Note 1. Introduction

Introduction to Statistical Programming

Chanmin Kim

Department of Statistics
Sungkyunkwan University

2022 Spring

Computing in Statistics

- Statistics = Mathematics + Computer Science.
- Size of data $\uparrow \Rightarrow$ Data analysis.
- Complex methods $\Rightarrow$ Numerical/Sampling methods.
- Big data $\Rightarrow$ Parallel/Cloud computing.
- Computationally intensive methods.

Statistical Softwares

- SAS: Commercial and analytic software (Business solution).
- SPSS: Commercial software for social science.
- R: Free software for statistical computing & graphics.
- Minitab, S-plus, Stata, JMP, Octave, etc.

If you get an industry job related to statistics,

- Data handling:
  - Extracting data from database.
  - Missing data.
  - Changing variables.
  - Creating new variables.
  - Splitting / Merging datasets, etc.
- Data analysis:
  - Descriptive statistic.
  - Applying models to data.
  - Visualization.
- Report: Summarizing results.

If you get an academic job,

- Complex / unstructured data: Correlated data, functions, images, text.
- Complex models: No analytical solution $\Rightarrow$ Computational methods.
- Computational statistics: Development of Fast/efficient algorithms.
- Simulation.

History of R

- S language:
  - developed as an internal statistical analysis environment by AT&T lab in 1976.
  - S-plus (Commercial statistical software).
  - Data analysis.
  - Interactive environment.
  - For user & developer.

- R language:
  - Created in 1991.
  - Similar syntax to S.
  - Free software.
  - The R Core Group formed in 1997 controls the source code for R.

Features of R

- Available in various operating systems (e.g., Windows, Mac, Linux, etc.) ($\because$ open source nature).
- Sharing with many popular open source projects $\Rightarrow$ active development $\Rightarrow$ frequent releases.
- Statistical analysis purpose + General purpose programming.
- Interactive environment.
- Object-oriented & functional programming language.
- Active & vibrant user community $\Rightarrow$ Development of platforms or packages.

Object-oriented & Functional Programming

- Object-oriented programming:
  - You can pick and choose parts of an object.
  - Polymorphic $\Rightarrow$ Generic function (e.g., `plot()` can be used for both data objects and regression objects).

- Functional programming:
  - Avoidance of explicit iteration.
  - More compact code.
  - Potentially much fast execution speed.
  - Less debugging time.
  - Easier transition to parallel programming.

R System

- Base R system: Download from CRAN (Comprehensive R Archive Network).
- Packages:
  - Over 10,000 packages on CRAN.
  - Many packages from Bioconductor project.
  - Personal packages.

Limitation of R

- Memory problem: R objects must be generally stored in physical memory $\Rightarrow$ More memory hog than other statistical software.
- Relatively slow speed, especially for loop statement.
- Functionality depends on customer demand and voluntary user contribution.