Estadística aplicada per a professionals de la salut

Introducció a "R"

Albert Roso. Estadístic. Tècnic de Recerca de l'IDIAP Jordi Gol.

Contingut

1. Funcionament d'R

1.1 Què és R?

R, és un llenguatge de programació i un entorn de desenvolupament de software per a l'obtenció de càlculs i gràfics estadístics. Fou creat originalment per Ross Ihaka i Robert Gentleman a la Universitat d'Auckland, Nova Zelanda, i actualment està desenvolupat per l'*Equip Central de Desenvolupament de R*.

R és àmpliament emprat per a desenvolupar programes estadístics i per anàlisi de dades, i ha esdevingut l'estàndard en el que els estadístics desenvolupen nou software. Els avantatges d'R en front d'altres programes habituals d'anàlisi de dades, com poden ser SPSS, SAS, Stata, són múltiples:

- És software lliure i per tant gratuït. Pot ser descarregat des del web http://www.r-project.org/.

- És multiplataforma. Existeixen versions per a Windows, MacOS, Linux i altres plataformes.

- Està avalat i en constant desenvolupament per una àmplia comunitat científica que l'utilitza com el programa estàndard per a l'anàlisi de dades.

- Compta amb multitud de paquets per a tot tipus d'anàlisi estadística i representacions gràfiques. Des de les més habituals, fins a les més recents i sofisticades que no inclouen altres programes. Els paquets estan organitzats i documentats en un repositori CRAN (Comprehensive R Archive Network) al qual es pot accedir des del web http://cran.r-project.org/.

- És programable, el que permet a l'usuari poder crear fàcilment les seves pròpies funcions o paquets per analitzar dades específiques.

- Existeixen multitud de llibres, manuals i tutorials lliures que permeten el seu aprenentatge i il·lustren l'anàlisi estadística de dades en diferents disciplines científiques com les matemàtiques, la física, la biologia, la medicina, etc.

Per defecte, l'entorn de treball d'R és en línia de comandes, el que significa que els càlculs i les anàlisis es realitzen mitjançant comandes o instruccions que l'usuari tecleja en una finestra de text. No obstant, existeixen diferents interfícies gràfiques d'usuari que faciliten els seu ús, sobre tot per a usuaris novells.

La interfície gràfica que s'utilitzarà para realitzar les pràctiques d'aquest curs serà R Commander, desenvolupada per John Fox, http://www.rcommander.com/.

1.2 Descàrrega i instal·lació

Per a instal·lar R, hi ha que seguir les passes següents:

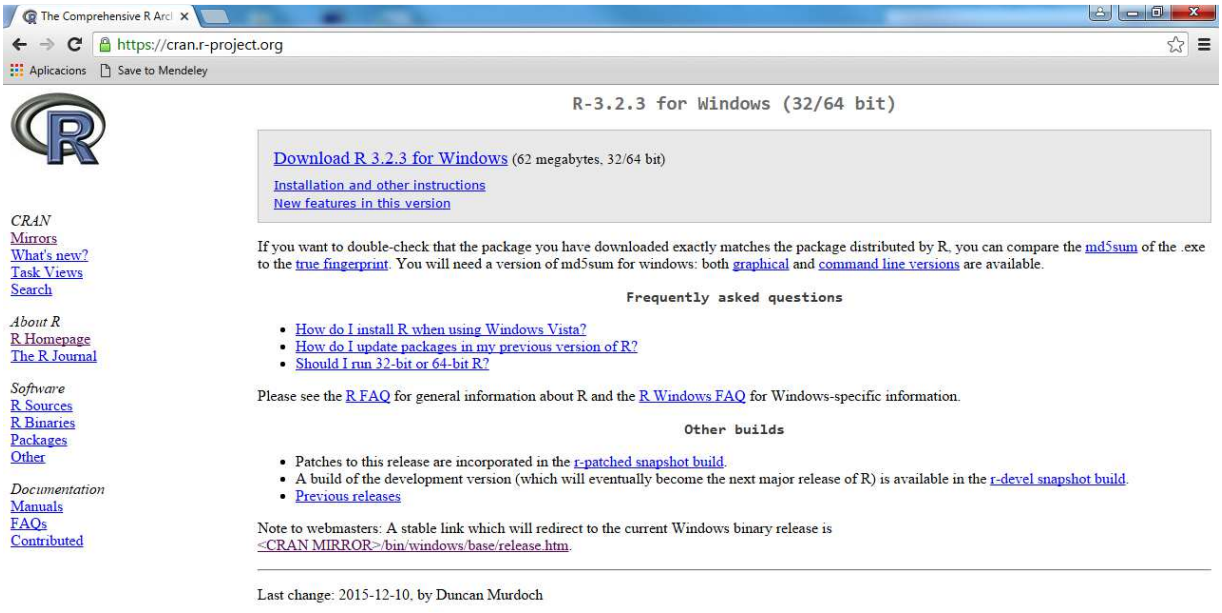
1. Obrir la pàgina web d'R: http://www.r-project.org/.

2. Fer clic en 'CRAN' i, a continuació escollir un dels servidors (mirrors) de CRAN (Comprehensive R Archive Network).

3. Segons el sistema operatiu en ús, fer clic en Linux, MacOS X o Windows i seguir les instruccions corresponents.

4. Si s'usa Windows, fer clic en '*install R for the first time*'.

5. Apareixerà la següent pantalla, on haurem de fer clic sobre *'download R 3.2.3 for Windows'* per descarregar el fitxer d'instal·lació d'R. La figura següent correspon a la versió oficial d'R a gener de 2016.



6. Un cop descarregat el fitxer, l'executem i seguim les instruccions d'instal·lació.

7. D'aquesta forma s'haurà instal·lat el programa d'R a l'ordinador. La instal·lació haurà creat un arxiu directe a l'Escriptori que podrem executar per tal d'accedir a R.

