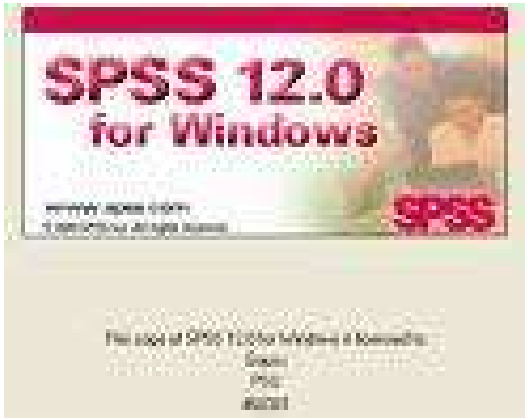
Laboratorio 0: Introducción a R y RCommander

Recomendado!!

Lo bueno..

Para la elección de R se han evaluado pues distintos aspectos, siendo especialmente destacables sus bondades en lo que se refiere a calidad, a la cantidad de técnicas y funciones implementadas, a que es libre y a la gran comunidad científica que lo usa como estándar para el análisis de datos. Dicha comunidad ha desarrollado y desarrolla herramientas integradas en paquetes–en la actualidad más de 800–, que dan solución a una gran variedad de problemas estadísticos.



Laboratorio 0: Introducción a R y RCommander

R-project

R es un lenguaje y entorno de programación para análisis estadístico y gráfico.

Se trata de un proyecto de software libre, resultado de la implementación GNU del premiado lenguaje S. R y S-Plus -versión comercial de S- son, probablemente, los dos lenguajes más utilizados en investigación por la comunidad estadística, siendo además muy populares en el campo de la investigación biomédica, la bioinformática y las matemáticas financieras. A esto contribuye la posibilidad de cargar diferentes librerías o paquetes con finalidades específicas de cálculo o gráfico.

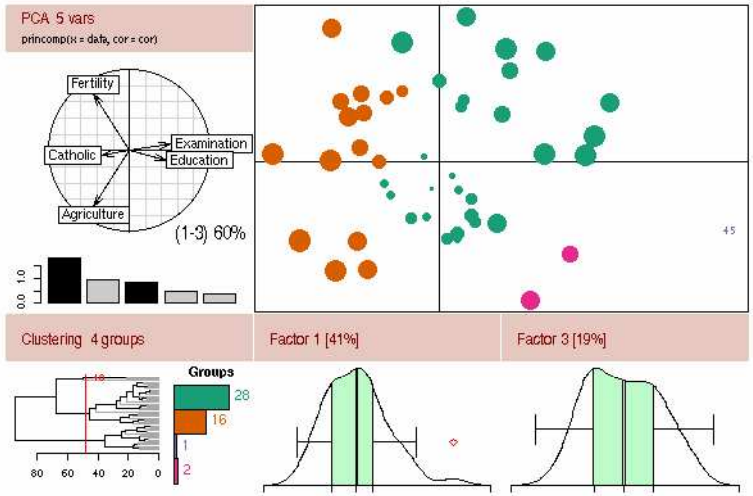
The R Project for Statistical Computing - Microsoft Internet Explorer

Archivo Edición Ver Favoritos Herramientas Ayuda

Atrás Búsqueda Favoritos

Dirección http://www.r-project.org/ Ir Vínculos

The R Project for Statistical Computing



About R
- What is R?
- Contributors
- Screenshots
- What's new?

Download
- CRAN

R Project
- Foundation
- Members & Donors
- Mailing Lists
- Bug Tracking
- Developer Page
- Conferences
- Search

Documentation
- Manuals
- FAQs
- Newsletter
- Wiki
- Books
- Certification
- Other

Misc
- Bioconductor
- Related Projects
- Links

Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.
- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News:

- **R version 2.8.1** has been released on 2008-12-22.
- **R News 8/2** has been published on 2008-11-03.
- **DSC 2009**, The 6th workshop on Directions in Statistical Computing, will be held at the Center for Health and Society, University of Copenhagen, Denmark, July 13-14, 2009.
- **useR! 2009**, the R user conference, will be held at Agrocampus Rennes, France, July 8-10, 2009.

Listo Internet

Lo bueno..

