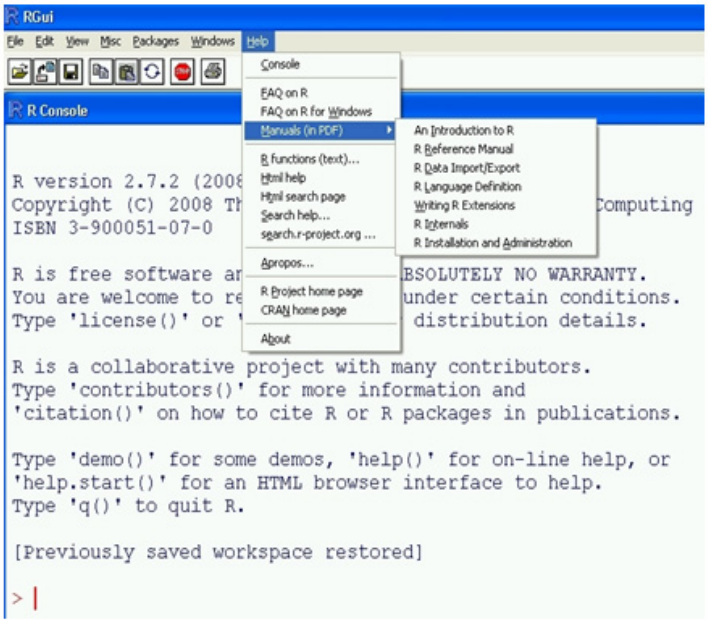
R – ЕЗИКА НА НАУКАТА

д-р Доньо Гънчев, Аграрен Университет – Пловдив

„R: A Language and Environment For Statistical Computing" е названието на един от най – популярните и разпространени в академичните среди по света езици за програмиране. Създаден през 1996 г. (самия проект стартира още през 1990 г.) от двама професори по статистика – Робърт Джентълмен и Рос Иака от Университета в Оукланд, Нова Зеландия като open-source алтернатива на програмния език S. В самото си начало той е озаглавен „GNU S" като по-късно е преименуван на R което произлиза от началните букви на малките имена на неговите създатели. S, като език за програмиране е творение на екипа на Джон Чембърс от лабораториите на Бел в Ню Джърси, САЩ и широко разпространен измежду статистиците и математиците. За разработването на S, през 1998 г. Чембърс е награден от ACM (Association for Computing Machinery) с Software System Award. Езика има множество имплементации като най-популярната е в основата на известните пакети за статистическа и математическа обработка на данни S-Plus и SAS. По своята същност R се явява като своеобразен диалект на S, доказателство за което е факта че голяма част от сорс кода за S-Plus и SAS без проблем може да се използва и под R. Колкото и да е странно, въпреки че R е своеобразен безплатен аналог на S и в този смисъл негов конкурент, Джон Чембърс и колегите му не само че не опълчват срещу проекта, а напротив – приветстват го и в момента повечето от тях, включително и Чембърс са членове на R Foundation – организацията занимаваща се с поддръжката и разпространението на езика. И все пак какво е R ? Както се разбира от неговото пълно име: програмен език и развойна среда за статистическа и математическа обработка на данни – ни повече, ни по-малко, мощен инструмент използван от хиляди учени, преподаватели и студенти по света, от най-различни дисциплини и направления. През последните години изучаването на езика се превъща в стандарт в катедрите по статистика и математика на повечето от университетите в Западна Европа и САЩ. Все повече и повече научни и бизнес институции където се провеждат статистически и математически изследвания и обработка на данни изискват от настоящите и бъдещи си служители умения за работа с R. Започнал своето начало като научен проект на двама професори по статистика с цел създаването на open-source алтернатива на комерсиалния S, днес R се радва на все по-нарастваща популярност. Сред поддържащите и използващите езика институции (донори и бенефициенти) са компании като: Google; Astra Zeneca – Швеция; Merck – САЩ; AT&T Labs, Ню Джърси – САЩ; Baxter Healthcare Corporation, Калифорния – САЩ; Ef-prime Incorporation – Япония; ViaLactia Biosciences - Нова Зеландия; Telecom New Zealand - Нова Зеландия; а от научните институции се открояват имената на Университета в Нантес – Франция; Станфорския Университет, Калифорния – САЩ; Факултета по Икономика на Университета в Грьонинген – Холандия; Департамента по Икономика на Университета в Милано – Италия; Центъра за борба с рака на гърдата към Колежа по Медицина в Хюстън, Тексас – САЩ; Института Дана – Фарбер за борба с рака в Бостън - САЩ; Департамента по Биостатистика към Университета Джон Хопкинс в Мериленд - САЩ; Департамента по Биостатистика към Медицинския Университет в Вандербилт – САЩ; Департамента по Икономика на Стохолмския Университет – Швеция; Департамента по Статистика на Университета Уисконсин – Мадисън – САЩ; Департамента по Статистика на Университета на щата Айова – САЩ; Центъра по Медицинска Биостатистика на Масачузетския Университет – САЩ; и много други. По думите на Дарил Прегибон, изследовател в Google и цитиран в статията във вестник New York Times посветена на R, "Езика за компанията е толкова важен че е трудно да си представим какво би било ако него го нямаше. R позволява на учените на провеждат сложни анализи без да е необходимо те да притежават задълбочени познания в в областта статистиката, математиката, както и на компютрите и компютрите системи". Освен всичко друго, съществуват и множество допълнителни проекти свързани с R като