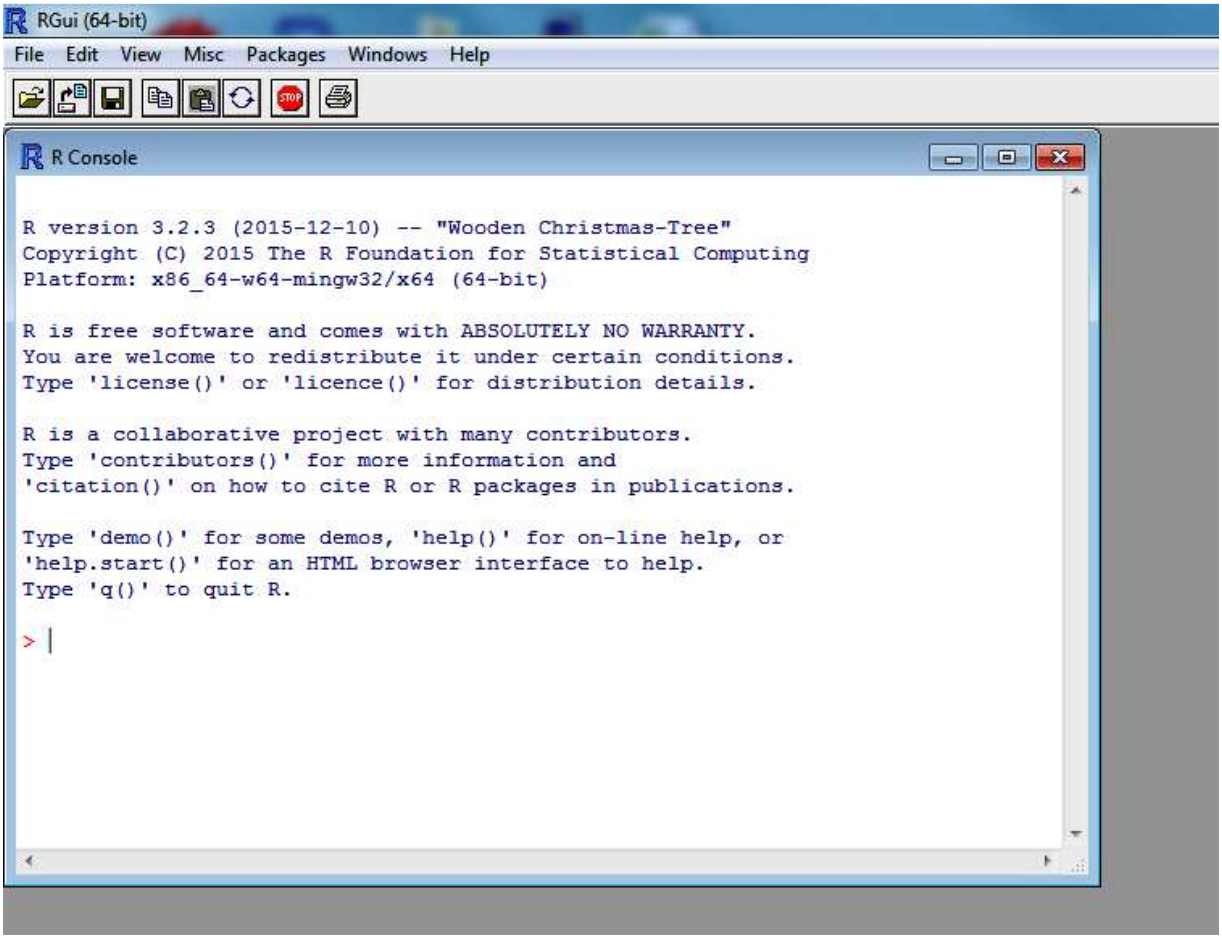
1.3 Paquets o llibreries contribuïdes

Juntament amb la versió estàndard és possible instal·lar paquets o llibreries contribuïdes, que contenen noves funcions i eines estadístiques i gràfiques.

No obstant, molts d'aquests paquets ha estan fets per a realitzar anàlisis molt específiques, i és recomanable instal·lar posteriorment només aquells paquets que l'usuari realment necessiti.

La manera més fàcil per a obtenir-los, és tenint el programa obert i anar a la barra d'eines d'R:

1. Farem clic a la pestanya *Packages→Install Package(s)...*



2. Se'ns desplegarà un llistat de CRAN *mirrors* i seleccionarem la ubicació que ens interessi.

3. Un cop desplegada la llista de *packages*, seleccionarem el que vulguem instal·lar.

4. Seguim les instruccions i esperem que acabi d'instal·lar tots els paquets necessaris.

5. Un cop instal·lats ja podrem utilitzar-los, accedint a *Packages→Load Packages...* i seleccionant el paquet descarregat.

1.4 Característiques d'R

1. En el mode per defecte, obrint R, s'obre una sola finestra, la consola o finestra de comandaments de R, en la qual es poden entrar les ordres i on es veuran els resultats de les anàlisis. L'indicador o *prompt* del sistema és el signe >. Cada instrucció, en principi, acaba amb un signe de punt i coma (;) i *Enter* que indica la seva execució. Si no s'utilitza el punt i coma i intentem executar l'ordre amb *Enter*, l'intèrpret de comandaments provarà de traduir la instrucció i, si és correcta, l'executarà; si no és correcta mostrarà un missatge d'error, `Error`, i si és incompleta quedarà a l'espera de completar l'ordre en la línia següent mostrant com a indicador un signe +. A la pràctica, s'utilitza l'*Enter* per acabar una instrucció o per a dividir la línia, si la instrucció és llarga.

El signe # indica la introducció d'un comentari. Per exemple, la següent instrucció sumarà 3 i 4 i ignorarà el comentari:

```
3 + 4 #és un exemple
```

La tecla *Esc* permet reiniciar l'actual línia en edició. Per tornar a executar instruccions utilitzades en la mateixa sessió, es pot utilitzar la tecla de moviment del cursor ↑.

2. R és un llenguatge a través d'objectes i funcions. Les instruccions bàsiques són expressions o assignacions. Per a realitzar una assignació es poden utilitzar els signes `<-`, `->` i `=`.

```
n<-5*2+sqrt(144)
m=4^-0.5
n+m->p
```

Per a visualitzar el contingut d'un objecte només és necessari escriure el seu nom. Si l'objecte és una funció es mostrarà en pantalla el programa que la funció executa.

```
n
m
p
(x <- log(7))
x
log
```

3. Les comandes `objects()` y `ls()` ens permeten visualitzar el llistat d'objectes d'R presents en l'actual espai de treball (workspace) del directori de treball.