Una diferencia importante entre R, y también S, con el resto del software estadístico es el uso del `objeto` como entidad básica. Cualquier expresión evaluada por R tiene como resultado un objeto. Cada objeto pertenece a una clase, de forma que las funciones pueden tener comportamientos diferentes en función de la clase a la que pertenece su objeto argumento. Por ejemplo, el resultado de la función `print` evaluada sobre un vector da como resultado la impresión de todos los elementos

```
> x <- sample ( 6, 10, replace=T ) # take a sample of
                                   # size 10 and call it x
> x # print the ten values
[1] 6 4 2 3 4 4 3 6 6 2

> sum ( x == 3 ) # how many are equal to 3?
[1] 2
>
```

Lo bueno..

Documentación en español

[editar]

- R para Principiantes, la versión en español de "R for Beginners", traducido por Jorge A. Ahumada (PDF).
- Versión en español de "An Introduction to R" por Andrés González y Silvia González (PDF).
- Estadística Básica con R y R-Commander (libro libre) en knuth.uca.es.
- "Gráficos Estadísticos con R" por Juan Carlos Correa y Nelfi González (PDF).
- Cartas sobre Estadística de la Revista Argentina de Bioingeniería por Marcelo R. Risk (PDF).
- Introducción al uso y programación del sistema estadístico R por Ramón Díaz-Uriarte, transparencias preparadas para un curso de 16 horas sobre R, dirigido principalmente a biólogos y especialistas en bioinformática (PDF).

Libro libre: Estadística Básica con R y R-Commander
WWW: http://knuth.uca.es/ebrcmdr
ISBN: 978-84-9828-186-6

(En copisterías de la UCA y en librerías)
(Precio 6€, impreso en color)

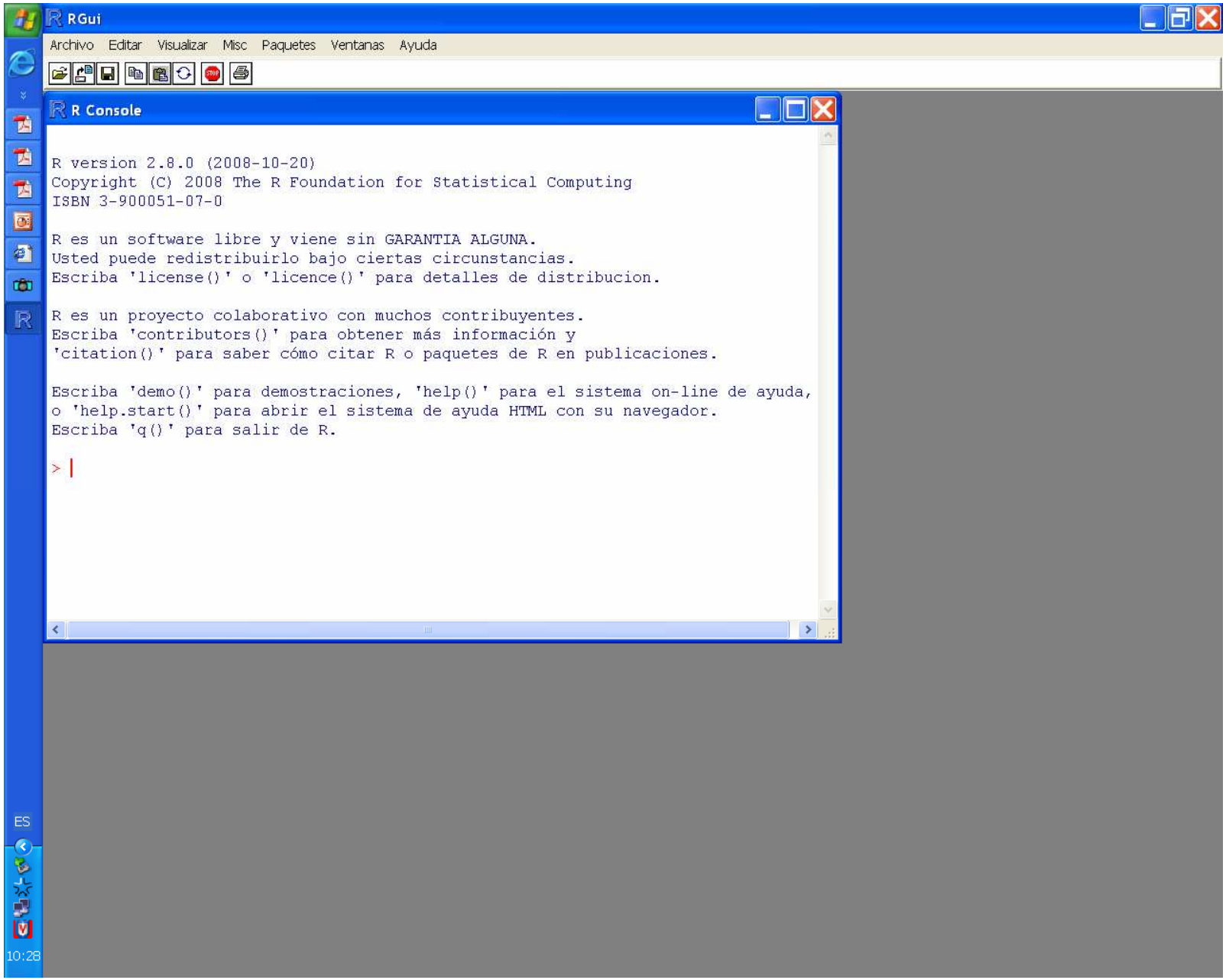
Fuentes
Bajo licencia: GNU Free Documentation License
(Licencia de documentación libre)
Más información en el prólogo

