IDSC Workshop

R Workshop 1

Agenda

- R Universe
- Workshop Plan
- Math
- Getting Acquainted with R
- R Packages
- Probability Distributions
- References

R Universe

- Command line from CRAN (Comprehensive R Archive Network)
  - Regularly updated.
  - Quick to start, efficient to use.
  - Each update wipes out your installed packages.
- RStudio
  - Useful for “projects”, i.e. several scripts together with data and graphs.
  - Great for documentation and interactive applications
    - R Markdown
    - Shiny
  - Leverages the base installation from CRAN.
- Git and GitHub
- Tidyverse

R Workshop Evolution

- Initial: 2Q - 2017
  - Math
    - Probability
    - Distributions
  - Base R
  - Graphics
  - R Markdown
- Intermediate: 2H - 2017
  - Math
    - Statistics
    - Sampling
    - Inference
  - Base R
  - Tidyverse
  - Shiny
- Advanced: 2018
  - Math
    - model analysis
    - bias and variance
    - singular value decomposition
    - matrix formulations
  - Machine Learning
    - regression
    - clustering
    - trees
  - Tidyverse
  - APIs

Math

- It’s easy to blindly plug data into a model and obtain professional-looking results.
  - It’s so easy that many people do.
  - It can be very embarrassing later.
  - It could even lead to legal liability if conclusions are reached with negligence.
- A degree of mathematical maturity is required to evaluate the range of applicability for a model.
- We’ll devote some part of each workshop to developing mathematical skills related to probability and statistics.

Getting Acquainted

- Help
  - ?this
  - ??that
  - version
- Setup
  - getOption('max.print')
  - options(max.print=200)
- Types
  - numeric, character
  - Boolean (TRUE, FALSE)
  - NA – Not Available
  - factor
- Fancy calculator
  - assignments
  - vectors
  - indexing
  - seq and rep
- Blocks
  - if (length(x) > 3) { print(x) }
  - for (x in y) { }
  - It's not as common to loop over arrays given the abundance of available vector operations.
- Data Structures
  - list – heterogeneous contents
  - data.frame
  - data sets

R Packages

- There are many add-on packages that are useful.
- `install.packages()` - downloads and installs package
- `library()` – loads the package into an environment and inserts it as the parent of the global environment.
- Tips
  - library() does not use quotes, the install command does
  - some packages have native components that can require C compilation (not common)
  - sometimes a package can obscure commands from other packages

Probability Distributions

- Event Space
- Random variable
  - Discrete
  - Continuous
- CDF – Cumulative Distribution Function
  - Discrete – PMF (probability mass function) is the probability of the outcome for each value of the random variable.
  - Continuous – PDF (probability density function) is the derivative of the CDF.
  - Sometimes discrete case is also called PDF.
- Distribution Workshop

References

- R workspace has plenty build-in documentation
- Hadley Wickham
  - Advanced R
  - R for Data Science