Az R statisztikai és grafikai környezet

Jeszenszky Péter
Debreceni Egyetem, Informatikai Kar
jeszenszky.peter@inf.unideb.hu

Mi az R?

- Nyílt forrású statisztikai és grafikai környezet
  - Programozási nyelv és interaktív környezet egyben
- Az S programozási nyelv implementációjának tekinthető
  - A megvalósításon a *Scheme* programozási nyelv hatása érezhető
- Nyílt forrású szoftverként bárki számára ingyenesen hozzáférhető
  - Ez nem gátja az üzleti célú felhasználásnak

Jellemzők (1)

- Eszköztára megszámlálhatatlan klasszikus és modern statisztikai és matematikai eljárást tartalmaz
  - Az alaprendszer részeként vagy csomagokban
- Kimagasló grafikus lehetőségekkel rendelkezik
- Kitűnően dokumentált
- Többféle platformon működik
  - Futtatható (bináris) formában tölthető le a szoftver Windows, Mac OS X és Linux operációs rendszerekre
  - Egyéb platformok esetében a forrásokból kell a rendszert lefordítani

Jellemzők (2)

- Interpretált nyelv
  - A felhasználók által látható függvények nagy része R-ben készült
  - Számításigényes feladatok megoldásához C, C++ és FORTRAN eljárásokat is meg lehet hívni
- Mint programozási nyelv a funkcionális és objektum-orientált programozási paradigmákat ötvözi
- Elegáns és tömör szintaxis jellemzi

Jellemzők (3)

- Tudás és sebesség tekintetében is méltó versenytársa a hasonló célú kereskedelmi programoknak
- Egyszerűen elsajátítható a használata, hamar megszerethető (szubjektív vélemény)

Fejlesztés

- Az eredeti fejlesztők Ross Ihaka és Robert Gentleman (Department of Statistics, University of Auckland, Új-Zéland)
- A fejlesztést 1997 óta az *R Development Core Team* tartja kézben
  - A csoport tagja John Chambers, az S programozási nyelv atyja
- Az *R Foundation* az *R Development Core Team* tagjai által alapított non-profit szervezet
  - Célja az R projekt támogatása
  - Az R-rel kapcsolatos copyright jogok tulajdonosa (szoftver és dokumentáció)

Elnevezés

- Az R az eredeti szerzők nevének kezdőbetűje
- Játék az S programozási nyelv nevével
- GNU S-nek is nevezik, mivel az R hivatalosan is a GNU projekt része

Nyílt forrású szoftver

- GNU GPL licenc alatt terjesztik http://www.gnu.org/copyleft/gpl.html
  - A legismertebb és legelterjedtebb nyílt forrású szoftver licenc
- A legtöbb csomag szintén ilyen licenc alatt hozzáférhető, de lehet eltérés
  - Esetenként előfordulhat, hogy az üzleti célú felhasználás korlátozott (üzleti felhasználás esetén érdemes megnézni a csomag licencét)
  - A nem üzleti célú felhasználás azonban nincs korlátozva (természetesen a GPL-nek megfelelően kell eljárni)

GNU GPL

- A program szabadon másolható, terjeszthető (akár pénzért is) és módosítható
- Az eredeti programot és módosított változatait is forráskód formájában, a licenc alatt kötelező terjeszteni
- Ez biztosítja, hogy valamennyi az eredeti programból származtatott mű is szabad maradjon
- Nem zárja ki az üzleti célú felhasználást
  - Csak a másolás, terjesztés és módosítás tartozik a licenc hatálya alá

CRAN

- The Comprehensive R Archive Network
  http://cran.r-project.org/
- Ugyanazt a tartalmat kínáló, földrajzilag a világ különböző részein található FTP és webszerverek hálózata, valamennyi szerveren elérhető:
  - A legutóbbi stabil és az összes korábbi R verzió több platformra, bináris (telepíthető) formában
  - Az összes korábbi és jelenleg fejlesztés alatt álló verzió forráskódban
  - A teljes dokumentáció
  - Az összes csomag

Telepítés

- A jelenleg aktuális verzió az R 2.14.1 (megjelenés: 2011. december 22.)
- A CRAN szerverekről lehet letölteni bináris, futtatható formában a telepítőcsomagokat Windows, Linux és Mac OS X platformra
  - Más platformra forrásból telepíthető
- Windows környezetben a telepítés után rendelkezésre áll egy RGui nevű grafikus felhasználói felületet, amelyet a Linux verzió nem tartalmaz