Agradecemos tu opinión: encuesta

Novedades

1 Estadística Básica con R y R-Commander
Comentarios
Libro por capítulos
Libro completo
Ficheros de datos

Lo menos bueno..



```
> mean(runif(10000))
[1] 0.4977687
>
```

Genera 10.000 de números aleatorios de una distribución uniforme y calcula su media

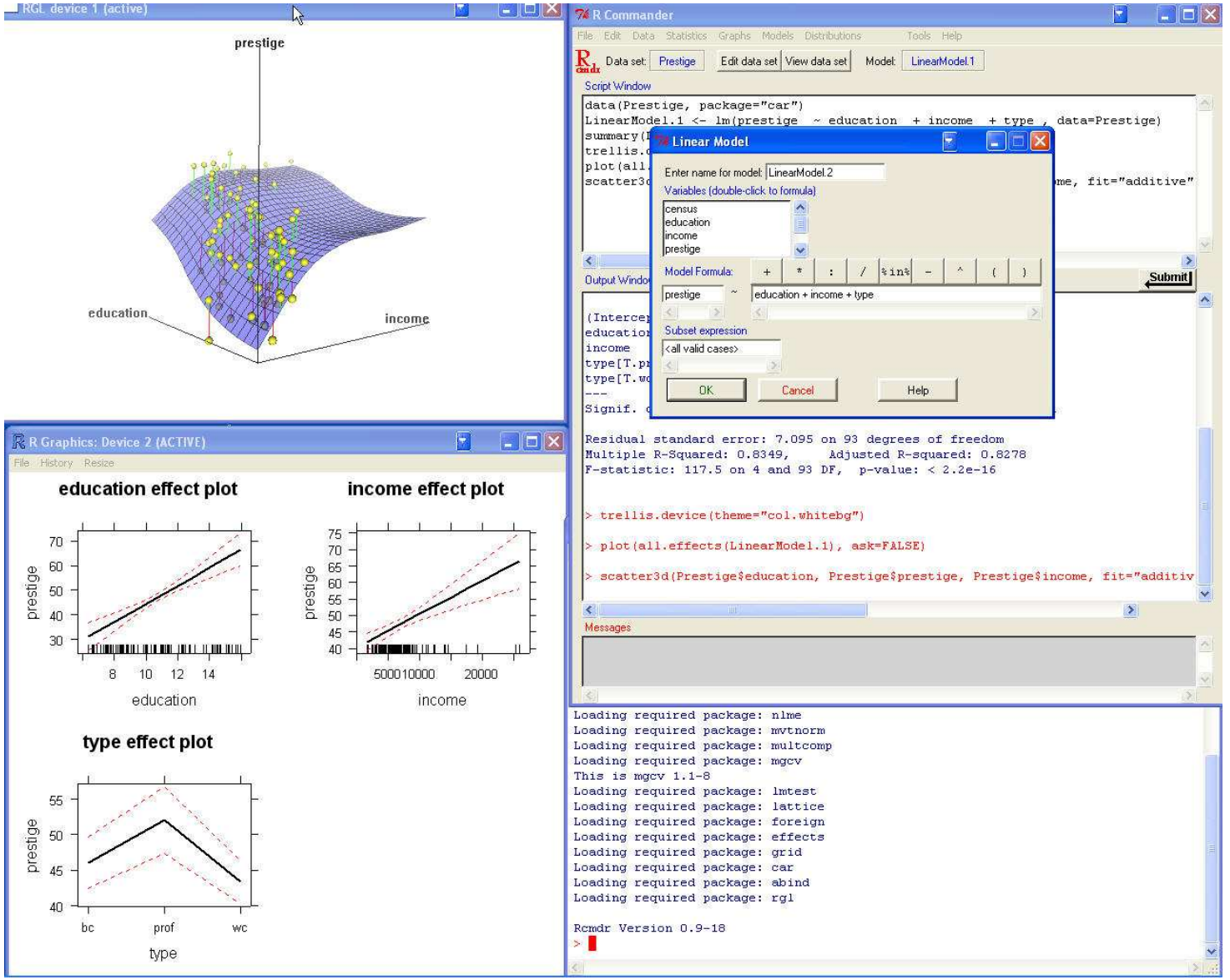
Tiempo empleado =0.000001 segundos

```
> mean(runif(10000))
[1] 0.4977687
> mean(runif(1000000))
[1] 0.4999676
>
```

