R y otros recursos

F. Tusell*

Curso 2.010-2.011

1. Introducción

El sistema y lenguaje de programación S fue un proyecto innovador en Bell Labs (ATT) de lenguaje para análisis estadístico y gráfico.

S-Plus es una versión comercial que añade algunas cosas a la original de ATT. Ha sido periódicamente actualizada, y es hoy un programa que recoge el estado del arte en multitud de técnicas estadísticas, ausentes de otros muchos paquetes.

R es una re-implementación libre del lenguaje S que, a los efectos del trabajo en las asignaturas de Estadística de la Facultad, es más que suficiente (y, en muchos aspectos, superior a S-Plus). La versión actual es la 2.11.1.

R, sobre el que encontrarás información más abajo, es el instrumento para realizar tus prácticas en varias asignaturas. El tiempo que dediques a aprender su uso te compensará con creces.

2. Transcripción de una sesión en R

El Apéndice A es la transcripción de una sesión, mostrando el resultado de ejecutar algunos de los comandos más usuales. No es sustituto de la imprescindible consulta de los manuales que recoge la bibliografía. El mismo código es ejecutable bajo S-Plus, sin diferencias aparentes.

3. Recursos

En general, aunque se mencione S-Plus o R, la bibliografía sobre uno es también utilizable con el otro.

*Actualización del día 19 de octubre de 2010. La última versión de este documento, posiblemente más reciente, puede obtenerse en `http://www.et.bs.ehu.es/~etptupaf`.

3.1. Bibliografía

Básica. Es útil leer los primeros capítulos de Becker et al. (1988) (el "Libro Azul") u otro libro introductorio como Krause and Olson (1997) o Dalgaard (2002), al tiempo que se practica sobre la máquina. El manual Chambers and Hastie (1992) (el "Libro Blanco") describe muchas funciones añadidas al primitivo `S`, que mostrarán su utilidad, sobre todo, en regresión avanzada y no paramétrica.

Hay bastantes libros de Estadística construidos sobre la base de ejemplos en S-Plus o R también interesantes: son Venables and Ripley (1999a), Militino (2001), Ugarte et al. (2008), el ya citado Dalgaard (2002), Faraway (2005) y Everitt (1994).

Especializada. Puedes también encontrar de interés Fox (2002) (regresión lineal con R), Huet et al. (1996) (regresión no lineal en S-Plus), Härdle (1990) (suavizado en S-Plus), Bruce and Gao (1996) (wavelets en S-Plus), Bates and Pinheiro (2000) (modelos de efectos mixtos), Tableman and Kim (2003) (análisis de supervivencia), Spector (2008) (lectura y manipulación de datos en R), Bivand et al. (2008) (análisis de datos espaciales).

El sistema gráfico de R es de gran potencia y sofisticación. Soporta desde los gráficos más simples y tradicionales hasta el estado del arte en gráficos ligados y dinámicos, vía en este último caso librerías externas. Murrell (2006) es una buena referencia sobre gráficos en R describiendo tanto el modelo convencional de gráficos como el "nuevo", basado en el paquete `grid`. Sarkar (2008) describe el paquete `lattice`, proporcionando gráficos Trellis, Wickham (2009) describe una implementación sobre R de los principios descritos en Wilkinson (1999) y Cook et al. (2008) describe elinterface con `ggobi` (gráficos dinámicos).

Crawley (2007) es un libro con vocación enciclopédica sobre R. Inevitablemente, con software que, como R, cambia con gran dinamismo, libros como éste rápidamente quedan obsoletos.

Libros bastante más avanzados, que seguramente no te interesarán a menos que desees programar con relativa sofisticación en S-Plus o R, son Chambers (1998), Chambers (2008) y Venables and Ripley (2000).

Todas las obras citadas están disponibles en la Biblioteca de Investigación de la Facultad[1]. Para facilitarte la búsqueda, se consigna junto a algunos libros la signatura topográfica, que está sujeta a cambio (hasta que acabe el proceso de reclasificación de los fondos bibliográficos de la Facultad). La información definitiva y fiable la puedes obtener mirando el catálogo de la biblioteca de la UPV/EHU[2].

[1] A la que tienes acceso en tu condición de estudiante de segundo ciclo.

[2] En `http://www.biblioteca.ehu.es`.

3.2. Software

3.2.1. S-Plus

S-Plus, una versión comercial de S, está, como se ha indicado, disponible para bastantes plataformas diferentes, incluidos ordenadores personales bajo Windows 95/98/XP/NT. Es un programa bastante caro, y normalmente fuera del presupuesto del estudiante medio. Puedes dirigirte al proveedor MathSoft, Inc.[3], a su distribuidor en España Add-Link, S.A.[4] si crees que a pesar de todo puede ser de tu interés.

