is R??**

- Some Examples
Example 1. Some Basic Ideas

CPU dataset

- Objects
- Data Frames
- Classes
- Search Path
- Graphs
- Linear Regression
- Matrices
Example 2. Some Data Structures

- Vectors
- Matrices
- Arrays
- Lists
- Data Frames
- Combinations of structures
- Your own structures
Example 3. Vectorized Arithmetic

- Vectorized arithmetic
- Some (naïve) alternatives
Simulate 100,000 uniform numbers in [0,1]

nsim<-100000

1. Working on the entire object—good!
   
   ```
   system.time(x<-runif(nsim))
   ```

2. Using a `for` loop—bad!
   
   ```
   x<-rep(NA,nsim)
   system.time(
       for(i in 1:nsim) x[i]<-runif(1) )
   ```

3. Using a `for` loop and building up an object—very bad!
   
   ```
   x<-c()
   system.time(
       for(i in 1:nsim) x<-c(x,runif(1)) )
   ```
Example 4. A Many-Files Problem

- Reading in a more complex file
- Cleaning up the file
- Rearranging data
- Reading in many files

See next page, TestMe.txt, and .R file
1. Scientist wants to work with data: o/p from profilometer.
2. Output: text file with header; x, then y values; trailer
3. What needs to be done
   a. Delete all records up to, including, 2\textsuperscript{nd} row of “EOR”
   b. Delete last two rows: “EOR” and “EOF”
   c. The remaining data should all be numeric, with one number per record. (Say numR records.)
   d. Split single column into two columns of length numR/2 (x=1st numR/2 numbers and y=2nd numR/2 numbers).
   e. Create third column, \( g(x, y) = x + y \).
   f. Write result to file, same as i/p but with “_op” on end.
4. An example file, TestMe.txt, can be used to test the code.
5. Also, investigate relationship of x and y, and look for any unusual values.
6. Then run the i/p\( \rightarrow \)o/p routine on all .txt files in a directory.
Example 5. Windows Files, Regular Expressions

- Accessing Windows file names
- Creating new file names
- Creating a new directory
- Copying files

See pings directory and PingFiles_Example.R
Example 6.  
Function Writing—Sieve of Erasthones’

R naturally lends itself to writing functions

- The ‘sieve of Erasthones’ determines whether a positive integer \(x\) is prime.
- Method: Check each integer \(y\) between 2 and \(\sqrt{x}\) to determine whether \(y\) evenly divides \(x\).
- Requirements
  1. Return TRUE if \(x\) is prime, FALSE otherwise
  2. Return the divisors of \(x\).

- Function writing
- sapply function (one of several *apply functions)
Example 7. More graphs

R has a wide variety of powerful graphic functions. You may also build a graph from more basic graphic calls.
Example 8. Packages

• 1752 at last count
• A wide variety of uses
  o Newest statistical techniques
  o Additions to base R
  o I/O, e.g. html, LaTex, Excel
  o Data sets from books
  o Interfaces to other libraries
  o Graphics
  o Utilities
  o Connections to editors
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADaCGH</td>
<td>Analysis of data from aCGH experiments</td>
</tr>
<tr>
<td>AER</td>
<td>Applied Econometrics with R</td>
</tr>
<tr>
<td>AIGIS</td>
<td>Areal Interpolation for GIS data</td>
</tr>
<tr>
<td>AIS</td>
<td>Tools to look at the data (&quot;Ad Inidicia Spectata&quot;)</td>
</tr>
<tr>
<td>ALS</td>
<td>Multivariate curve resolution alternating least squares (MCR-ALS)</td>
</tr>
<tr>
<td>AMORE</td>
<td>A MORE flexible neural network package</td>
</tr>
<tr>
<td>ARES</td>
<td>Allelic richness estimation, with extrapolation beyond the sample size</td>
</tr>
<tr>
<td>AcceptanceSampling</td>
<td>Creation and evaluation of Acceptance Sampling Plans</td>
</tr>
<tr>
<td>AdMit</td>
<td>Adaptive Mixture of Student-t distributions</td>
</tr>
<tr>
<td>AdaptFit</td>
<td>Adaptive Semiparametric Regression</td>
</tr>
<tr>
<td>Package</td>
<td>Description</td>
</tr>
<tr>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td>yest</td>
<td>Gaussian Independence Models</td>
</tr>
<tr>
<td>ZIGP</td>
<td>Zero Inflated Generalized Poisson (ZIGP) regression models</td>
</tr>
<tr>
<td>Zelig</td>
<td>Everyone's Statistical Software</td>
</tr>
<tr>
<td>zipfR</td>
<td>Statistical models for word frequency distributions</td>
</tr>
<tr>
<td>zoeppritz</td>
<td>Zoeppritz Equations</td>
</tr>
<tr>
<td>zoo</td>
<td>Z's ordered observations</td>
</tr>
<tr>
<td>zyp</td>
<td>Zhang + Yue-Pilon trends package</td>
</tr>
</tbody>
</table>
More Information on R?

www.r-project.org/

www.rit.edu/kgcoe/cqas/about/technicalreports.htm

(My Intro to R for Windows)

Thank you

Questions?