Development Data Boot Camp
Introduction and Preparation: Programming Languages

Ge Sun

University of Notre Dame

May 12, 2023

Outline

Outline

Why Programming?

- Why do we need to learn how to program? (Isn't Excel great?)
  - Big data and greater computer power are available to us.
    - Vast data set are less manageable in spreadsheets.
  - It's nice to process data through codes.
    - keep the original raw data unchanged
    - keep track of every step we make
    - easy to check and correct in the future
    - **can re-run it every time we need!**
  - Coding gives us flexibility to accomplish more tasks.
    - Not constrained by the previously designed "icon", on which we can just click.

Classification of different programming languages

- Dynamic Language
  - Examples: Matlab, R, Python, Stata, etc.
  - Benefits: ease of development
    - line-by-line execution in real-time
    - user-friendly features built-in
- Static Languages
  - Examples: Fortran, C, C++, Java
  - Benefits: speed
    - packages (complies) entire source code before executing
  - Costs: Painful user interface and long development time
- Something new and in-between: Julia

Outline

Stata

- Stata: A powerful software for data analysis.
  - to analyze, manage, and produce graphical visualizations of data
  - primarily used in economics, biomedicine, and political science
- Strengths:
  - It is easily accessible to beginners through its graphical user interface (GUI) and has many useful commands.
  - For economists, Stata is a suitable (or the best) tool for regression analysis, especially with handling fixed effects, clustering, and instrumental variable (IV) estimations.
  - It has good support from its community, very good documentation, and reliable results.
- Weakness:
  - It is a commercial software and is not available for free. It is not open source.
  - It can be slow to incorporate new methods.
  - It is not a true "programming" language. It has some limitations in completing certain tasks.
- Official website: https://www.stata.com

Stata

- How to access Stata in ND?

Matlab

- Matlab stands for Matrix Laboratory
  - widely used by engineers and scientists for various applications, including image processing, matrix manipulation, machine learning, and signal processing
  - for economists: macro modelling, time series data analysis
- Strengths:
  - Matlab is a programming language that allows users to write their own functions to perform sophisticated tasks.
  - It is known for being quick in numeric calculations and analysis.
  - MathWorks provides good support for Matlab.
- Weakness:
  - In order to fully leverage the power of Matlab, data often needs to be vectorized.
  - Matlab is a commercial software and is not available for free. It is not open source.
- Official website: https://www.mathworks.com

R

- R: commonly used by statisticians, as well as professionals in fields such as economics and political science.
- Strengths:
  * It has a vast collection of packages for statistical analysis. It is often the first place where new statistical technologies and concepts are introduced and implemented.
  * R is a free and open-source programming language, which means that there is a vast amount of open-source code available online for free.
- Weakness:
  * The abundance of packages in R can lead to differences in grammar and functionality across different packages, even for the same task.
  * The quality of documentation may vary depending on the individual developers creating the packages.
  * The accuracy and credibility of new packages may not always be guaranteed.
- Official website: https://www.r-project.org
- A must-know and game-changer package: Tidyverse (https://www.tidyverse.org)

Python

- Python: a high-level, general-purpose programming language
  - a versatile programming language that is good at machine learning, web crawling, and building datasets
  - For data science tasks, there are several Python libraries, including:
    - *Numpy* for handling large dimensional arrays
    - *Pandas* for data manipulation and analysis
    - *Matplotlib* for building data visualizations
- Strengths:
  - Python is an open-source programming language supported by large communities and favored by many programmers.
  - It emphasizes code readability, making it easier to learn as a multipurpose language compared to C++ and Java.
- Weakness:
  - Python is not designed specifically for statistical purposes, so not all statistical methods are available.
- Official website: https://www.python.org

Julia

- Julia: "as fast as C, but as easy for statistics as R"
  - runs dynamically, but compiles as it goes
  - slow in first-time execution, but very fast in subsequent executions
- Strengths:
  - open source and free
  - very good at handling "functions", making it a lot easier to build and solve a structural model
  - as it says, it is quick
- Weakness:
  - not very long history, syntax in old versions may change
  - the documentation and resources are still developing and not as extensive as some other languages
- Official website: https://julialang.org

Some blogs as reading materials:

- Why do we create Julia?
- Which numerical computing language is best: Julia, MATLAB, Python or R?
- What's the Best Statistical Software? A Comparison of R, Python, SAS, SPSS and STATA