R bootcamp

What, when and where

This is a full-day, hands-on, intensive introduction to R and its use for statistical analysis, graphics and programming, hosted by the CUNY School of Public Health. It will take place in a computer lab and you do not need your own computer. However you are welcome to use your own laptop; if you do, please come with R (www.cran.r-project.org) and RStudio (www.rstudio.com) pre-installed.

Location: Hunter College 68th St Campus, room 1001G North (Lab G)
Date: Saturday August 9th, 2014
Time: 9am-5pm

Instructors

Levi Waldron PhD, Assistant Professor of Biostatistics, CUNY School of Public Health, Hunter College, New York NY
Timothy Tickle PhD, Senior Software Developer, Broad Institute, Cambridge MA

Registration

There is a suggested minimum donation of $20 to the CUNY School of Public Health “SPH Fund” (https://www.paybill.com/v2/CUNYSchoolofPublicHealth/login.aspx) to support this course and other activities of the CUNY SPH Predictive Modeling Club.

RSVP by August 6 at http://goo.gl/WqsLmO. Please notify Levi Waldron (levi.waldron@hunter.cuny.edu) if must cancel.

Prerequisites

This course will assume knowledge of introductory statistics, but no prior experience in R.
Learning Objectives

- recognize key object classes in R
- Perform basic operations and function calls in R
- import / export data in common formats
- explore and describe an unfamiliar large dataset
- perform linear regression and analysis of variance
- write reproducible scripts to perform data analysis
- create figures and tables for presentation and publication
- create html reports and PDF presentations of analyses using “literate code”

Course Outline

Introduction (9-9:30)

- Survey of RStudio
- Markdown and Slidify for creating reports and presentation
- arithmetic, algebra
- Navigating directories
- Getting help

Data object classes (9:30-10)

- vector types: integer, numeric, character, factor, logical
- arrays, including matrices
- lists and dataframes
- S3 and S4 objects - how to recognize and find something out about them

Importing and exporting diverse data (10-10:30)

- to/from csv, txt, rda
- importing from Excel using library(xlsx) and from SAS/SPSS using library(foreign)
- Brief notes on: library(sqldf), library(data.table)
- Troubleshoot problem text files (mis-aligned columns, header and row names, variety NA strings, incorrect assumption of variable type)
Morning assignment (10:30-12)

Load datasets from several formats, do data cleaning, explore variable classes.

Lunch with an R video (12-1)

Data exploration and plotting (1-1:30)
- interactive data exploration through the R prompt
- principal components analysis
- histograms, boxplots, frequency + density plots, mosaic plots, contingency tables, stripcharts, heatmaps
- pairwise variable plots

Afternoon assignment #1 (1:30-3)
- Describe and prepare a short exploratory report about a complex dataset
- Customize figures, modify fonts and font sizes, choose different graphics devices.

Linear modeling HOWTO (3-3:30)
- Statistical sections will not teach statistical concepts, only how to do them in R
- model formulae
- t-test
- ANOVA/ANCOVA/multiple regression using lm()
- GLM interface

Afternoon assignment #2 (3:30-5:00)
- linear modeling
- diagnostic plots (e.g. residuals)
- present results using xtable