EDA R workshop. Day 1

Goals

* Be familiar with R just enough to perform basic importing and graphing. You won’t be able to do nontrivial analysis or data maniupation.
* Learn how to get help and context of functions.
* Understand how to start a simple graph with minimal code. And then if desired, slowly tweak it to expose unseen patterns or communicate a visible pattern more clearly in a publication.

1. Open R. Start a new script through the File menu. Type “18+20”. “Ctrl+R” executes the highlighted text. Use “Ctrl+A” to selected everything first. If nothing is highlighted, the current line is executed. Output is sent to the Console or the Graphics window.
2. Add two scalars. Create two vectors and add them.

rm(list=ls(all=TRUE)) *#Clears variables*

x <- c(0, 3, 6, 10, 18, 20) *#Create array of IV*

y <- c(6, 2, 4, 10, 8, 6.5) *#Create array of DV*

jerks <- c(1, 0, 0, 1, 0, 1) *#Create a binary variable*

*#Or: <- c(T, F, F, T, F, T) #Create a binary variable*

1. Histogram of one of the vectors.

hist(y)

1. Discuss help files and change the cosmetics of the histogram.

“?hist” and later “?par”

1. Simple scatterplot with existing data

plot(x, y) *#Simple Scatterplot when both variables are continuous*

abline(lm(y ~ x)) *#Draw the "L"inear "M"odel line.*

cor(x, y) *#Calculate correlation.*

1. Examine the linear model itself. Notice how the ‘plot’ function behaves when passed a linear model instead of being passed two vectors.

summary(lm(y ~ x)) *#The model being plotted*

plot(lm(y ~ x)) *#Visualize deviations from linear model*

1. Create graph from empty palette

plot(NA, xlim=c(0, 20), ylim=c(0, 10), xlab="IV", ylab="DV")

xJerks <- x[jerks==1]

yJerks <- y[jerks==1]

points(xJerks, yJerks, col="red", pch=16)

abline(lm(yJerks ~ xJerks), col="red")

xDustins <- x[jerks!=1]

yDustins <- y[jerks!=1]

points(xDustins, yDustins, col="blue", pch=17)

abline(lm(yDustins ~ xDustins), col="blue")

cor(cbind(xJerks, yJerks, xDustins, yDustins))*#Some cells are trash.*

1. Import csv files into a dataframe. Matrix of scatter plots.

*#Use forward slashes (can’t directly paste from Windows Explorer)*

ds <- read.csv(file="D:/Users/wibeasley/Documents/Day1DataFrame.csv")

plot(ds)

summary(ds)

1. Smoothing

plot(ds$x1, ds$x3)

lines(lowess(ds$x3 ~ ds$x1), col="darkorchid1")*#Also see the new ‘loess’ function.*

lines(lowess(ds$x3 ~ ds$x1, f=.09), col="salmon2")

1. Packages. Set of software released by other researchers. They provide specialized or cutting-edge technqiues that aren’t in the base installation of R. It’s most straight-forward to exit out of R. The open it with administrative privileges (right click in windows). Then “install” it once for the lifetime (of that version) of R. Each time R is opened, the package will need to be loaded with the “library” statement (eg., “library(foreign)”).  
     
   Or programmatically:  
   install.packages(c("lme4", "colorspace", "NlsyLinks"))
2. Use the “RODBC “ when importing database tables, and “foreign” for SPSS or SAS. If it’s in Excel, save it as a csv (comma separated file) first.
3. Lattice (includes coplots)

<http://www.stat.auckland.ac.nz/~paul/RGraphics/simplegrid-latticecomplex.png>

1. Colors
2. Any technqiues you want to see Thursday?
3. Dates are a big pain. Read the chapter on dates in Phil Spector’s book.

### Resources

**Websites**

[www.stat.auckland.ac.nz/~paul/RG2e/](http://www.stat.auckland.ac.nz/~paul/RG2e/) Website for Pat Murrell’s book. Good gallery & code.

[addictedtor.free.fr/graphiques/](http://addictedtor.free.fr/graphiques/) Big gallery of graphs & their code; slow loading

**Books dedicated to graphics in R**

*R Graphics*, Pat Murrell, 2011  
My favorite R graphing book

*ggplot2: Elegant Graphics for Data Analysis*, $ Hadley Wickham  
Implements Wilkison's Grammar of Graphics

*Lattice multivariate data visualization with R*, $ Deepayan Sarkar, 2008  
Sarkar wrote the ‘Lattice’ package, which replicates Cleveland’s old ‘trellis’ software.

**Good general R books**

*Introductory Statistics with R*, $ Peter Dalgaard, 2008   
A good basic R book. It provides a good context/motivation for the procedures and graphs.

*Data Manipulation with R*, $ Phil Spector   
No graphs, but a great concise book for the plumbing you’ll use (that we don’t cover at all today).

*A Beginner's Guide to R*, $ Alain Zuur, 2009  
I haven’t read this one. It seems really dumbed down if that what you want to start with.

*An Introduction to R*, Venables & Smith, 2010; it’s included in R for free. Go to Help -> Manuals  
It has a lot of basic content. The common complaint is that it reads like a compilation of help pages. In contrast, Dalgaard provides the context of when a function might be used.

**Help Pages & Vignettes**

Preface a function with a question mark to see its help page. For example, “?lm” or “?lowess”  
It’s great if you already know the function name, and want to know the arguments or details.

Otherwise, you’ll need a real search engine. For Google, include the term “r-project” or “CRAN”. For example, “CRAN nonparametric smoothing” or “CRAN CART”. I’ve had some success with [www.rseek.org/](http://www.rseek.org/)

If you search with the name of a package (e.g., “CRAN foreign” or “CRAN lme4”), you’ll get a lot of stuff, possibly a vignette. If there’s a vignette, you won’t be wasting time skimming it.

A $ means you can download it as a pdf from our library. In the OU Advanced book Search, type “Use R” in the series field. We have 30+ of these for free.

**Color selection:**[statmath.wu.ac.at/~zeileis/papers/Zeileis+Hornik+Murrell-2009.pdf](http://statmath.wu.ac.at/~zeileis/papers/Zeileis+Hornik+Murrell-2009.pdf)  
[cran.r-project.org/web/packages/colorspace/vignettes/hcl-colors.pdf](http://cran.r-project.org/web/packages/colorspace/vignettes/hcl-colors.pdf)