Genera 1.000.000 de números aleatorios de una distribución uniforme y calcula su media

Tiempo empleado =0.000001 segundos

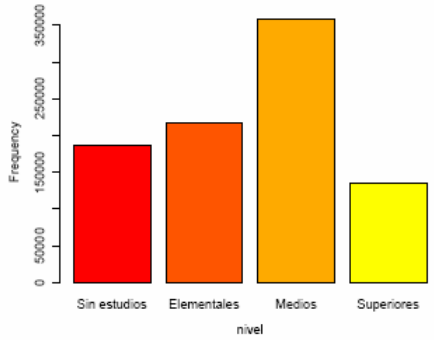
The R Commander



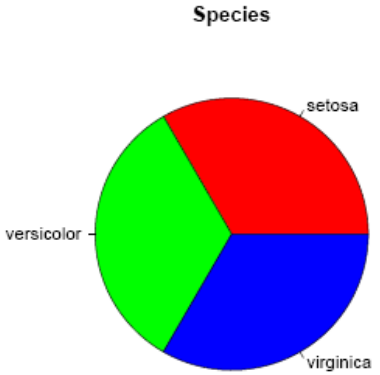
Análisis Exploratorio de Datos Unidimensional

En este módulo, a través de una serie de medidas, gráficos y modelos descriptivos, se caracterizará a un conjunto de individuos, intentando descubrir regularidades y singularidades de los mismos y, si procede, comparar los resultados con los de otros grupos, patrones o con estudios previos. Se podría considerar que este estudio es una primera entrega de un estudio más completo o, por contra, tener un carácter finalista; en cualquier caso, se trata de un análisis calificable como de *exploratorio*, y de ahí el nombre del capítulo.

```
> numSummary(Niv_estudios[,''niv_num''],
statistics=c(''quantiles''))
0%   25%   50%   75%   100%
0     1     2     2     3
```




```
> .Table <- table(iris$Species)
> .Table # counts for Species
setosa      versicolor      virginica
50             50              50
> 100*.Table/sum(.Table) # percentages for Species
setosa      versicolor      virginica
33.33333    33.33333        33.33333
```



Formalmente, se podría definir el AED como un conjunto de técnicas estadísticas cuya finalidad es conseguir un entendimiento básico de los datos y de las relaciones existentes entre las variables analizadas; aunque esta primera entrega se centrará en un análisis de tipo unidimensional.

Las conclusiones obtenidas serán aplicables exclusivamente a los individuos considerados explícitamente en el estudio, sin que puedan hacerse extrapolaciones con validez científica fuera de ese contexto. Los resultados del Análisis Exploratorio de Datos (AED) sí que podrían emplearse para establecer hipótesis sobre individuos no considerados explícitamente en dicho análisis, que deberían ser posteriormente contrastadas.

1. La organización de la información

Al conjunto de individuos físicos considerados en un análisis se le denominará *Colectivo* o *Población*, aunque también se utilizarán esos mismos términos para referirse a la(s) característica(s) de esos individuos que son objeto de estudio. De hecho, desde un punto de vista estadístico, los individuos sólo interesan como portadores de rasgos que son susceptibles de marcar diferencias entre ellos. La obtención y materialización en formato analógico o digital de las características consideradas constituirá el conjunto de datos que será estadísticamente analizado.

1.1. La matriz de datos

En una primera instancia se supondrá que, sobre un conjunto de n individuos físicos, se obtienen una serie de k caracteres u observaciones de igual o distinta naturaleza. Es importante tener en cuenta, ya desde este momento, que la calidad del análisis que se realice, va a depender de la habilidad que se tenga a la hora de seleccionar los caracteres que se obtendrán del conjunto de individuos seleccionados.