например: Bioconductor – софтуер използващ R за анализ на данни в областта на генетиката. Този проект в момента се разработва от самия проф. Робърт Джентълмен. Друг подобен е софтуера Zoo/Phyto Image за автоматичен таксономичен анализ на проби от зоо и фито планктон на база на дигитални снимки.

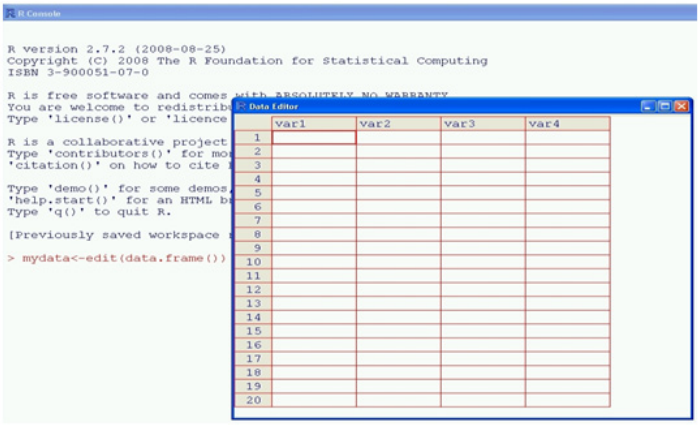


Общ вид на R IDE под Widows

Основни характеристики:

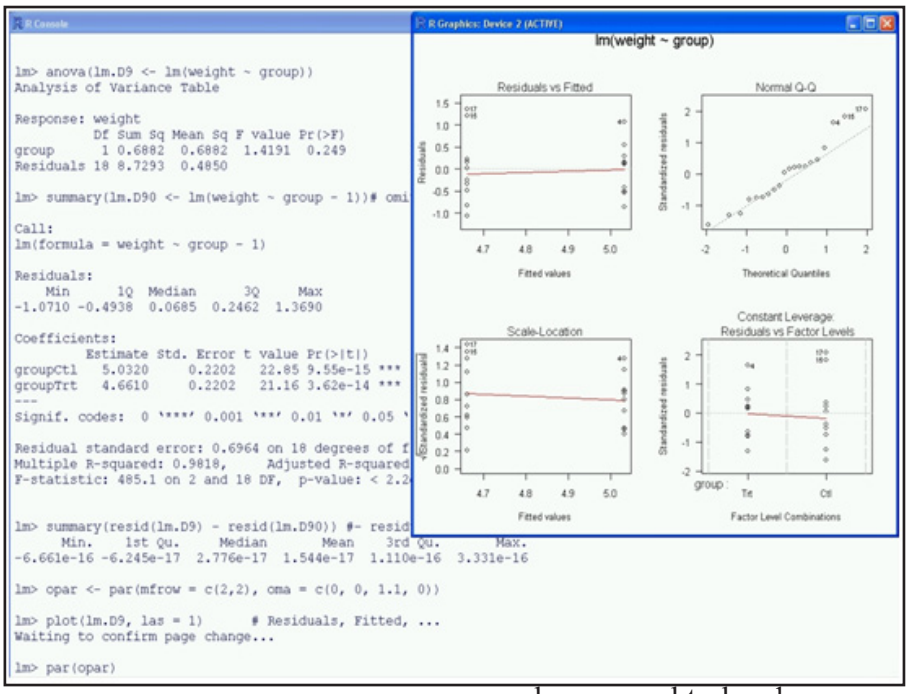
- Обектно-ориентиран, прост и лесен за научаване програмен език наподобяващ по синтаксис на C, но семантично подобен на Scheme, като притежава и прилики с Perl, Python и Lisp. Езика включва възможности за работа с различни типове данни - както числови, така и стрингови и логически (булеви). Обектни структури като: вектори, фактори, списъци, матрици и датафреймове (data frames). Наличие на условни преходи от рода на if-else, както и switch която при R е просто функция (switch (statement, list)) и функцията eval за оценка дали е изпълнено дадено условие или не. Цикли от типа на for, while и repeat и възможности за създаване на потребителски функции. Езика поддържа в пълна степен обектно-ориентираното програмиране аналогично на C++. Възможностите на R в това отношение по настоящем се използват широко при изграждането на математически симулации и модели на различни биологични и природни системи.
- Разнообразни механизми за въвеждане на първоначалните данни – R IDE притежава вграден редактор за данни, подобен на този на програмите за електронни таблици, който се извиква от интерпретатора чрез функцията edit ( ) - например mydata<-edit(data.frame()).

Вградения редактор на данни в действие



Друга възможност използвана най – често от начинаещите потребители е чрез функцията scan( ) - присвояването на функцията на дадена променлива (например mydata<-scan()) води до стартиране на цикъл за въвеждане на данни, които ще бъдат записани във вектора mydata. Натискането на Enter без въвеждане на стойност терминира цикъла и въвеждането приключва. Друг начин, също често използван, е чрез пакета RODBC благодарение който става възможност да се импортират xls файлове. Освен това с функцията read.table() може да се четат csv и txt файлове а когато като аргумент на същата функция, не се подаде конкретния csv файл а съдържанието на клипборда – data1<-read.table(file("clipboard")), то тогава в интерпретатора на R се импортира копираното съдържание на клетките от програмата за електронни таблици. Освен в официалния сайт на езика има препратки към външни модули чрез които може да се извикват обекти на R (функции) директно от програмите за електронни таблици (поддържа се както Ms. Excel, така и Open Office Calc и Gnumeric)

- Възможност за разширяване на езика чрез т. нар. packages, които са аналогични на модулите в Python например. Стандартната инсталация на R IDE включва 9 допълнителни пакета, а от сайта на фондацията (http://mirrors.dotsrc.org/cran/) разработваща езика са достъпни за сваляне над 1600 !! обхващащи почти всички области на науката и практиката. Съгласно принципите на отворения код, всеки може да създаде пакет за R (налице е подробно