4. Per a les ubicacions del disc dur s'ha d'utilitzar la direcció entre cometes i barra (*/*) entre subcarpetes. Per obtenir el directori de treball podem utilitzar la comanda `getwd()`. Per exemple:

```
getwd()
```

```
"C:/Mis Documentos"
```

5. R disposa d'una ajuda molt completa sobre totes les funcions, procediments i elements que configuren el llenguatge. També disposa de manuals que es poden accedir via la barra d'eines d'R:

Help→Manuals (in PDF)→ ...

A més a més de les opcions de menú pròpies d'R, des de la finestra de comandes es pot accedir a la informació específica sobre les funcions amb la comanda `help` o mitjançant '?':

```
help(objects)
help(log)
?ls
```

6. També és possible obtenir ajuda sobre diferents temes mitjançant la funció `help.search`. R buscarà ajuda sobre el tema escollit en totes les llibreries instal·lades. Per exemple, per a obtenir informació sobre regressió logística:

```
help.search("logistic regression")
```

7. La majoria de les llibreries disponibles en la versió local d'R han de ser carregades abans de que es puguin utilitzar. Per exemple, per a carregar la llibreria `survival` es pot executar la següent instrucció

```
library(survival)
```

o escollir `survival` des de la barra d'eines:

Packages→Load Packages…

8. Durant una sessió d'R es pot guardar l'històric de totes les comandes executades fins al moment des de la barra d'eines:

File→Save History...

El fitxer guardat és un fitxer en format ASCII que pot ser editat per altres softwares si interessa. A més a més, és possible carregar l'històric en un altra sessió mitjançant (en la barra d'eines):

File→Load History...

D'aquesta manera es pot tornar a executar les comandes de la sessió anterior.

9. És possible obrir varies sessions d'R i treballar simultàniament en elles.

10. Per a sortir d'R es pot executar l'ordre `q()` o clicar el botó de tancar del programa. En aquest moment, R preguntarà a l'usuari si vol guardar l'actual espai de treball. Si hem guardat l'espai de treball anteriorment, no cal tornar-ho a fer. No obstant, si contestem que 'Sí', es

guardarà l’actual sessió en el fitxer amb extensió .RData a la carpeta de treball actual conjuntament amb l’històric de la sessió.

2. Maneres de treballar amb R

2.1 R bàsic

1. Hem vist que, treballant amb la consola d'R, és possible navegar entre les comandes executades anteriorment mitjançant les tecles ↑ i també ↓. No obstant, si l'usuari vol (tornar a) executar una sèrie de comandes, és més pràctic i eficient executar-les des d'una finestra *script* que es pot obrir des de la barra d'eines mitjançant:

File→New script

A les finestres *script* es poden entrar diferents comandes, separades o per ';' o per línies, que es poden executar conjuntament anant a la barra d'eines:

Edit→Run all

Si es vol executar només una selecció de les comandes de la finestra script, haurem de marcar-los i executar-los mitjançant 'Control-R'. Les comandes no s'esborraran i els resultats apareixeran a la finestra de comandes. Els scripts es poden guardar i utilitzar en qualsevol moment. El postfix per defecte és '.R'.

2. Mitjançant la funció `source()` es pot carregar un *script* d'R sencer, per exemple:

```
source("C:/Archivos de programa/R/script.R")
```

Podem aconseguir el mateix des de la barra d'eines:

File→Source R code...

3. Per defecte, tots els resultats apareixen en la consola d'R. Existeix, no obstant, la possibilitat d'enviar els resultats directament a un fitxer extern (de format ASCII) utilitzant la funció `sink()`. Veiem un exemple:

```
sink("C:/Archivos de programa/R/prova.txt")
n<-5*2+sqrt(144)
n
sink()
n
```

A partir de la instrucció `sink()`, els resultats apareixen de nou en la consola d'R.

4. Per defecte, el directori de treball d'R és 'C:/Mis documentos/', depenent de l'ordinador. Aquest es pot canviar en el quadre que s'obre mitjançant (barra d'eines)

File→Change dir...

5. En qualsevol moment d'una sessió d'R es pot guardar el seu contingut. Això és molt recomanable si volem tornar a utilitzar els objectes d'R en ús. La funció per a guardar l'àrea de

treball és `save.image()`. Una altra forma, és usar el quadre de diàleg corresponent accessible des de la barra d'eines:

File→Save Workspace...

Si volem guardar només alguns dels elements de l'àrea de treball, per exemple els objectes `x` i `y`, tenim dos possibilitats: o eliminar primer la resta d'objectes amb la funció `rm()` i després usar la funció `save.image()`, o usar la funció `save`:

```
save(x,y,file="nomarxiu.RData")
```