Dokumentáció

- A telepítés után elérhető egy nagyon részletes és átfogó online dokumentáció, amely több kézikönyvből áll, valamint a telepített csomagok dokumentációjából
  - A dokumentáció többféle módon kereshető, böngészőprogramban is megjeleníthető
- Az R saját, a LaTex-hez hasonló dokumentációs formátumot használ
  - A dokumentációs oldalak átalakíthatók LaTeX (így PDF) és HTML formátumba, közönséges szöveges állományokká

Felhasználói közösség

- Népes a felhasználók tábora
  - Számtalan statisztikus, biológus, közgazdász, orvos és pszichológus használja
- Az *R-help* levelezési listán kérhető segítség (feliratkozás szükséges) http://stat.ethz.ch/mailman/listinfo/r-help
  - Merjünk kérdezni, szívesen segítenek, általában nem kell sokat várni a válaszra
  - Nagy forgalom, naponta akár 100 levél

Kapcsolódó folyóiratok

- Journal of Statistical Software http://www.jstatsoft.org/
  - *Volume 18 – Special Volume: Spectroscopy and Chemometrics in R*
  - *Volume 20 – Special Volume: Psychometrics in R*
  - *Volume 22 – Special Volume: Ecology and Ecological Modelling in R*
  - *Volume 27 – Special Volume: Econometrics in R*
  - …
- The R Journal http://journal.r-project.org/

Használat

- A legtöbb felhasználó interaktív módon használja, de természetesen lehetőség van programok készítésére és futtatására
  - Hibák felderítéséhez van nyomkövetési lehetőség

S

- Statisztikai programozási nyelv, amelynek kifejlesztése elsősorban John Chambers nevéhez fűződik
- Története az 1970-es évek közepéig nyúlik vissza
- Több verzió létezett az idők folyamán, a legutóbbi az S Version 4 (S4), amelynek leírása az alábbi könyvben („*Green Book*”):
  - John M. Chambers (1998), *Programming with Data: Guide to the S Language*. New York: Springer.
- Két mai, modern implementációja létezik a nyelvnek, az R és az S-PLUS

S-PLUS

- Az S programozási nyelv kereskedelmi implementációja
  - Az Insightful Corporation terméke http://www.insightful.com/
  - A jelenleg aktuális verzió a 2007-ben megjelent S-PLUS® 8
- Comprehensive S-PLUS Archive Network http://csan.insightful.com/
  - A legtöbb csomag nyílt forrású és elérhető a CRAN szervereken is (eleve úgy készülnek, hogy működjenek mindkét környezetben)

Különbségek az R és S között (1)

- Mivel az R az S implementációjának tekinthető, az eltérés az R és az összes többi S implementáció (így a különböző S-PLUS verziók) között értendő
- Bizonyos eltérések oka az, hogy a fejlesztők az S viselkedését esetenként nem tartották letisztultnak (logikusnak, konzekvensnek, pontosan tisztázottnak, …)
  - A cél egy letisztultabb, ugyanakkor az S-sel a lehető legnagyobb mértékben kompatibilis implementáció létrehozása volt

Különbségek az R és S között (2)

- Olyan grafikus lehetőségekkel is rendelkezik az R, amelyekkel a többi implementáció nem:
  - Nem csupán a beépített vonaltípusok állnak rendelkezésre, hanem tetszőleges vonaltípus megadható
  - Az S-PLUS 8-ban csak 8 beépített vonaltípus használható
  - Fejlettebb színkezelés (például gamma-korrekció)
  - Az R a TeX-hez hasonlóan képes matematikai formulákat megjeleníteni
  - …
- Viszont az S-PLUS is rendelkezik néhány olyan grafikus lehetőséggel, amelyekkel az R nem

Különbségek az R és S között (3)

- A legfontosabb eltérés az, hogy az összes többi implementációhoz képest az R hatáskörkezelése statikus
  - A statikus hatáskörkezelés következményeként az R valamennyi objektumot a memóriában tárol
    - Emiatt gyorsabb
    - Azonban R összeomlása esetén valamennyi adat elvész, amennyiben nem végeztünk explicit módon mentést

Az R-re épülő kereskedelmi szoftverek

- Revolution Analytics
  http://www.revolutionanalytics.com/
  - Revolution R Community, Revolution R Enterprise: bináris formában elérhető, optimalizált R disztribúciók
  - ParallelR: az RPro többprocesszoros környezetre szabott, párhuzamos feldolgozást támogató változata
- R-PLUS (XLSolutions Corporation)
  http://www.experience-rplus.com/

Összehasonlítás hasonló szoftverekkel (1)