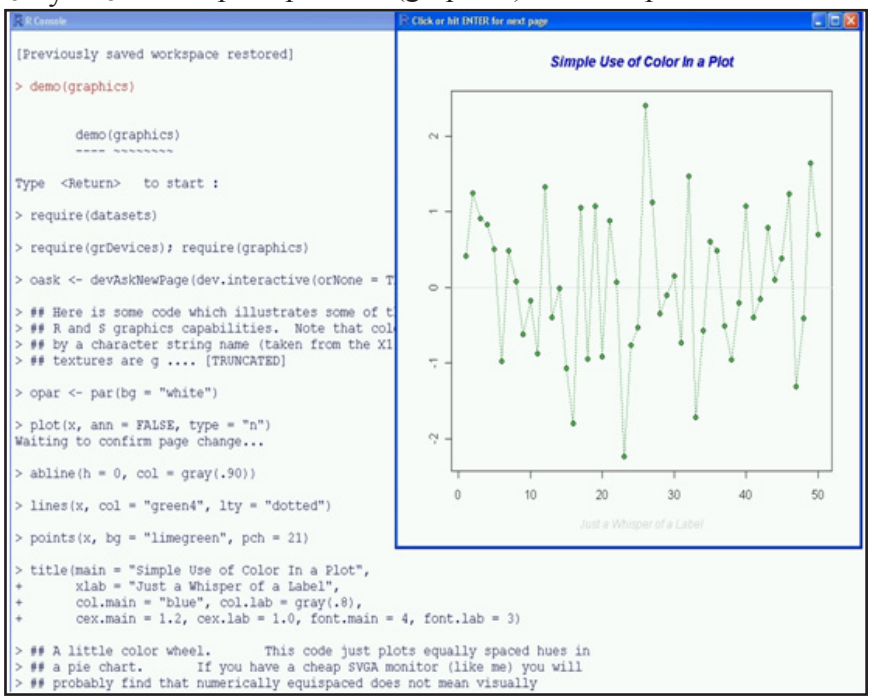


и безплатно ръководство за това) и да го публикува. Освен това всеки може да модифицира съществуващия код на в даден пакет съгласно своите нужди. Всеки пакет съдържа подробен help файл под формата на chm и pdf с детайлно описание на всички функции които съдържа и начина им на използване съпроводено със съответните примери. Така че казаното по-горе по отношение на потребителските функции, до голяма степен се обезсмисля в предвид на факта че е много по-вероятно да откриете нужните ви функции във вече създадени пакети, които ако все пак не ви удовлетворяват по начина по който работят, може безпроблемно да модифицирате. Пакетите могат да се свалят автоматично при наличие на връзка с Интернет. Под Windows и Mac OS това става като се кликне на Install Packages от менюто Packages. При всички операционни системи, включително и Unix – чрез install. packages (“”), като вътре в кавичките се пише названието на пакета който желаем да инсталираме. Ако нямате връзка с Интернет, спокойно може да си свалите нужните ви пакети които са в zip (не е нужно предварително да ги разхивирате – R IDE се грижи автоматично за това) и да ги инсталирате off line – просто избирате опцията „Install Package (s) from local zip file” и от менюто Open избирате модула zip файла (пакета) който желаете да добавите.

- Възможности за графично представяне на информацията – говорим за всемогъщата функция plot ( ) и нейните разновидности, която може да приема над 30 аргумента. Наистина, в това отношение способностите на R далеч надхвърлят тези на останалите инструменти за статистическа обработка на данни.
- Поддръжка – една дума – СТРАХОТНА. Езика се разработва от R Foundation с президенти Робърт Джентълмен и Рос Иака, със седалище във Виена. Самия език пристига с 7 ръководства в pdf формат, подробна подробна помощна документация в chm и html формат. Чрез функцията example ( ) директно от интерпретатора може да се получи информация за даден обект (функция) под формата на реален пример – example (lm) проведе едно факторен AVOVA анализ с графичното представяне включително.

Вградения пример за провеждането на ANOVA анализ чрез функцията example ()

Друг начин за получаване на подобна информация е чрез функцията demo( ) - например demo (graphics) ще ви представи по