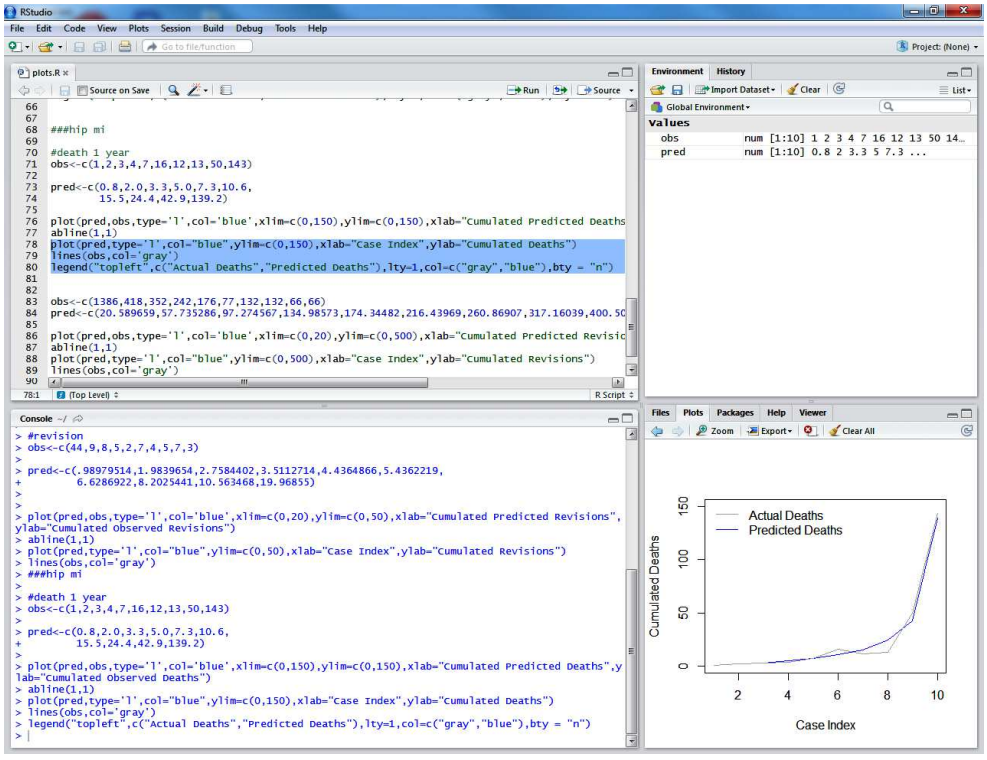
Com hem dit abans, encara que s'hagi guardat l'àrea de treball d'R, al sortir d'R el programa sempre pregunta si es vol guardar l'àrea de treball. Evidentment, al ja haver-ho fet, no cal tornar-ho a fer. Podem obrir un espai de treball amb la funció `load()`o anant a la barra d'eines:

File→Load Workspace...

Noteu que durant una sessió es poden carregar diferents àrees de treball.

6. Existeixen diferents programes editors que poden facilitar el treball amb R. Un d'aquests és l'RStudio, disponible en Windows, MacOS i Linux. El programa pot ser descarregat des del web http://www.rstudio.com/ .

A RStudio es poden obrir i editar diferents scripts, i ràpidament enviar-los a execució a R. Els avantatges d'aquest editor són que ofereix una sèrie de opcions no existents a R. Per exemple, es pot comprovar ràpidament si existeixen parèntesis sense tancar, millor gestió dels objectes i dels gràfics, creació de documents en pdf o word, etc.



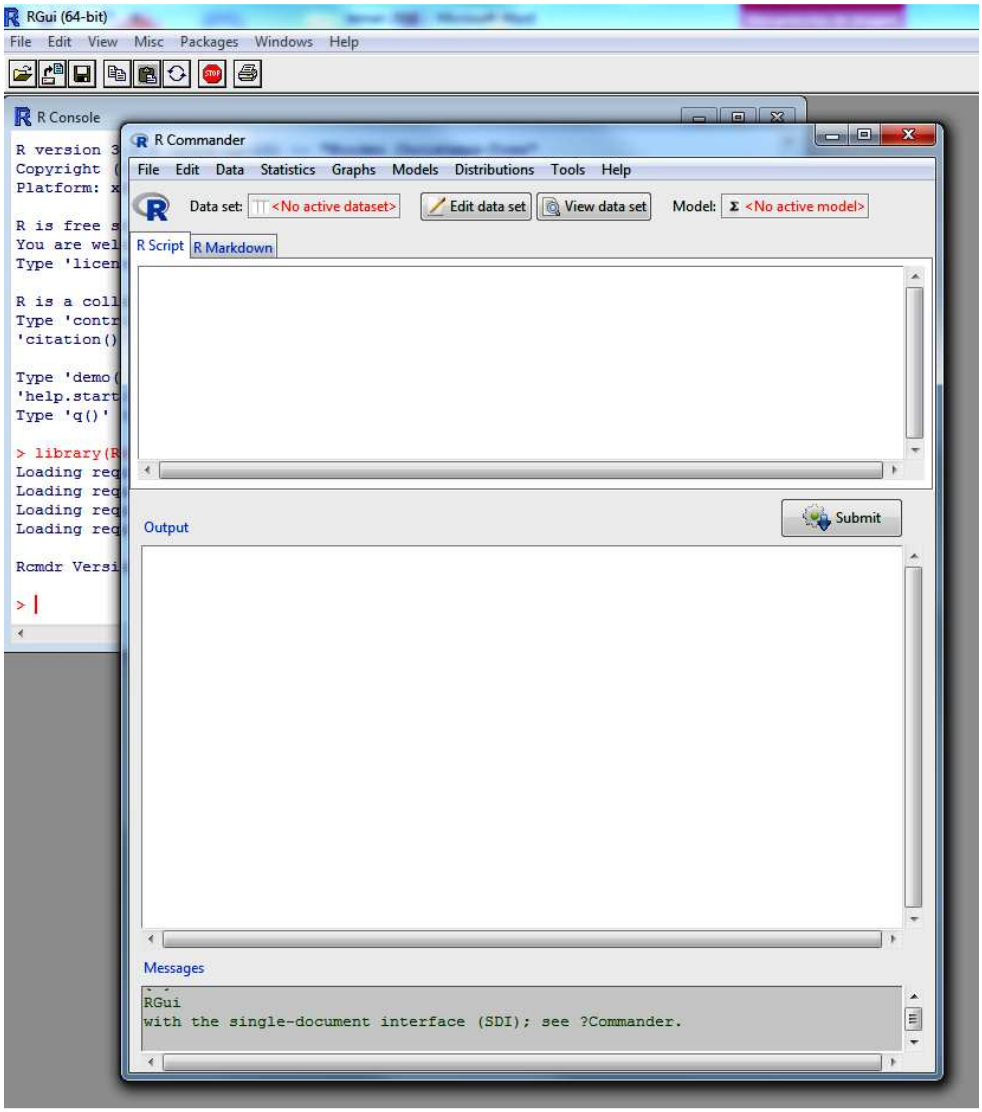
2.2 R Commander

R Commander és una interfície gràfica tipus finestra que cobreix una gran part de les anàlisis estadístiques més habituals en uns menús desplegables similars als de la majoria de programes que utilitzem en qualsevol sistema operatiu.

