Introduction

R for Data Science

Juan R Gonzalez
juanr.gonzalez@isglobal.org

BRGE - Bioinformatics Research Group in Epidemiology
ISGlobal - Barcelona Institute for Global Health
http://brge.isglobal.org

Credits

R for Data Science: http://r4ds.had.co.nz/index.html

Outline of the course

- Day 1: Reproducible Research (1h)
  - knitr
  - R markdown
- Day 1: Tidyverse - Data wrangling (1h)
  - Tibbles
  - Data import
  - Filter rows
  - Arrange rows
  - Select columns
  - Add columns
  - Grouped summaries
- Day 1: Tidyverse - Data visualization (1h)
  - Exploratory data analysis
  - Advanced graphics

- Day 2: Big data analysis (1h)
  - Parallelization in R
  - MapReduce
  - Linear regression for Big Data
- Day 2: Model fitting (1h)
  - Multivariate linear regression
  - General rules for variable selection
  - Stepwise variable selection
  - Comparing models
  - Automatic variable selection

- Day 3: Model fitting (2h)
  - Cross validation
  - K-fold and bootstrap cross validation
  - Missing data imputation
  - Regularization (Lasso and Elastic Net)

Course Material and tasks

- Slides, pdf, supplementary info at Moodle
- Material available at https://github.com/isglobal-brge/TeachingMaterials/tree/master/Master_Modelling
- Tasks
  - Daily exercises (Slides + Moodle)
  - Final exercise (Moodle)

Evaluation

- Tasks (50%)
- Final real data analysis (model building) (50%)