ATENEO DE MANILA UNIVERSITY

DEPARTMENT OF ECONOMICS

Code & Conquer:

Crash Course on Alternative Econometric Software

November 15, 2024

SOSS Sanggu / AEA / MEcO

dsaplala@ateneo.edu

(edited Nov 16, 2024)

Outline

- Introduction: STATA vs R vs Python

Python (Disclosure: lecture materials and activities were prepared with the help of ChatGPT)

- Basic Commands: Loading Data and Basic Descriptive Statistics
- Basic Data Wrangling: Generating new variables, Merging Datasets
- Basic Econometrics: OLS Estimation and Diagnostics
- (time permitting) Data Visualization (plotly) and Time Series Analysis

Sources and Further Reading

Introduction

About STATA (from Wikipedia)

Stata is a general-purpose statistical software package developed by StataCorp for data manipulation, visualization, statistics, and automated reporting.

- It is used by researchers in many fields, including **economics**, sociology, political science, biomedicine, and epidemiology.
- Stata was initially developed by Computing Resource Center in California and the first version was released in 1985.
- In 1993, the company moved to College Station, TX and was renamed Stata Corporation, now known as StataCorp.
- The latest version of Stata (18) was released last April 2023.

STATA: Pros and Cons

- **”Gold standard” for economic research**
  - Written and programmed by Econometricians (from STATA Corp)
  - Most employers at top economic research institutions recognize the program.
- **Excellent STATA documentation & customer support**
  - including with a lot of economics examples
  - With free webinars organized by STATA Corp
- **Beginner-friendly (point-and-click features available)**
- At the same time, has a corresponding programming language (.do files, good for reproducibility purposes)

Disadvantage / Cons

- **Proprietary (i.e. not free)**
- Good speed for most datasets used by economists (e.g. PSA data) but can slow down in the case of “big data”
- For tasks beyond economics, subpar (or at least developing)
  - Examples: Machine learning, Geospatial analysis, web scraping, etc
  - Note: Python integration available starting STATA 16

About R (from Wikipedia, ChatGPT)

R is a free software environment for **statistical computing and graphics**. It is widely used among statisticians and data miners for developing statistical software and data analysis.

- R is maintained by the R Core Team and is a GNU project, reflecting its commitment to free and open-source software collaboration.
- It is highly extensible and features a wide array of packages for various types of data analysis, contributed by users worldwide.
- The R programming language was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and the first version was released in 1995.
- R is part of the R Foundation and the R Consortium, which work to support the R community and foster the development of R.

R: Pros and Cons

Profile / Pros

- R is a FREE programming language primarily designed for **Statistics**
- Fast and intuitive syntax especially for statistical tasks
- Like for stat, R has extensive features for econometrics too (especially modern R packages)

Disadvantage / Cons

- Greater learning curve compared to STATA. (in some cases, compared to Python too)
- Fragmented approach to tasks (e.g. different approaches from different packages especially when comparing different eras of R)
  - R is fairly old (1995)

About Python (from Wikipedia, ChatGPT)

Python is a high-level, interpreted programming language renowned for its simplicity and versatility, enabling rapid development of diverse applications.

- Created by Guido van Rossum and first released in 1991, Python emphasizes readability and clear syntax, making it accessible to beginners and powerful for advanced users.
- It supports multiple programming paradigms, including procedural, object-oriented, and functional programming, ensuring flexibility for developers.
- Python is maintained by the Python Software Foundation (PSF) and operates under an open-source license, encouraging community-driven development and collaboration.
- The language boasts an extensive ecosystem of libraries and frameworks
  - e.g., NumPy, Pandas, Plotly, TensorFlow, Pytorch

Python: Pros and Cons

Profile / Pros

- Python is a FREE programming language primarily designed for **Computer Science** tasks
- Popular industry standard especially for machine learning, computer science, **and artificial intelligence (e.g. most GPT models written in Python)**
- Intuitive syntax (e.g. shortest for printing statements, compared to C++, Java, etc)

Disadvantage / Cons

- Good statistical and econometrics packages, although not primarily designed for such (so some simple stat tasks can require more lines, etc than R/STATA counterparts)
- Greater learning curve compared to STATA. In some cases, compared to R too.
- Python packages are more uniform compared to R, but overall, still fragmented compared to STATA

Summary

	STATA	R	PYTHON
Primary Disciplines Involved	Economics / Econometrics - Also popular in epidemiology and other social sciences	Statistics - and any field where Statistical tools are frequently used	Computer Science (CS) – including Machine Learning (ML) & Artificial Intelligence (AI)
Cost	Proprietary	Free (open-source)	
Ease of Use	**Beginner-friendly** with both point-and-click and .do programming	**steeper learning curve** -but may vary depending on the packages used	**medium learning curve** -but may vary depending on the packages used
Other Strengths	**Economic research,** reproducible workflows with simple syntax	**Stats-focused** (simple syntax with some stat tasks) **Fast, modern packages are flexible**	**Versatility. ML & AI industry standard.** Simple syntax with some basic CS/ML/AI tasks.

Summary

	STATA	R	PYTHON
Other Weaknesses	**Limited/subpar beyond economics** (or still developing)	**Fragmented ecosystem** • diff. packages from different eras may totally differ in terms of syntax/approach even for same task	**Not the most efficient with some econometric tests.** -Not primarily designed for statistics and econometric tasks.
Speed	**Good for typical economic datasets.** Slower and more expensive for much larger ones.	**Faster with "big data"** (e.g. image datasets, typical ML-size datasets)	
Community Support (e.g. Stack Overflow responses, Software Documentation, etc)	**Strongest in economics,** usually more unified in terms of approach	**Strong in various disciplines** although support types can vary or get fragmented	**ML, AI industry standard.** **- Lots of support available online especially with popular packages**
Integration Capabilities	**With Python integration within STATA**	**With Python integration** within Google Colab, RStudio, VS Code, various IDE	**Integrates with many tools**
Best For	**Econometric analysis**	**Statistical modeling**	**CS, ML, AI**

Sources and Further Reading

Sources and Further Reading

PYTHON

- Arthur Turrell. “Coding for Economists”. https://aeturrell.github.io/coding-for-economists/intro.html

R

- Benjamin S. Baumer, Daniel T. Kaplan, and Nicholas J. Horton. “Modern Data Science with R”. https://mdsr-book.github.io/mdsr3e/

STATA

- www.stata.com, especially the video tutorials: https://www.stata.com/links/video-tutorials/ and the free webinars

Sources and Further Reading

Comparing the Programs/Languages

- https://www.youtube.com/watch?v=ZFsPEnsq3bQ&ab_channel=Econometrics%2CCausality%2CandCodingwithDr.HK

R, Python, Julia (with Jeffrey Wooldridge's book)

- Florian Heiss. Using R, Python, Julia for Introductory Econometrics. http://upfie.net/

Sources and Further Reading

Extras

Comparing the Programs/Languages (if you have time watching a 1.5 hour debate show)

- **https://www.youtube.com/watch?v=-z67WeJUdkY&ab_channel=CentralEuropeanUniversity**
- **There's a Slido poll answered by the audience of the debate at the end.**

Sources and Further Reading

Artificial Intelligence

Try prompting ChatGPT (or any AI), questions like:

Stata vs R vs Python

Shortest and most efficient? In terms of:

- General Statistics
- Testing for multicollinearity, heteroskedasticity
- Web scraping
- Geospatial analysis
- Machine learning
- Deep learning