- Speed comparison of various number crunching packages (version 2) (08/03/2003) http://www.sciviews.org/benchmark/
  - Sajnos régi
  - Az alábbi programok összehasonlítása: R 1.9.0, S-PLUS 6.1, Matlab 6.0, O-Matrix 5.6, Octave 2.1.42, Scilab 2.7, Ox 3.30
  - Sokféle teszt (FFT, sajátérték számítás, mátrix invertálás, rendezés, …)
  - Nagyon kedvező eredmények az R-re nézve

Összehasonlítás hasonló szoftverekkel (2)

- Matlab vs. R performance benchmarking http://mlg.eng.cam.ac.uk/dave/rmbenchmark.php
  - Matlab 2008b és R 2.8.0 összehasonlítás (mátrixműveletek, FFT, sajátérték számítás, …)
- Egy sokrétű összehasonlítás (rendelkezésre állás, támogatott platformok, eszköztár) http://en.wikipedia.org/wiki/Comparison_of_statistical_
  - Sajnos nincs forrásmegjelölés, máshol pedig nem lelhető fel

Csomagok (1)

- A rendszer telepítése során a számítógépre kerülnek az alap- és ajánlott csomagok
  - Számos további csomag érhető el a CRAN szerverekről és egyéb helyekről, amelyek további képességekkel bővítik a rendszer
- Csomagok telepítéséhez használjuk az interpreterben az `install.packages()` függvényt
  - Windows platformon csomagok telepítése elvégezhető az RGui **Packages** menüjének **Install package(s)** menüpontjával is

Csomagok (2)

- Windows és Mac OS X platformra a csomagok bináris formában kerülnek letöltésre
  - A források természetesen megtalálhatók a CRAN szervereken
- Linux platformra forrásban történik a csomagok letöltése, a C, C++ és FORTRAN források lefordítása lokálisan
  - Ehhez tipikusan rendelkezésre állnak a gcc (C, C++) és g77 (FORTRAN) fordítóprogramok
- Csomagfüggőségek kezelése

Alap- és ajánlott csomagok

- Az alábbi 12 alapcsomag alkotja az R-t:
  - `base`, `datasets`, `grDevices`, `graphics`, `grid`, `methods`, `splines`, `stats`, `stats4`, `tcltk`, `tools`, `utils`
- Valamennyi bináris disztribúció tartalmazza továbbá az alábbi 15 ajánlott csomagokat:
  - `boot`, `class`, `cluster`, `codetools`, `foreign`, `KernSmooth`, `lattice`, `MASS`, `Matrix`, `mgcv`, `nlme`, `nnet`, `rpart`, `spatial`, `survival`

További csomagok (1)

- További csomagokat a CRAN szervereken találunk
  - Ezeken jelenleg 3600-nál több csomag áll rendelkezésre!
  - Egyéb projektek keretében készült csomagok nincsenek feltüntetve, így az összes R csomag száma nagyobb ennél!

További csomagok (2)

- Bioconductor (bioinformatikai R csomagok) http://www.bioconductor.org/
  - 500-nál több további professzionális R csomag
- R-Forge http://r-forge.r-project.org/
- The Omega Project for Statistical Computing http://www.omegahat.org/

Csomagok tematikus csoportosítása

- Az átláthatóság érdekében a CRAN szervereken a csomagokat tematikusan csoportosítva is lehet böngészni (nézetek)
  - Jelenleg 28 nézet (`Finance`, `Graphics`, `MachineLearning`, `MedicalImaging`, `NaturalLanguageProcessing`, `Optimization`, …)
- Érdemes telepíteni a nézetek kezelését támogató `ctv` (CRAN Task Views) csomagot

Egy hasznos csomag csomagok telepítéséhez

- Nézetek kezelését támogatja a `ctv` csomag
  - A rendelkezésre álló nézeteket az `available.views()` függvénnyel lehet listázni
  - Az `install.views()` függvénnyel lehet telepíteni az adott csoportba tartozó csomagokat
    - Például az `install.views("MachineLearning")` paranccsal lehet telepíteni a gépi tanuláshoz kötődő csomagot

IO lehetőségek

- Objektumok tárolása állományokban
- Szöveges állományok
- Csatlakozás adatbázis-kezelő rendszerekhez
- Hálózati kommunikáció
- Excel állományok feldolgozása
- Egyéb IO lehetőségek

Objektumok tárolása állományokban

- Az objektumokat állományokba lehet menteni
  - A `save()` függvény az argumentumként adott objektumokat menti
  - A `save.image()` függvény valamennyi a memóriában tárolt objektumot menti
- Az így elmentett állományokat a `load()` függvénnyel lehet a memóriába betölteni
- A tárolás az R saját bináris formátumában történik