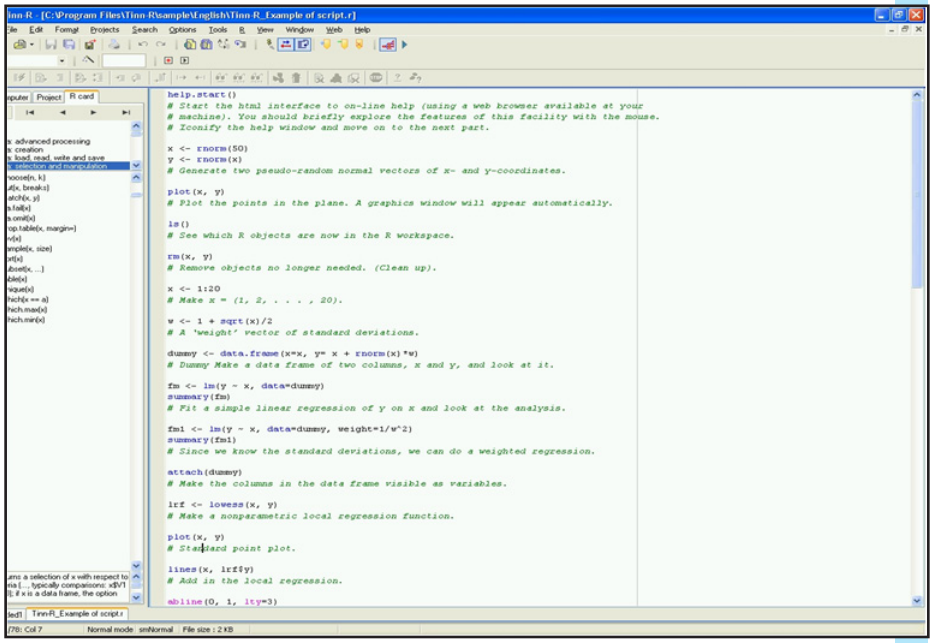
интерактивен начин възможностите на езика в областта на визуализацията на научни данни.

Вградената демонстрация на графичните възможности на R чрез demo ()