Podem dir que és una manera d'utilitzar R sense necessitat d'aprendre el seu codi o quasi res d'ell, fet que el fa bastant pràctic quan s'està aprenent a usar-lo.

Per a activar-lo, primer cal instal·lar el paquet `Rcmdr` i després sortir d'R guardant l'àrea de treball. Seguidament, reiniciem R i carreguem R Commander mitjançant `library(Rcmdr)` des de la barra d'eines:

Packages→Load Packages→ Rcmdr



Una vegada carregat R Commander veurem una finestra com la de la Figura anterior. En ella podem distingir 4 parts:

1. El menú de finestres desplegables, amb les opcions *File, Edit, Data, …*

És un menú de finestres amb entrades bastant intuïtives, que no requereixen coneixements d'R, però sí d'estadística.

2. La finestra d'instruccions.

Cada vegada que executem alguna acció del menú, R Commander traduirà aquesta acció a codi d'R i ho escriurà en aquesta finestra. Això permet anar aprenent aquest codi i, a més a més, facilita la possibilitat de tornar a executar la mateixa acció o una lleugera variant de la mateixa retocant el codi, sense haver de tornar a utilitzar el menú.

Per altra banda, aquesta finestra d'instruccions és equivalent a l'editor de R. Per exemple, podem escriure 2+2, clicar el botó d'executar (equivalent a 'Control-R') obtenint el resultat.

3. La finestra de resultats.

Si hem realitat aquest senzill exemple en la finestra d'instruccions, haurem vist que el resultat apareix en aquesta finestra. En general, qualsevol resultat de R Commander serà mostrat aquí.

4. La finestra de missatges.

Es la més inferior de totes i apareix lleugerament ombrejada. Serveix per a que R Commander ens informi de qualsevol aspecte, especialment d'errors comesos.

Com ja s'ha comentat anteriorment, utilitzarem R Commander per a realitzar les pràctiques d'aquest curs.

3. Manipulació de dades amb R

3.1 Tipus de dades en R

En R hi ha diferents tipus de dades. Els més bàsics són:

Numeric: És qualsevol nombre decimal. S'utilitza el punt com a separador de decimals. Per defecte, qualsevol nombre que es teclegi prendrà aquest tipus.

Integer: És qualsevol nombre enter. Per convertir un nombre de tipus *Numeric* en un enter s'utilitza la comanda `as.integer()`.

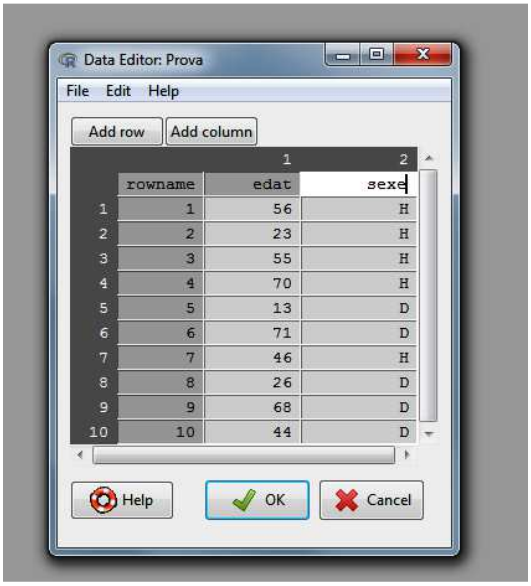
Logical: Pot prendre qualsevol dels dos valors lògics TRUE (cert) o FALSE (fals).

Character: És qualsevol cadena de caràcters alfanumèrics. S'han d'introduir entre cometes. Per convertir qualsevol nombre en una cadena de caràcters s'utilitza la comanda `as.character()`.

3.2 Creació i edició de dades

Per introduir noves dades en R Commander triem *New data set...* del menú *Data*. Això obre l'editor de dades que, en primer lloc, ens demanarà un nom per a la matriu de dades triat, en aquest cas l'anomenarem *Prova*, i a continuació obrirà una finestra amb caselles semblant a un full de càlcul d'Excel. En aquest full hem d'introduir les dades amb la mateixa estructura que té una matriu de dades, amb els individus en les files i les variables en columnes. Per afegir individus clicarem sobre *Add row* i per afegir variables sobre *Add colom.* En el cas de que vulguem eliminar certes files o columnes, tenim l'opció *Delete current row* o *Delete current column* a la pestanya *Edit* de la barra d'eines de la finestra oberta. Per exemple, introduirem dues variables amb valors possible d'edat i sexe.