Szöveges állományok

- Táblázatos adatokat lehet beolvasni szöveges állományokból a `read.table()`, `read.csv()` és `read.delim()` függvényekkel
- A `write.table()` és `write.csv()` függvényekkel lehet táblázatos adatokat szöveges állományokba írni
- Ez a legbiztonságosabb módja a más szoftverekkel állományokon keresztül történő kommunikációnak
  - Ilyen állományokat minden szoftver tud írni és olvasni

Csatlakozás adatbázis-kezelő rendszerekhez (1)

- Több csomag áll rendelkezésre, amelyek különböző absztrakciós szintűek
  - Mindegyik lehetővé teszi SQL lekérdezések végrehajtását és a teljes eredménytábla beolvasását, vagy az eredménytábla részekben beolvasását
  - Némelyik lehetővé teszi adatok beolvasását és kiírását SQL közvetlen használata nélkül

Csatlakozás adatbázis-kezelő rendszerekhez (2)

- RODBC csomag:
  - Csatlakozás ODBC interfészen keresztül (szinte minden adatbázis-kezelő rendszer támogatja)
    - MS SQL Server, MS Access, MySQL, Oracle PostgreSQL, …
  - A Windows még Excel és szöveges állományokhoz is biztosít ODBC meghajtót
    - Ez lehetővé teszi Excel táblák feldolgozását (nem szükséges, hogy az Excel telepítve legyen)
  - Windows rendszerekben az ODBC támogatás általában alapértelmezésben telepítve van
  - Linux környezetben ODBC meghajtó kezelő program telepítése szükséges
    - Két nyílt forrású megoldás: unixODBC http://www.unixodbc.org/, iODBC http://www.iodbc.org/

Csatlakozás adatbázis-kezelő rendszerekhez (3)

- DBI csomag:
  - Speciálisan az R és az adatbázis-kezelő rendszerek közötti kommunikáció megvalósításához készült interfész csomag
  - Front-end csomag, valamennyi adatbázis-kezelő rendszerhez megfelelő back-end csomag szükséges
    - Jelenleg rendelkezésre álló back-end csomagok: `RmySQL`, `ROracle`, `RPostgreSQL`, `RSQLite`

Hálózati kommunikáció

- Az állományokat beolvasó függvényeknek, mint például a `read.csv()`, `read.table()` és a `scan()` állománynevek helyett URL-eket is meg lehet adni
- A `download.file()` függvény állományok letöltésére szolgál
- Az `url.show()` pedig URL-ekkel adott állományok tartalmának megjelenítésére

Excel állományok feldolgozása (1)

- A legbiztonságosabb megoldás szöveges állományokba exportálni az Excel állományokat
- Windows rendszerekben egy lehetőség az RODBC csomag használata
  - Akár több munkalap is lehet az állományban

Excel állományok feldolgozása (2)

- Kizárólag Windows környezetben használható az `xlsReadWrite` csomag
  http://www.swissr.org/software/xlsreadwrite
  - Ez írni és olvasni is képes Excel állományokat
- A `gdata` csomag `read.xls()` függvénye Excel állományok beolvasására szolgál
  - A függvény az Excel állományt CSV állománnyá alakítja egy Perl programmal
  - Perl telepítése szükséges

Excel állományok feldolgozása (3)

- A `WriteXLS` csomag `WriteXLS()` függvénye adatok kiírását teszi lehetővé Excel 2003 állományokba
  - Perl szükséges a használathoz

Excel állományok feldolgozása (4)

- `XLConnect` csomag http://www.mirai-solutions.com/, `xlsx` csomag
  - Platformfüggetlen megoldások Excel állományok olvasásához és írásához
  - Használatukhoz Java szükséges
  - Az Apache POI könyvtáron alapulnak http://poi.apache.org/

Egyéb IO lehetőségek

- A `base` csomag `gzcon()`, `gzfile()`, `bzfile()` és `unz()` függvényei: adattömörítés (gzip, bzip2, ZIP)
- `rimage` csomag (digitális képfeldolgozás): JPEG állományok beolvasása
- `ReadImages` csomag: JPEG és PNG képek beolvasása
- `sound` csomag: WAV állományok írása és olvasása
- `XML` csomag: XML dokumentumok írása és beolvasása
- Csomagok speciális bináris formátumok kezeléséhez: `hdf5`, `RNetCDF` és `ncdf` csomagok

Hivatkozások

- The R Project for Statistical Computing http://www.r-project.org/
- Kurt Hornik (2010), *The R FAQ*. ISBN 3-900051-08-9 http://cran.r-project.org/doc/FAQ/R-FAQ.html
- Brian Ripley et. al (2010), *R Data Import/Export*. ISBN 3-900051-10-0 http://cran.r-project.org/doc/manuals/R-data.html
- Különböző csomagok dokumentációja http://cran.r-project.org/web/packages/