Los datos obtenidos se organizarán en una matriz $n \times k$, donde cada fila representa a un individuo o registro y las columnas a las características observadas. Las columnas tendrán naturaleza homogénea,

	TFR	contraception	infant.mortality	GDP	region
Afghanistan	6.90	NA	154	2848	Asia
Albania	2.60	NA	32	863	Europe
Algeria	3.81	52	44	1531	Africa

Trabajo con RCommander

Una vez que ha arrancado R, para arrancar RCommander se *carga* primero la librería Rcmdr tecleando:

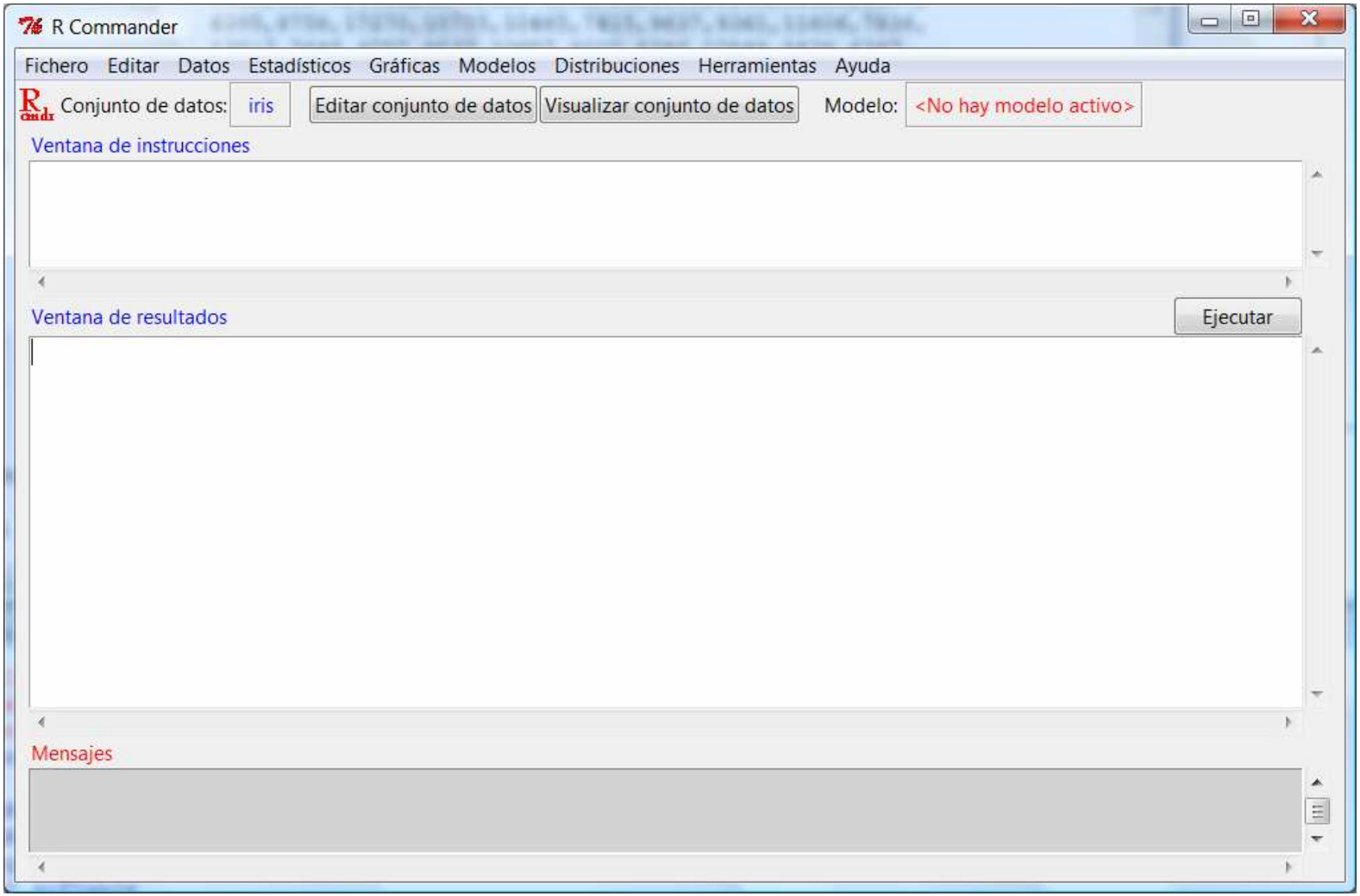
```
library(Rcmdr)
```

Alternativamente se puede hacer desde el menú superior de ventanas de R:

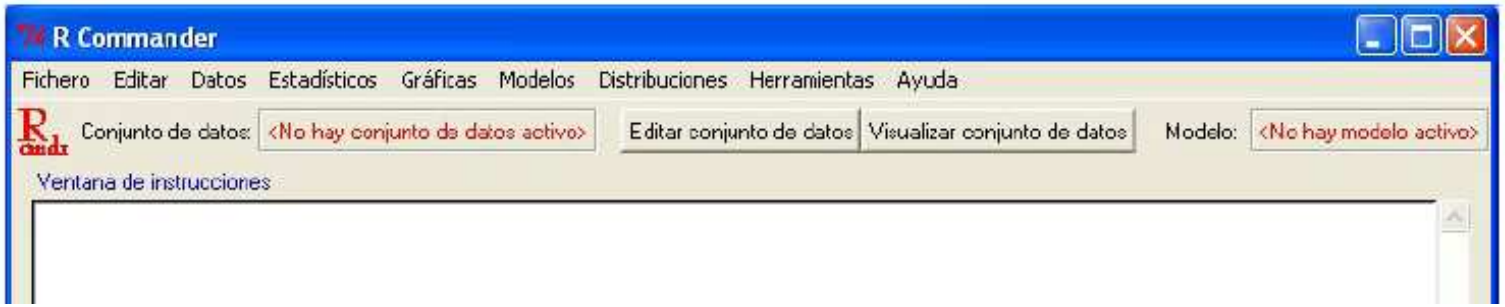
Paquetes → Cargar paquete → Rcmdr

Aparece un entorno de trabajo con tres partes:

- Ventana de instrucciones
- Ventana de resultados
- Mensajes

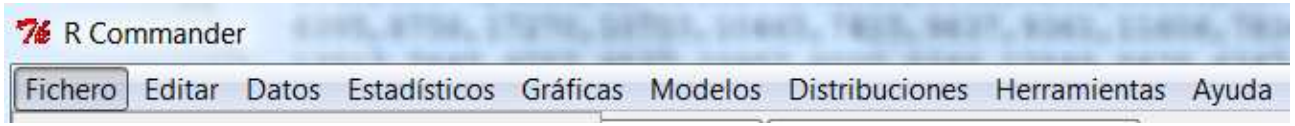


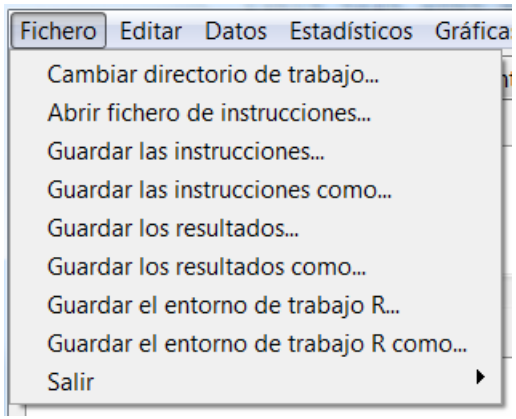
Menús

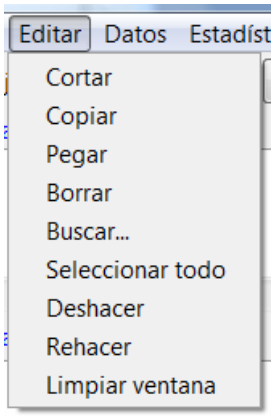


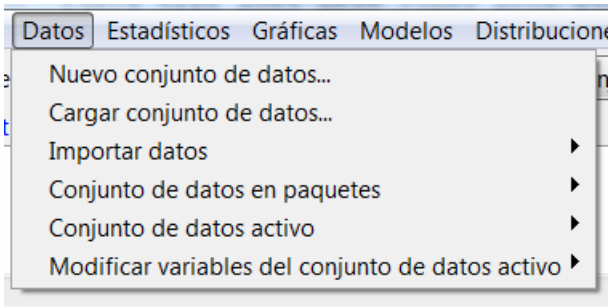
- FICHERO: Menú para cargar y grabar ficheros *script*, salidas y el R workspace.
- EDITAR: Submenús para editar los contenidos de los scripts y salidas. Pulsando con el botón derecho encima de las ventanas *script* y *output* aparece también un menú contextual.
- DATOS: Submenús para lectura y manipulación de datos.
- ESTADÍSTICOS: Submenús para realizar numerosos análisis estadísticos.