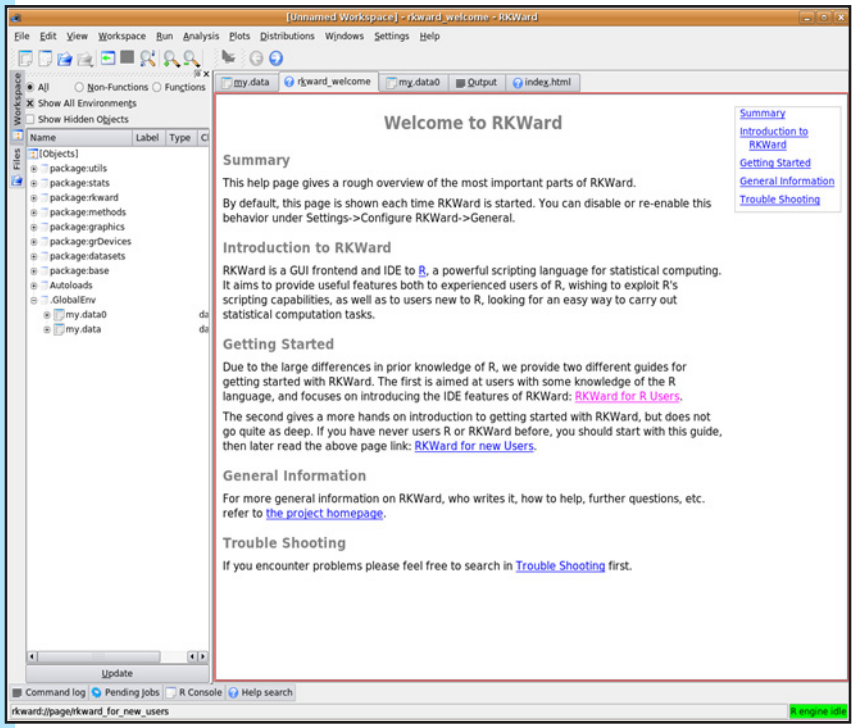
•

От сайта на фондацията http://www.r-project.org/ са достъпни за безплатно сваляне множество допълнителни ръководства и безплатни книги посветени на програмирането с R. До момента има издадени над 50 учебници и ръководства както в областта на статистиката и математиката като цяло, така и посветени на спецификата при различните науки като физика, биология, екология, медицина, сеизмология, геология и др. Примерни глави от доста голяма част същите са достъпни чрез Google Books, като разбира се не е никакъв проблем да си ги и закупите от Amazon.com. Множество уеб страници и блогове също са посветени на R, обхващащи различни аспекти и акцентиращи върху спецификата на използването му в отделните научни дисциплини и направления. Налице е и научнен журнал – R Journal (http://journal.r-project.org/) издаван от R Foundation който е безплатен – т.е статиите са достъпни без ограничения в pdf формат. Освен това всеки желаещ може да изпрати своя статия за публикуване създадена чрез LaTex, като списанието е със значителен Импакт фактор и се реферира от международно признати авторитетни учени в областта на статистиката и математиката. Като черешка на тортата е и провеждащата ежегодно конференция за R – UseR! (http://www2.agrocampus-ouest.fr/math/useR-2009/), с участието на представители от видни научни и бизнес институции като Google, Springer, CRC Press, INRA, Сименс, Университета в Принстон, Института „Макс Планк“, Университетите в Харвард, Оксфорт и Иелс, Европейския център за превенция контрол над болестите, Центъра за анализиране и моделиране на данни към Университета в гр. Фрайбург – Германия, Технологичния и Медицинския Университети във Виена, Университета и Университетската болница в гр. Копенхаген – Дания, Кралски Колеж Лондон, Отдела по биостатистика към федералния център за изследване на рака – Германия, Института „Пастьор“ - Франция, Колумбийския Университет в САЩ, дори Майкрософт както и още много други. По-голяма част от материалите от конференциите като презентации, постери и сборници със статии са също достъпни за безплатно сваляне от Интернет. От търсачката на Google за сорс код, в опцията за разширено търсене също можете да специфицирате разширението „r” .

• Наличие на допълнителни редактори и разширения на езика осигуряващи графичен интерфейс. Може би основния и най-дразнещ недостатък на R IDE е че вградения редактор не позволява оцветяване на сорс кода, което е е една от основните критики към разработчиците. Въпреки това допълнително може да се инсталира софтуер от рода на Tinn-R (http://www.sciviews.org/Tinn-R/), Rkward или SciViews http://www.sciviews.org/SciViews-R/index.html – SciViews-R). Ако използвате Ubuntu – просто щракнете върху отметката на Rkward в Add Remove Programs и операционната система ще инсталира както него, така и R language автоматично. Освен това Gedit – основния текстов редактор на Gnome поддържа безпроблемно синтаксиса на R. Под KDE може да използвате редактора Kate. Потребителите на Windows могат да се възползват от безплатния Crimson Editor. Под Линукс, за разлика от Windows, езика се интегрира със системата и стартирането на R IDE става като се напише R в командния интерпретатор,



Популярния редактор на R сорс код Tinn-R



Интегрираната среда с визуален интерфейс за R - Rkward под Ubuntu

- Интеграция с други езици като C/C++, Fortran, Python, като специално за последния може да се изтегли допълнителен модул за Python и пакет за R чрез който двата програмни езика, респ. техните интерпретатори да се обединят. Друга интересна и малко позната възможност е поддръжката на Tcl/Tk. Благодарение на това става възможно чрез Tcl и Tk да се изграждат GUI интерфейси т.е интерпретатора на R става аналог на wish. Графични интерфейси могат да бъдат изграждани и чрез wxPython чрез инсталирането на едноименния пакет и пакета RSPython.
- Мултиплатформеност – първоначално R бил създаден само за Unix операционни системи, като в последствие, след нарастване на популярността му се появили и версии за Windows и MacOS.

Когато за първи път се представя R Program Language пред дадена аудитория, един от основните въпроси които неминуемо възниква е, защо да се използва при условие че има достатъчно платени и безплатни програми от типа на електронните таблици с визуален интерфейс които могат да се използват за математическа и статистическа обработка на данни. Този въпрос и спора породен от него доста прилича на спора между LaTex и визуалните текстови редактори. В основни линии предимствата на R в това отношение могат да бъдат накратко обобщени така:

