Reproducible Research (2)

Data import and manipulation

Zhuanghua Shi

28 April 2016

Github repository

```
git clone https://github.com/strongway/
seminar-reproducible-research.git
```

Data import

- We encounter many types of data formats
  - csv, excel, matlab, spss, web table, Sql
- R {utils} provides basic import methods
  - read.csv
  - read.delim
  - read.table
- data.table package provides powerful function
  - fread()
- readr package (fast)
  - read_csv
  - read_delim
  - read_tsv

import text file basic functions

read.table(filename, header=TRUE, sep=',', stringsAsFactors = FALSE, col.names = c(...))

- specific functions:
  - read.csv()
  - read.csv2()
  - read.delim()
  - read.delim2()
- helper function
  - file.path()
  - past0()

Comparison between utils and readr

utils	readr
read.table()	read_delim()
read.csv()	read_csv()
read.delim()	read_tsv()

fread() from data.table

- super powerful, fast
- automatically manage column names
- infer column types and separators
- select or drop columns

Practise

Hands-on data from github

Import matlab file

Two options for importing matlab data

- convert data format to csv within Matlab
- using package 'R.matlab'
    - readMat

Import multiple data

A good way to import multiple data set is:

- write a function of importing individual data
- using lapply() or Map()

Some helper functions

- dir() list the files in a specific folder
- do.call(rbind, a_list) combine data together

Data cleaning and exploration

- data exploration
  - hist()
  - boxplot()
  - summary()

Data manipulation

Two important packages

- dplyr - grammar-like manipulation
- data.table - sql-like manipulation

I prefer the former for its clearity.

Functions in dplyr

- filter and arrange
- group_by
- select and mutate
- summarise
- pipe operator (Important feature)

tidyr functions

- gather(data, key, value, . . .)
- spread(data, key, value)
- separate(data, key, colnames)
- unite(data, col, . . . , sep)

Course schedule

dates	contents
2016-04-14	Version control system - git
2016-04-28	Data import and manipulation
2016-05-12	Publication-ready figures
2016-05-26	Statstics and Modeling
2016-06-09	Reusable data analysis
2016-06-30	RMarkdown and writing