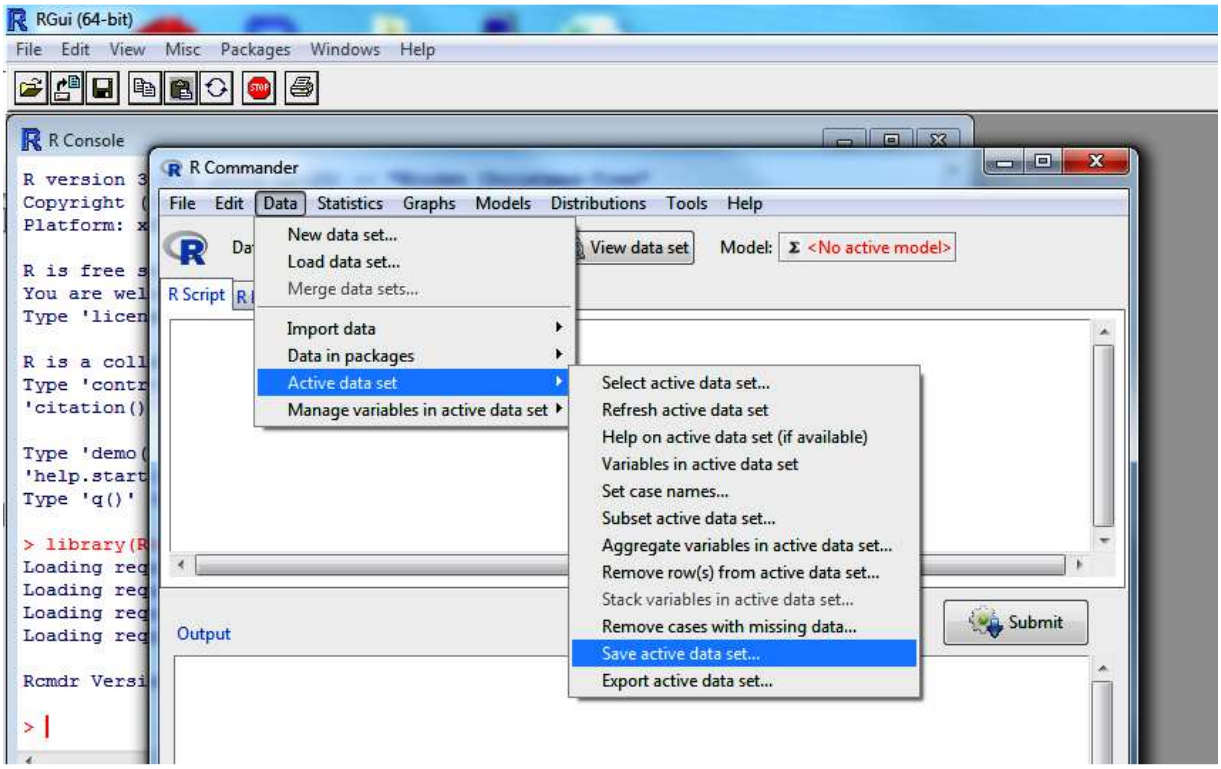
Un cop introduïdes les dades, hem d'anomenar les variables, és a dir, les columnes, amb noms senzills que ens recordin a quina variable correspon cada columna. Per a això cliquem amb el ratolí sobre la part superior de cada columna, on R Commander anomena per defecte les variables com var1, var2, etc. i escrivim altres noms més d'acord amb les nostres dades de *Prova*. En aquest cas hem nomenat les variables com *edat* i *sexe*.



Per acabar, tanquem la finestra de l'editor de dades. En aquest moment, a R hi haurà emmagatzemat les dades introduïdes convertint-les en el que R Commander diu el conjunt de dades. Observeu que just a sobre de la finestra d'instruccions apareix ara una pestanya informativa que posa Data set: Prova. Aquesta finestra especifica que, en efecte, el conjunt de dades actiu en aquest moment és el que nosaltres hem anomenat *Prova*.

Finalment, podem retocar aquestes dades prement la pestanya *Edit data set* que hi ha just sobre la finestra d'instruccions o simplement visualitzar les dades fent clic a la pestanya *View data set.*

Per desar un full de dades en R Commander, seleccionem al menú *Data* l'opció *Active data set* i, dins d'aquesta, *Save active data set...* A continuació ens demanarà un nom i un directori on emmagatzemar el fitxer, l'extensió per defecte serà .rda.

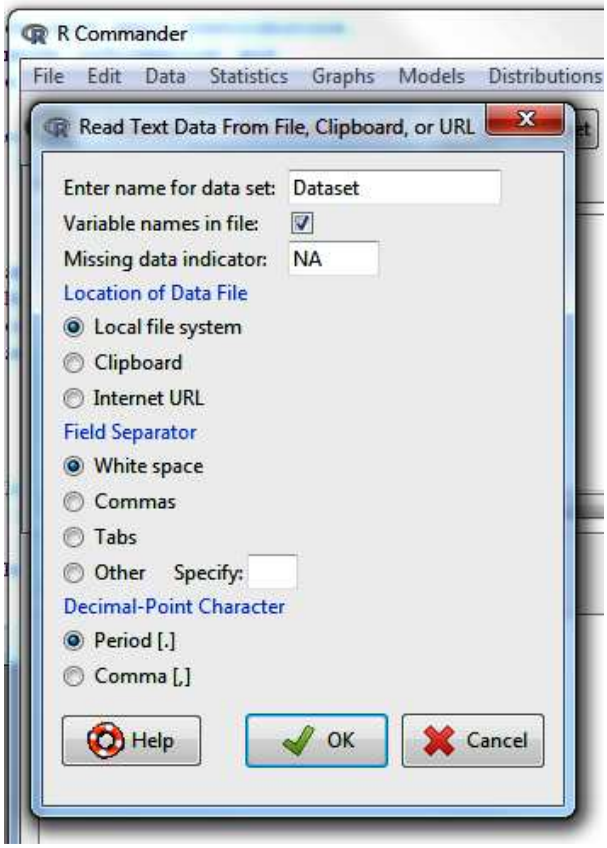


Si posteriorment volem carregar aquestes dades, només hem de fer servir l'opció del menú *Data→Load data set* i buscar l'arxiu corresponent mitjançant la finestra del navegador que s'obre.

3.3 Importació de dades

3.3.1 Fitxers de text

Per carregar un fitxer de text anem a l'opció del menú *Data→ Import data → from text file, clipboard,...* S'obre una finestra com la de la figura següent en la qual hem de triar les opcions del fitxer *dades.txt*:



Enter name for data set: Per exemple, Dades.

Variables names in file: activat.

Missing data indicator: el deixem igual.

Field indicator: White space.

Caràcter decimal: period.

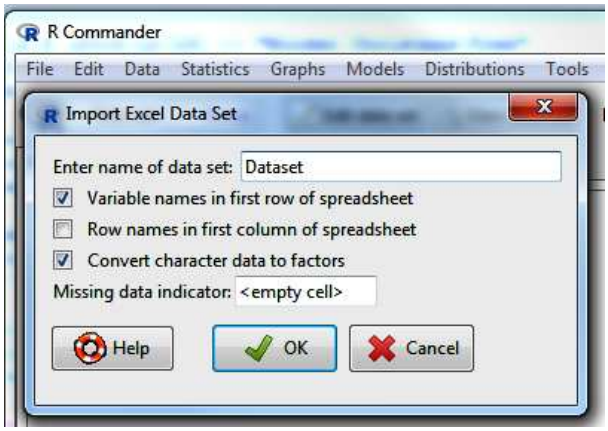
Com veiem, es pot escollir entre buscar les dades a un arxiu del nostre disc dur (sistema d'arxiu local) o bé des del porta papers. En el primer cas, s'obre una finestra de l'explorador perquè trobem l'arxiu i el seleccionem. Ara el conjunt de dades actiu és Dades. Si ho desitgem, podem

guardar aquest conjunt de dades actiu amb format .rda o perquè la pròxima vegada no l'haguem d'importar de nou.