1. Конзолния вход и изход използван от R осигурява пълен контрол на действията от страна на потребителя (програмиста).
2. R притежава много по-големи възможности за математическа и статистическа обработка колкото програмите с визуален интерфейс, които възможности са неимоверно разширени чрез допълнителните пакети с функции. Така например чрез пакета deSolve може да решавате диференциални уравнения, пакети като drc и drfit се използват от токсиколози от целя свят за изчисляване на показателите LD50 и LD90, провеждане на пробит и логит анализ, както и за построяване на т.нар dose – response curves. По настоящем езика се използва масово от математиците работещи в областта на Теорията на Хаоса и Недетерминираните Линейни Системи - налице са два специализирани пакета за това – fNonlinear и tseriesChaos. Така могат лесно и бързо да бъдат създавани фрактални графики, атрактори, бифуркационни диаграми и фазови модели на различни системи. Например, едно от най – новаторските направления в областта на кардиологията е моделирането на сърдечната дейност, чрез изграждането на съответните фазови модели с използването на R Program Language. Най - новите постижения в лечението на епилепсията също се дължат на обработката на получените енцефалограми с R.
3. Изходната информация от проведените анализи, която извежда R е далеч по-детайлна отколкото резултатите извеждани от повечето от специализираните статистически програми с графичен интерфейс.
4. R притежава по-мощни методи за четене (въвеждане) на големи съвкупности от данни особено характерно за сложните научни изследвания и тестове. Поддържат се множество разнообразни методи за въвеждане на първоначалната информация, включително и благодарение на пакета RODBC директно имтортиране на xls файлове, както и интегриране на R IDE с по-

известните програми за електронни таблици. Така например, разработчиците на една от най – популярните програми с графичен интерфейс за статистическа обработка на данни „STATISTICA" в последните версии, рекламират именно интеграцията на софтуера с R.

5. Както беше споменато и по – горе за провеждането на сложни математически и статистически тестове чрез R не е необходимо детайлното и пълно познаване на съответните алгоритми и методи, което прави езика много подходящ за учени представители на не математическите науки, като химици, физици, биолози, медици и др.

6. R поддържа в пълна степен обектно-ориентираното програмиране, което се използва много при създаването на сложни математически модели и симулации.

7. R притежава далеч по – добра поддръжка като помощна информация от която и да е програма с визуален интерфейс.

8. Обработката на данни в R не е усложнена от объркващ визуален интерфейс. Всеки които е работил с големите GUI софтуерни пакети за статистическа обработка на данни ще ме разбере. В повечето случей потребителя се изправя пред цял лабиринт от менюта, подменюта и подподменюта със множество опции и настройки, за чието разбиране е нужно значително повече време и усилия от колкото за разучаването на R.

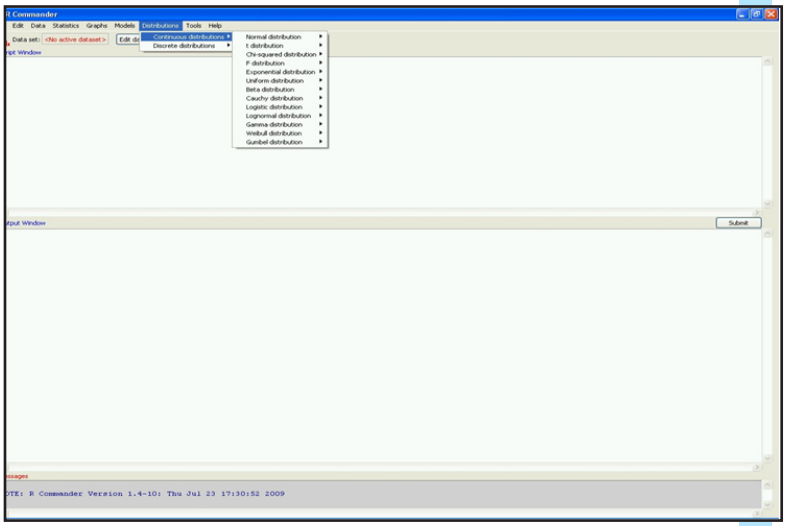
9. Възможност за пълно автоматизиране на математическата (статистическата) обработка. Чрез R всеки може да създаде свой собствен скрит описващ обработката на конкретния, специфичен тип експерименти. Интерпретатора автоматично ще прочете първоначалните данни намиращи се в зададената работна директория или въведени в програмата за електронни таблици, ще проведе съответните анализи и като финал ще запише създадените графики (поддържат се по-известните графични формати, както и pdf) в посочена от потребителя директория. След което готовите вече графики без проблем могат да се вмъкнат в научната статия, монография, дисертация, доклад и пр.

