R
Története, alapjellemzők

Csima Judit

BME, VIK,
Számítástudományi és Információelméleti Tanszék

2018. február 5.

Az R gyökerei

- az R az S nyelv egy dialektusa
- az S egy programozási nyelv és környezet statisztikai számításokhoz
- John Chambers (az S nyelv fő fejlesztője):
  "[W]e wanted users to be able to begin in an interactive environment, where they did not consciously think of themselves as programming. Then as their needs became clearer and their sophistication increased, they should be able to slide gradually into programming, when the language and system aspects would become more important."
  http://www.stat.bell-labs.com/S/history.html

Az R története

- 1991: Ross Ihaka and Robert Gentleman (Új-Zéland)
- 1995: meggyőzik Ihaka-t és Gentleman-t hogy az R legyen open source, szabad szoftver (GNU)
- 1996: Az R-help és R-devel levlisták megszületnek
- 2000: R 1.0.0 verzió
- 2017. november 30.: R 3.4.3. (gyakori frissítések)

Az R jellemzői

- szinte minden platformra van (Windows, Linuxok, Mac)
- rendszeres fejlesztések, hibajavítás
- moduláris felépítés, package-ek
- fejlett grafikai lehetőségek
- interaktív és programozási használat (fokozatos átmenet)
- aktív közösség (R-help, R-devel listák)
- ingyenes (http://www.fsf.org)

Az R felépítése

- package-ek
- "alap" R: ebben minden benne van, ami kell, ha nem akar az ember spéci dolgokat
- például az "alap" R-ben van a **base** package, ami a legalapvetőbb függvényeket tartalmazza
- további "alap" package-ek: **utils, stats, datasets, graphics, grDevices, grid, methods, tools, parallel, compiler, splines, tcltk, stats4**
- letölthető még kb. 12000 package (CRAN, Bioconductor, gitHub)
- letöltések, sok hasznos infó: http://cran.r-project.org
- R főoldal: http://www.r-project.org/

Az rstudio

- egyben minden, ami kellhet, ha az ember R-t akar használni
- szövegszerkesztő kódíráshoz
- konzol az interaktív munkához
- help-ablak
- mindenféle platformra, letölthető: http://www.rstudio.com/

▸ Különböző adatelemzési eszközök összehasonlítása, blog

AI and Social Science – Brendan O'Connor
cognition, language, social systems;
statistics, visualization, computation

Comparison of data analysis packages: R, Matlab, SciPy, Excel, SAS, SPSS, Stata

Posted on February 23, 2009

Lukas and I were trying to write a succinct comparison of the most popular packages that are typically used for data analysis. I think most people choose one based on what people around them use or what they learn in school, so I've found it hard to find comparative information. I'm posting the table here in hopes of useful comments.

Name	Advantages	Disadvantages	Open source?	Typical users
R	Library support; visualization	Steep learning curve	Yes	Finance; Statistics
Matlab	Elegant matrix support; visualization	Expensive; incomplete statistics support	No	Engineering
SciPy/NumPy/Matplotlib	Python (general-purpose programming language)	Immature	Yes	Engineering
Excel	Easy; visual; flexible	Large datasets	No	Business
SAS	Large datasets	Expensive; outdated programming language	No	Business; Government
Stata	Easy statistical analysis		No	Science
SPSS	Like Stata but more expensive and worse			

[**7/09 update:** tweaks incorporating some of the excellent comments below , esp. for SAS, SPSS, and Stata.]

There's a bunch more to be said fo r every cell. Among other things:

- Two big divisions on the table: The more programming-oriented solutions are R, Matlab, and Python. More analytic solutions are Excel, SAS, Stata, and SPSS.
- Python "immature": matplotlib, numpy , and scipy are all separate libraries that don' t always get along. Why does matplotlib come with "pylab" which is supposed to be a unified namespace for everything? Isn' t scipy supposed to do that? Why is there duplication between numpy and scipy (e.g. numpy .linalg vs. scipy.linalg)? And then there's package compatibility version hell. Y ou can use SAGE or Enthoug ht but neither is standard (yet). In terms of functionality and approach, SciPy is closest to Matlab, but it feels much less mature.

▶ Különböző adatelemzési eszközök összehasonlítása

NYU
DataServices

Which Statistical
Software to use?

Summary

Each statistical software is different with their own advantages and disadvantages. The important part is choosing which software is right for you and your data.

Software Access

Software	Mac/Windows	DS Lab	VCL	HPC	Personal Access at NYU
SPSS	Both	✔	✔		Purchase via NYU Computer Store
Stata	Both	✔	✔	✔	Purchase via Stata Grad Plan
SAS	Windows	✔	✔	✔	Purchase via NYU Computer Store
R	Both	✔	✔	✔	Free download via CRAN website
MATLAB	Both	✔	✔	✔	Contact hpc@nyu.edu

Software Features and Capabilities

Software	Interface*	Learning Curve	Data Manipulation	Statistical Analysis	Graphics	Specialties
SPSS	**Menus** & Syntax	Gradual	Moderate	Moderate Scope Low Versatility	Good	Custom Tables, ANOVA & Multivariate Analysis
Stata	Menus & **Syntax**	Moderate	Strong	Broad Scope Medium Versatility	Good	Panel Data, Survey Data Analysis & Multiple Imputation
SAS	Syntax	Steep	Very Strong	Very Broad Scope High Versatility	Very Good	Large Datasets, Reporting, Password Encryption & Components for Specific Fields
R	Syntax	Steep	Very Strong	Very Broad Scope High Versatility	Excellent	Packages for Graphics, Web Scraping, Machine Learning & Predictive Modeling
MATLAB	Syntax	Steep	Very Strong	Limited Scope High Versatility	Excellent	Simulations, Multidimensional Data, Image & Signal Processing

* The primary interface is bolded in the case of multiple interface types available.

Learning Curve

Fordítás