3.3.2 Fitxers Excel

En el cas dels arxius tipus Excel, R Commander no necessita que li diguem res, ja que detecta automàticament els noms de les variables si estan presents. No obstant això, aquests no han d'incloure caràcters estranys, i han d'estar tots els noms de totes les variables o cap; en qualsevol altre cas, la importació podria ser invàlida.

Tan sols hem d'utilitzar l'opció del menú *Data → Import data → from Excel file...,* triant després l'arxiu a través de la finestra del navegador.

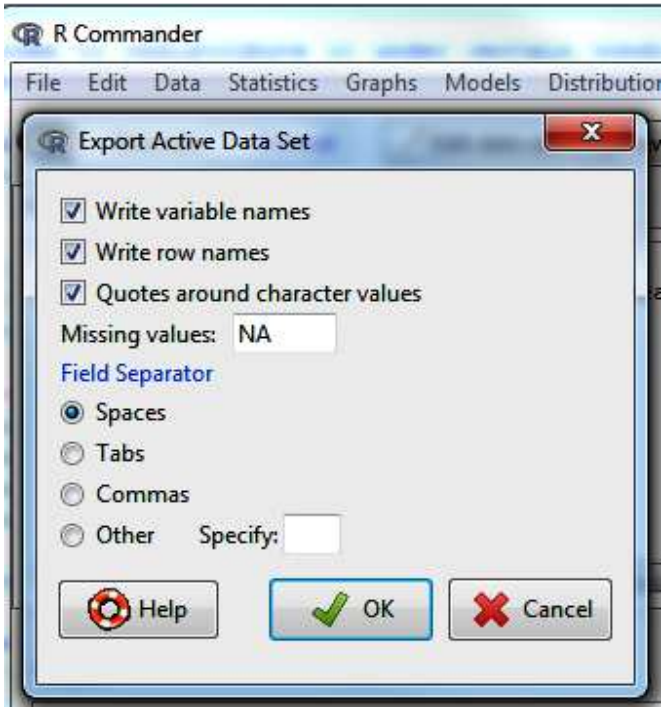


3.3.3 Fitxers SPSS

En el cas dels arxius de dades de SPSS, R Commander tampoc necessita que li diguem res. Si triem les opcions per defecte, tan sols hem d'utilitzar l'opció del menú *Data → Import data → from SPSS file...,* triant després l'arxiu a través de la finestra del navegador.

3.4 Exportació de dades

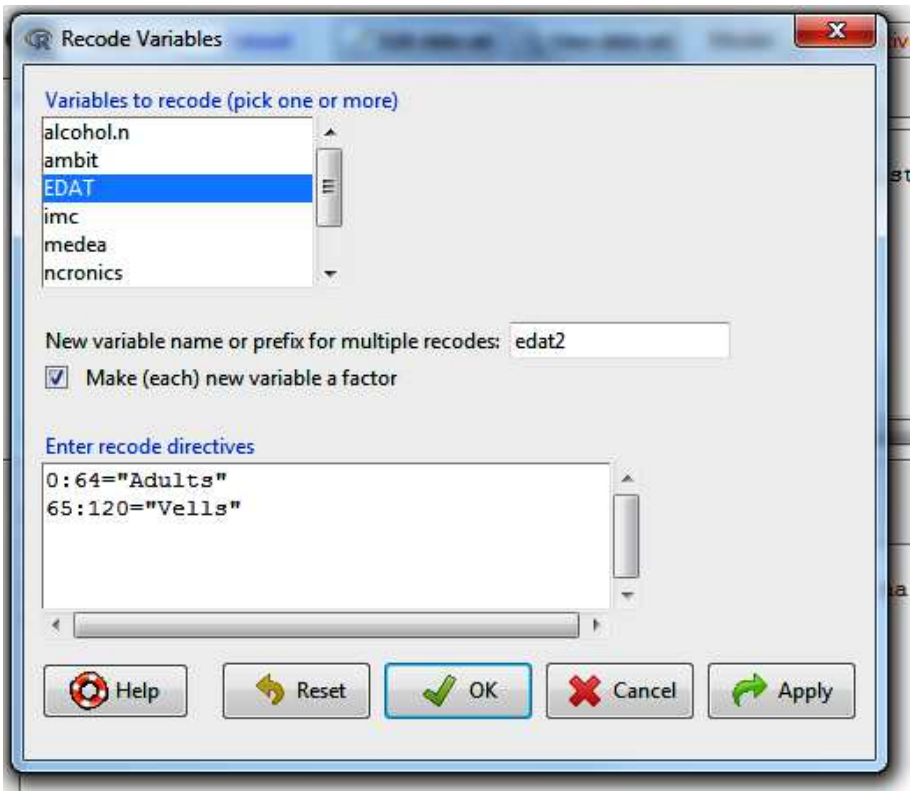
Es pot exportar el conjunt de dades actius a fitxers de text. Utilitzarem l'opció del menú *Data→Active data set→Export active data set...* Les opcions poden ser modificades, però les que apareixen per defecte són les més recomanables.



De moment, amb l'R commander bàsic no és possible exportar dades a altres tipus de fitxers. Però un fitxer de text pot ser obert per qualsevol software estadístic.

3.5 Recodificar variables

Anem a veure com es fa una recodificació mitjançant R Commander. Importem en primer lloc el fitxer *dades.txt*. Seleccionem l'opció *Data → Manage variables in active data set → Recode variables...* Ens apareix la finestra de la Figura següent.