3.2.2. R

Una alternativa es el paquete R, que replica en buena medida la sintaxis de S-Plus y es libremente copiable. Lo tienes a tu disposición para ordenadores personales bajo Windows y otros muchos sistemas operativos[5]; encontrarás referencias en en las páginas Web del curso. Además de la bibliografía detallada en la Sección 3.1, hay mucha información libremente disponible: existe un pequeño manual (en inglés) Venables et al. (1997), una traducción del mismo al castellano y un manual de referencia extensivo documentando cada función. Claros y bien escritos, Maindonald (2000) y Kuhnert and Venables (2005) complementan bien a los anteriores[6]. Los libros recientes sobre S-Plus tratan también de R. Por ejemplo, Venables and Ripley (1999a) (con sus complementos Venables and Ripley (1999b)).

Para realizar las tareas de curso apenas notarás la diferencia entre trabajar con S-Plus o con R. Si haces programación avanzada, sí hay diferencias sustanciales, sobre todo en lo que se refiere a la visibilidad de los objetos[7].

3.2.3. Colecciones de añadidos a S-Plus y R

Si necesitas hacer uso de técnicas que no forman parte de R standard, tienes buenos motivos para mirar en `http://cran.at.r-project.org`; encontrarás casi todo lo imaginable.

3.2.4. Quantian

Si preferirías usar Linux pero te resulta difícil instalarlo en un ordenador, una solución a medio camino es emplear un DVD "live": se puede arrancar Linux desde un DVD sin necesidad de instalar nada en el disco duro.

[3] En `http://www.insightful.com`.

[4] En `http://www.addlink.es`

[5] Notablemente, para Linux, que te recomendamos si puedes utilizarlo.

[6] Todo puedes tomarlo de `http://www.et.bs.ehu.es`.

[7] R emplea *lexical scoping*, una aproximación radicalmente diferente a la de S-Plus. No notarás la diferencia hasta que comiences a escribir funciones que definen (y llaman) a otras. Entonces te convendrá leer Venables and Ripley (2000).

Hay muchas distribuciones de Linux "live". Knoppix[8] es una de las más conocidas. Quantian[9] es Knoppix ampliado con casi cualquier aplicación que pueda concebiblemente ser de utilidad para el cálculo científico. Entre ellas, R y todas las citadas más arriba.

3.3. Listas de correo

Hay una lista de correo que sirve de instrumento de comunicación entre los usuarios de R. Es un recurso valioso: se aprende mucho leyendo las preguntas y respuestas, que muchas veces tocan no sólamente cuestiones relativas a R, sino también de metodología estadística.

Cualquiera puede suscribirse, pero para evitar la proliferación de mensajes iguales dirigidos a personas en una misma institución, hay instalada[10] una colección de los mensajes recientes. Puedes acceder a ellos por tema o hacer búsquedas.

3.4. Laboratorio de Economía Cuantitativa

Es un espacio de trabajo en el que puedes disponer de una mesa y un ordenador para tu uso exclusivo. Dispones también de acceso a un servidor de mayor potencia que tu máquina de sobremesa para trabajos de cierta entidad.

Tus ficheros son accesibles externamente, de manera que puedes también trabajar desde tu casa o desde cualquier otro lugar[11].

4. Forma de uso de R

4.1. Sobre Linux

Probablemente lo más cómodo y efectivo. Pide información en clase.

4.2. Sobre Windows 95/98/XP/NT

4.2.1. Uso interactivo

Para ejecutar R lo normal será pinchar sobre el icono correspondiente. Crea un acceso directo en tu Escritorio para simplificarte la vida. Si emplear el CD ROM que te ha suministrado tu profesor, el icono de acceso a R es creado por la instalación.

[8] Puedes ver información en `http://www.cylnux.org/knoppix-es/`.

[9] Información en `http://dirk.eddelbuettel.com/quantian.html`.

[10] En `http://www.et.bs.ehu.es`.

[11] El acceso se realiza exclusivamente sobre conexiones seguras con SSH. Hay clientes gratuitos disponibles para Windows, y es standard en Linux.

4.2.2. Uso en batch

Abre una ventana de comandos y teclea: `R CMD BATCH` *fichero.R*. R procesará los mandatos en *fichero.R* proporcionando los resultados en *fichero.Rout*. Para que esto funcione, R debe estar en el "path". La manera de lograrlo difiere según se emplee una versión u otra de Windows.