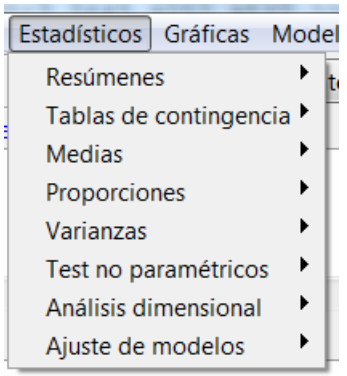
Menús

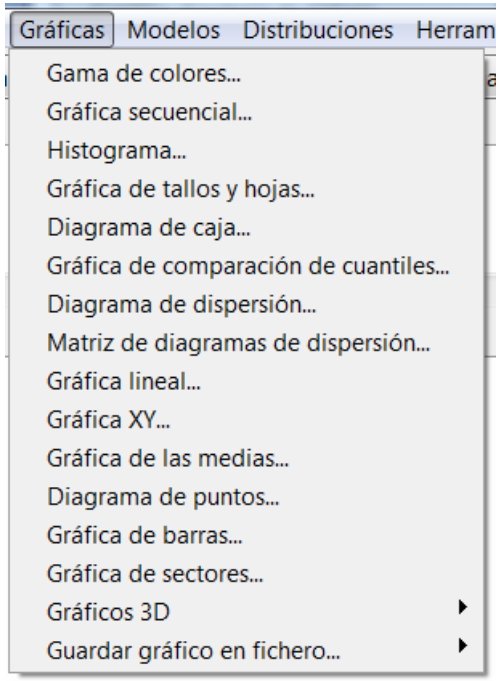

R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda

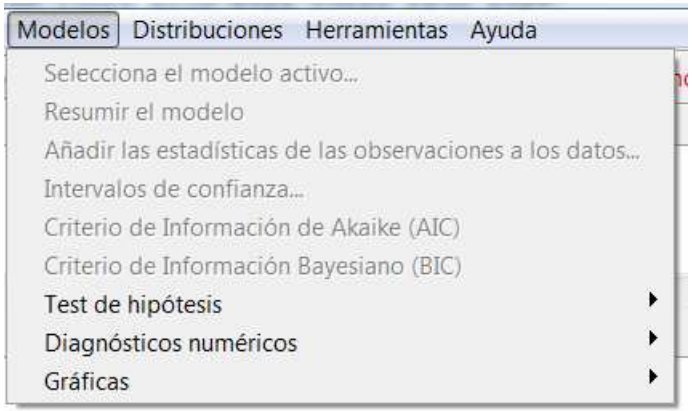

Fichero Editar Datos Estadísticos Gráfica
Cambiar directorio de trabajo...
Abrir fichero de instrucciones...
Guardar las instrucciones...
Guardar las instrucciones como...
Guardar los resultados...
Guardar los resultados como...
Guardar el entorno de trabajo R...
Guardar el entorno de trabajo R como...
Salir

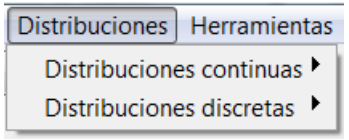

Editar Datos Estadíst
Cortar
Copiar
Pegar
Borrar
Buscar...
Seleccionar todo
Deshacer
Rehacer
Limpiar ventana

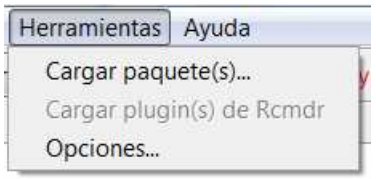

Datos Estadísticos Gráficas Modelos Distribucione
Nuevo conjunto de datos...
Cargar conjunto de datos...
Importar datos
Conjunto de datos en paquetes
Conjunto de datos activo
Modificar variables del conjunto de datos activo

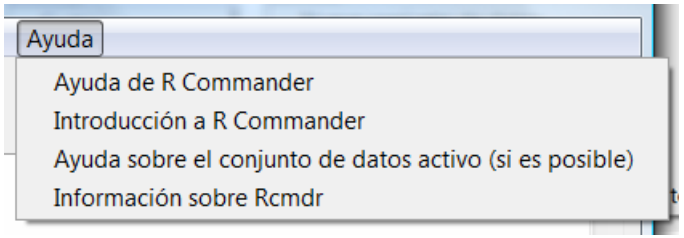

Estadísticos Gráficas Model
Resúmenes
Tablas de contingencia
Medias
Proporciones
Varianzas
Test no paramétricos
Análisis dimensional
Ajuste de modelos


