Introduction to Data Analysis and Visualization with R

David Buch

Department of Mathematics
West Virginia University

February 27, 2019

Installing R and R Studio, Orientation to the Interface

Installation:

- Install R
- Install RStudio

Customization: options(), width, digits, etc.
Packages:
library() function
chooseCRANmirror()
install.packages("packagename") - ggplot2

Data Types, Operations, Manipulations, Functions

Primitive Types: Logical, Numeric, Character, and Factor
R is *object oriented*
Constructing Sequences with c(),rep(),seq() or colon shortcut
For homogeneous collections, R has types called vectors, matrices, and arrays
dim()
Selecting elements by position, exclusion, condition, membership (%in%), name
Element wise vs. Matrix operations
For heterogeneous collections, R offers types called lists and data frames
dim, class, and str functions

Built in Functions

get help ("?" or help())
write comments (#)
Transpose - t() - and Inverse - solve() - of matrices sort, rev, table, unique
min, max, mean, median, sum, sd, cor
floor, ceiling, trunc, round, sin, cos, exp, log, etc.
apply

Working with Dataframes

summary()
Loading and Saving Variables (with "load" and "save")
save all with save.image()
remove variables with rm()(remove *all* variables with ls())
read.csv(), write.csv(),read.table(),write.table()
Subsetting or "Slicing" Data - similar to vectors
Transforming Data: transform(dataframe, field=fieldtransformation)

Programming

.R files
custom functions, for loops, while loops, if else
source('filename.r')

What is Statistics?

Two Meanings - A discipline or a functional
Discipline of Statistics:

- Collection
- Summarization
- Inference

Modern discussion of Data Science as a field places some additional emphasis on data cleaning, computations, and communication (for example, through graphics)
This distinction is fraught

Base Graphics and ggplot2

Base graphics - plot, lines, barplot/hist, pairs, boxplot (admits formula, can be horizontal), qqplot
?par
ggplot2 - geoms, aesthetics, and mappings

Some Summarization and Inference

Basic Descriptive Statistics
mean, median, sd
Conditional Expectations

$$f(x) = Pr(X = x|Y = y)$$

Linear Models
Much more flexible than they sound due to available transformations
lm() function
t.test(), aov()
predict()
"All models are wrong. Some are useful." - George Box

knitr and R Markdown

Very useful for conducting reproducible research
Compatible with latex, more discussion next week
Specially indicated "code chunks" are added to a markup file
Sweave/knitr convert R code and output into a usable format for markup languages

Additional (Free) Resources

- R Studio Cheat Sheets! - Scroll to the Bottom for "Contributed Cheatsheets" including Base R
- R for Data Science
- An Introduction to Statistical Learning with Applications in R
- Dynamic Documents with R and knitr