4.3. Usando Emacs como interface

Pincha el icono de Emacs que se habrá creado durante la instalación en tu Escritorio. Cuando se abra el editor, teclea: `ESC x R`. Se iniciará una sesión R dentro de Emacs, lo que suministra un entorno de trabajo particularmente cómodo y productivo. Puedes alternar interactividad y proceso batch. Verás una sesión demostrativa en clases prácticas.

Referencias

D. M. Bates and J. C. Pinheiro. *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag, 2000. Signatura: 519.233.4.

R. A. Becker, J. M. Chambers, and A. R. Wilks. *The New S Language. A Programming Environment for Data Analysis and Graphics*. Wadsworth & Brooks/Cole, Pacific Grove, California, 1988.

R. S. Bivand, E. J. Pebesma, and V. Gómez-Rubio. *Applied Spatial Data Analysis with R*. Springer Verlag, 2008.

A. Bruce and H.-Y. Gao. *Applied Wavelet Analysis with S-Plus*. Springer Verlag, 1996.

J. M. Chambers. *Software for Data Analysis: Programming with R*. Springer Verlag, 2008.

J. M. Chambers. *Programming with Data*. Mathsoft, 1998.

J. M. Chambers and T. J. Hastie. *Statistical Models in S*. Wadsworth & Brooks/Cole, Pacific Grove, Ca., 1992.

D. Cook, D. F. Swayne, A. Buja, and D. T. Lang. *Interactive and Dynamic Graphics for Data Analysis: With R and GGobi*. Springer-Verlag, 2008.

M. Crawley. *The R Book*. Wiley, 2007. ISBN 0-470-51024-2. Signatura: 519.682 CRA.

P. Dalgaard. *Introductory Statistics with R*. Statistics and Computing. Springer-Verlag, 2002. Signatura: 519.682 DAL.

B. S. Everitt. *A Handbook of Statistical Analysis in S-Plus*. Chapman and Hall, London, 1994.

J. J. Faraway. *Linear Models with R*. Chapman & Hall/CRC, 2005. Signatura: 519.233 FAR.

J. Fox. *An R and S-Plus Companion to Applied Regression*. Sage Pub., 2002.

W. Härdle. *Smoothing Techniques with Implementations in S*. Springer Verlag, New York, 1990.

S. Huet, A. Bouvier, M. A. Gruet, and E. Jolivet. *Statistical Tools for Nonlinear Regression with S-Plus*. Springer Verlag, 1996.

A. Krause and M. Olson. *The Basics of S-Plus*. Springer Verlag, 1997. Signatura: 519.682 KRA.

P. Kuhnert and W. Venables. *An Introduction to R: Software for Statistical Modelling and Computing*. CSIRO Mathematical and Information Sciences, Cleveland, Australia, 2005.

J. H. Maindonald. Data analysis and graphics using R - An introduction. January 2000. URL `\url{http://www.maths.anu.edu.au/\~{}johnm/r-notes/r-notes.pdf}`.

M. D. U. y. A. F. Militino. *Estadística Aplicada con S-Plus*. Universidad Pública de Navarra, 2001.

P. Murrell. *R Graphics*. Chapman and Hall/CRC, 2006.

D. Sarkar. *Lattice. Multivariate Data Visualization with R*. Springer, 2008. doi: 10.1007/978-0-387-75969-2.

P. Spector. *Data Manipulation with R*. Springer, 2008. doi: 10.1663/978-0-387-74731-6.

M. Tableman and J. S. Kim. *Survival Analysis Using S*. Chapman & Hall/CRC, 2003. 519.2.001.6 TAB.

M. Ugarte, A. Militino, and A. Arnholt. *Probability and Statistics with R*. CRC Press, 2008.

B. Venables, D. Smith, R. Gentleman, and R. Ihaka. *Notes on R: A Programming Environment for Data Analysis and Graphics*. Dept. of Statistics, University of Adelaide and University of Auckland, 1997. Available at `http://cran.at.r-project.org/doc/R-intro.pdf`.

W. Venables and B. Ripley. *Modern Applied Statistics with S-Plus*. Springer-Verlag, New York, third edition, 1999a.

W. Venables and B. D. Ripley. *S Programming*. Springer-Verlag, 2000.

W. Venables and B. D. Ripley. R complements to *Modern Applied Statistics with S-Plus*. En `http://www.stats.ox.ac.uk/pub/MASS3`, 1999b.