Ja estan incloses les entrades necessàries per a la nostra recodificació:

1. La variable a recodificar: EDAT.

2. El nom de la nova variable: edat2.

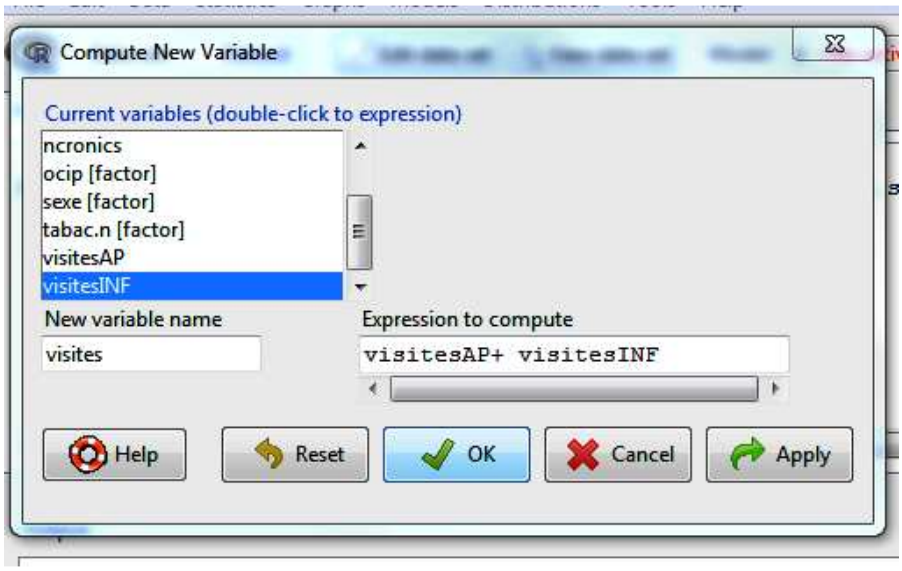
3. Les condicions que determinen la recodificació. Per especificar tots els números entre un valor *a* i un altre valor *b* s'ha de posar *a:b*. D'altra banda, com volem que els nous valors siguin caràcters (Jove, Adult), s'han d'escriure entre cometes.

4. L'opció *Make (each) new variable a factor* es deixa activada perquè la variable recodificada sigui considerada com un factor.

Aquest procediment també és vàlid per a les variables categòriques.

3.6 Calcular noves variables

Un cop importades les dades del fitxer *dades.txt*, utilitzem l'opció del menú *Data → Manage variables in active data set → Compute new variable...* Això ens obrirà una finestra que funciona bàsicament com una calculadora. El que farem és calcular la suma de visites a Metges i a Infermeria d'Atenció Primària.



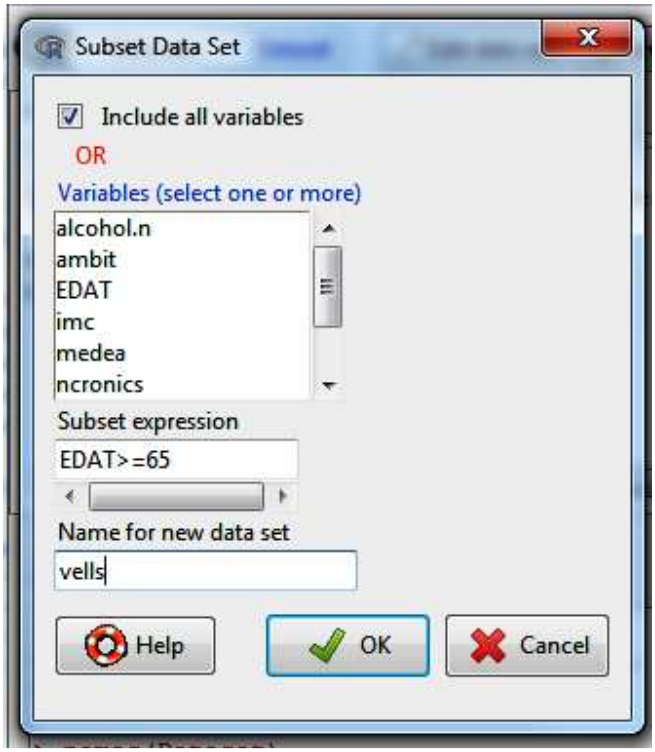
La Figura anterior ens mostra l'entrada de dades. En el càlcul de visites es pot visualitzar tot el contingut de la finestra.

3.7 Filtres de dades

Es pot crear un filtre de dades de la següent manera:

1. Seleccionant al menú *Data→Active data set→Subset active data set*.

2. A la finestra emergent podem seleccionar si volem quedar-nos amb totes les variables o triar només algunes.



3. La casella més important és la de *Subset expression*: aquí hem d'escriure l'expressió lògica que determini el nostre filtre. Per exemple, en el nostre cas podria ser EDAT>=65.

4. Finalment, és recomanable posar-li un nom al nou conjunt de dades filtrat diferent de l'original, per evitar sobreescriure'l. En el nostre cas l'hem anomenat *vells*.

3.8 Desar scripts i resultats

Anem a exemplificar aquest apartat amb l'exemple de crear una nova variable. Hem d'importar de nou el fitxer *dades.txt* i crear la variable visites.

A continuació seleccionem en el menú *File → Save script...* Ens demanarà el nom i la ruta on guardar el fitxer d'instruccions, que tindrà extensió .R. Una bona idea per nomenar els fitxers d'instruccions és posar com a nom la data del dia, per exemple, 16_01_16. No cal escriure l'extensió (però tampoc l'esborrem): ho farà el propi programa. Podem i hem de seguir guardant les instruccions amb posterioritat, triant de nou *Save script...*, però ja no ens demanarà de nou un nom, llevat que triem *Save script as...*

Ara anem a reiniciar R Commander i triem al menú *File→Open script file...* i seleccionem el fitxer d'instruccions que abans hem guardat. Com podem veure, apareixen les dues línies amb les que hem carregat les dades i hem creat la variable. Podem executar aquestes línies directament de la finestra d'instruccions sense haver d'utilitzar el menú. De la mateixa manera, podem guardar i recuperar posteriorment tot el que apareix a la finestra de resultats, mintjançant l'opció *File → Save output...* Els fitxers de resultats que R Commander crea són fitxers de text, amb extensió .txt.