10. Наличие на огромен брой допълнителни пакети с функции и алгоритми – в това отношение R няма аналог не само между програмите с графичен интерфейс, но и сред останалите програмни езици. Както беше уточнено, до момента пакетите са над 1600 !!! създадени от програмисти от целия свят и обхващащи почти всички направления на науката и практиката. Всеки един пакет съдържа средно около 20 – 40 отделни функции (някой, и много повече), така че преценете сами за какво количество става въпрос. Още нещо – тъй като това е софтуер на базата на отворения код, не е никакъв проблем да модифицирате съществуващите функции и алгоритми съгласно вашите нужди и изисквания.

11. R поддържа тясна интеграция с LaTex което е важно за учените с цел професионалното оформяне на техните статии за научните журнали, като се има в предвид факта че почти всички научни списания с висок импакт фактор, препоръчват (а някой дори задължават) статиите да бъдат написани на LaTex. Все пак не бива да се забравя факта че R е програмен продукт създаден от учени за учени.

12. R може да се интегрира с други програмни езици като C/C++, Fortran, Python, Tcl. Чрез ess (Emacs Speaks Statistics) модула той може да се интегрира и с редактора Emacs (http://ess.r-project.org/).

13. Благодарение на поддръжката на Tcl/Tk и wxPython (чрез допълнителен модул за последния) могат да се създават и графични потребителски интерфейси. Така например



споменатия по-долу Rcomander е създаден именно чрез Tcl/Tk.

R Commander

От сайта на фондацията може да изтеглите zip файл с примерни сорс кодове на Tcl/Tk.

14. Наличие на допълнителни редактори като Tinn-R, Rkward, Rcommander, Sciview_R

осигуряващи GUI интерфейс за въвеждането на данните и за провеждането на по – известните статистически обработки. Освен подобни „GUI редактори с общо предназначение“, съществуват и множество подобни, специализирани в областта на конкретната наука / научна област, предоставящи визуален интерфейс спрямо най – често използваните в дадената сфера анализи. Въпреки това, доста от потребителите на R предпочитат да не използват подобен софтуер. Според тях стандартните механизми за вход на данни и конзолния интерпретатор са напълно достатъчни, като е истина че почти всички са съгласни че вградения в R IDE редактор е необходимо да притежава оцветяване на изходния код . Една не малка част, но главно по-опитните предпочитат да използват Emacs като редактор на скриптовете. Популярен редактор на сорс код който поддържа R е и Crimson Editor (http://www.crimsoneditor.com), под Линукс, както беше споменато може да използвате Kate. Стандартният текстов редактор на Gnome – Gedit също осигурява синтаксисно оцветяване на езика. Други редактори притежаващи синтаксисно оцветяване за скриптове на R са: WinEdit, Textpad, Syn, Nedit, Jedit, UltraEdit, всеизвестния Vim, Bluefish; Eclipse като цялостна развойна среда написана на Java също поддържа R чрез съответния плъгин.

15. R е мултиплатформен, съществуват версии за Windows, Linix (има го в repositories на повечето от известните дистрибуции, или може да се вземе като сорс код и да се компилира чрез gcc), MacOS. Езика може да се свали от официалния сайт на R foundation http://www.r-project.org/. Под Ubuntu може да се инсталира чрез Synaptic или от Add Remove Programs като изберете отметката на Rkward. Скриптовете на R с разширение „.r” тръгват без проблемно под всяка една операционна система.

16. И не на последно място – R е абсолютно безплатен на фона на скъпо платените професионални пакети за математическа и статистическа обработка на данни като SAS, SPSS, S-Plus, Systat, Stigma и др. Допълнително предимство е че по отношение на хардуера изискванията на R IDE са далеч по – скромни от тези на останалите програми (пълната инсталация на последната версия заема около 70 MB дисково пространство, като размера на един допълнителен пакет варира средно между 500 KB и 2 MB), което също не е без значение като се има в предвид ограничените бюджети на научните организации и факта че парите винаги отиват за закупуване на скъпо струващата лабораторна техника и консумативи а за хардуер и софтуер „пари нема“.

Включи се в групата на БГ Наука в най-голямата социална мрежа. Следи новостите и знай какво се случва....