H. Wickham. *ggplot2 : elegant graphics for data analysis*. Springer-Verlag, 2009.

L. Wilkinson. *The Grammar of Graphics*. Springer, 1999.

A. Transcripción de una sesión

```
R version 2.11.1 Patched (2010-09-04 r52880)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.

R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.

Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.

[Previously saved workspace restored]

> 3 + 5                          # Suma directa de dos números.
[1] 8
> 4 / 3                          # División directa.
[1] 1.333333
> g <- 20                        # <- es el operador de asignación.
> g                              # Tecleando una variable,
[1] 20
>                                # vemos su valor.
> 10 + g                         # ... y lo podemos emplear
[1] 30
>                                # como operando.
> a <- c(2,3,9,1,6,7)            # c() concatena numeros
>                                # para dar vectores.
> a
[1] 2 3 9 1 6 7
> a[1]                           # Podemos referenciar un
[1] 2
>                                # elemento,
> a[c(1,3)]                      #   o varios,
[1] 2 9
> a[-3]                          #   o todos menos uno,
[1] 2 3 1 6 7
>                                # dando el subíndice, o un
>                                # vector lógico (valores T o F).
>                                # Esto último en particular
>                                # es un modo útil de recuperar
>                                # elementos verificando una
>                                # condición. Por ejemplo, los
```

```
>                                  # mayores de 3 en a se
>                                  # obtendrían así:
> a[a > 3]
[1] 9 6 7
>                                  # El argumento a>3 actua como
>                                  # selector lógico. Su valor es:
> a > 3
[1] FALSE FALSE  TRUE FALSE  TRUE  TRUE
> b <- a*a                         # Producto elemento a elemento.
> b
[1]  4  9 81  1 36 49
> d <- matrix(a,2,3,byrow=T)       # También podemos crear matrices.
> d
     [,1] [,2] [,3]
[1,]    2    3    9
[2,]    1    6    7
> d[2,2]                           # y referenciar elementos de ellas.
[1] 6
> d[2,c(1,2)]
[1] 1 6
> d[,1]                            # Cuando un subíndice se omite, se
[1] 2 1
> d[2,]                            # toma toda la fila (o columna).
[1] 1 6 7
>
> e <- t(d)                        # ... y trasponerlas.
> e
     [,1] [,2]
[1,]    2    1
[2,]    3    6
[3,]    9    7
> f <- d %*% e                     # %*% es el producto matricial.
> f
     [,1] [,2]
[1,]   94   83
[2,]   83   86
> h <- cbind(f,f)                  # Forma de empalmar por columnas.
> j <- rbind(f,f)                  #   y por filas. Los
>                                  #   resultados a continuación
> h
     [,1] [,2] [,3] [,4]
[1,]   94   83   94   83
[2,]   83   86   83   86
> j
     [,1] [,2]
[1,]   94   83
[2,]   83   86
[3,]   94   83
[4,]   83   86
```

```
> invf <- solve(f)                # La función solve con un
>                                 # sólo argumento invierte una
>                                 # matriz. Podemos comprobar que,
> af <- invf %*% f
>                                 # es (salvo truncamientos)
>                                 # la matriz unidad.
> af
             [,1] [,2]
[1,] 1.000000e+00    0
[2,] 8.881784e-16    1
>                                 # Hay varios tipos de datos.
>                                 # Podemos utilizar variables cuyo
>                                 # valor sea numérico (como más
>                                 # arriba) o alfanumérico:
>
> nombres <- c("Juan","Pedro" ,"Andres")
> nombres
[1] "Juan"   "Pedro"  "Andres"
>                                 # Podemos incluso crear "listas",
>                                 # objetos con elementos de
>                                 # diferentes tipos.
> ejemplo <- list(nombres,c(1,2,3))
> ejemplo
[[1]]
[1] "Juan"   "Pedro"  "Andres"

[[2]]
[1] 1 2 3

>                                 # Se puede leer un fichero, y
>                                 # almacenar su contenido en
>                                 # vectores o variables. Si
>                                 # fichero fuera un fichero ASCII
>                                 # (creado con el editor vi por
>                                 # ejemplo), y su contenido fuera,
>                                 #
>                                 #       5 4 2 3 4
>                                 #
>                                 # entonces,
>
> b <- scan("fichero",0)          # guarda dichos valores en  b
Read 5 items
> b                               #  ...como puede comprobarse.
[1] 5 4 2 3 4
> rm(ejemplo,b)                   # Este es el modo de borrar objetos
>                                 # cualesquiera. Hay que  hacerlo
>                                 # para no llenar el espacio de trabajo
>                                 # de cosas inservibles
> ls()                            # Para ver lo que queda
```

```
 [1] "a"       "af"      "d"       "e"       "f"       "g"       "h"
 [8] "invf"    "j"       "nombres"
>                                 # Este es el modo de abandonar
> q()                             # S-Plus ó R
> proc.time()
   user  system elapsed
  0.700   0.184   0.690
```