Gráficas Modelos Distribuciones Herram
Gama de colores...
Gráfica secuencial...
Histograma...
Gráfica de tallos y hojas...
Diagrama de caja...
Gráfica de comparación de cuantiles...
Diagrama de dispersión...
Matriz de diagramas de dispersión...
Gráfica lineal...
Gráfica XY...
Gráfica de las medias...
Diagrama de puntos...
Gráfica de barras...
Gráfica de sectores...
Gráficos 3D
Guardar gráfico en fichero...


Modelos Distribuciones Herramientas Ayuda
Selecciona el modelo activo...
Resumir el modelo
Añadir las estadísticas de las observaciones a los datos...
Intervalos de confianza...
Criterio de Información de Akaike (AIC)
Criterio de Información Bayesiano (BIC)
Test de hipótesis
Diagnósticos numéricos
Gráficas


Distribuciones Herramientas
Distribuciones continuas
Distribuciones discretas


Herramientas Ayuda
Cargar paquete(s)...
Cargar plugin(s) de Rcmdr
Opciones...


Ayuda
Ayuda de R Commander
Introducción a R Commander
Ayuda sobre el conjunto de datos activo (si es posible)
Información sobre Rcmdr

En `Rcmdr` se asume que existe un conjunto de datos *activo*, y cuando comienza el programa no se asume ninguno por defecto por lo que lo primero que hay que hacer es cargar un conjunto de datos para poder trabajar.

Si hay varios conjuntos de datos en memoria, sólo se puede seleccionar uno al mismo tiempo para trabajar con él.

Formas de introducir los datos:

- Directamente escribiéndolos en el programa (para pequeños conjuntos de datos) usando

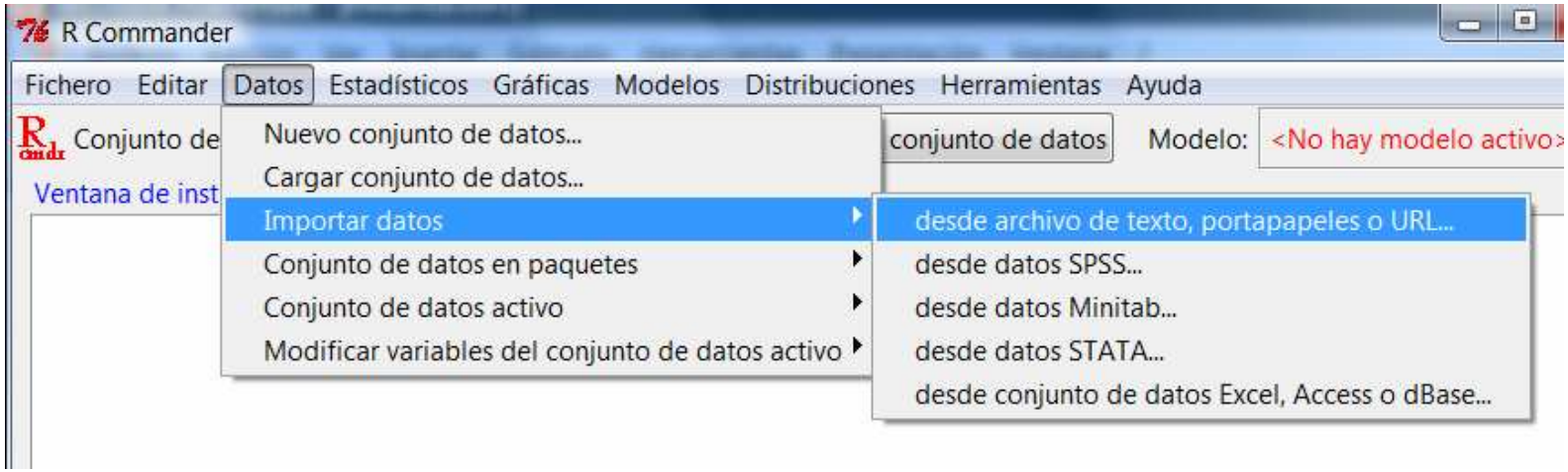
  Datos → Nuevo Conjunto de Datos

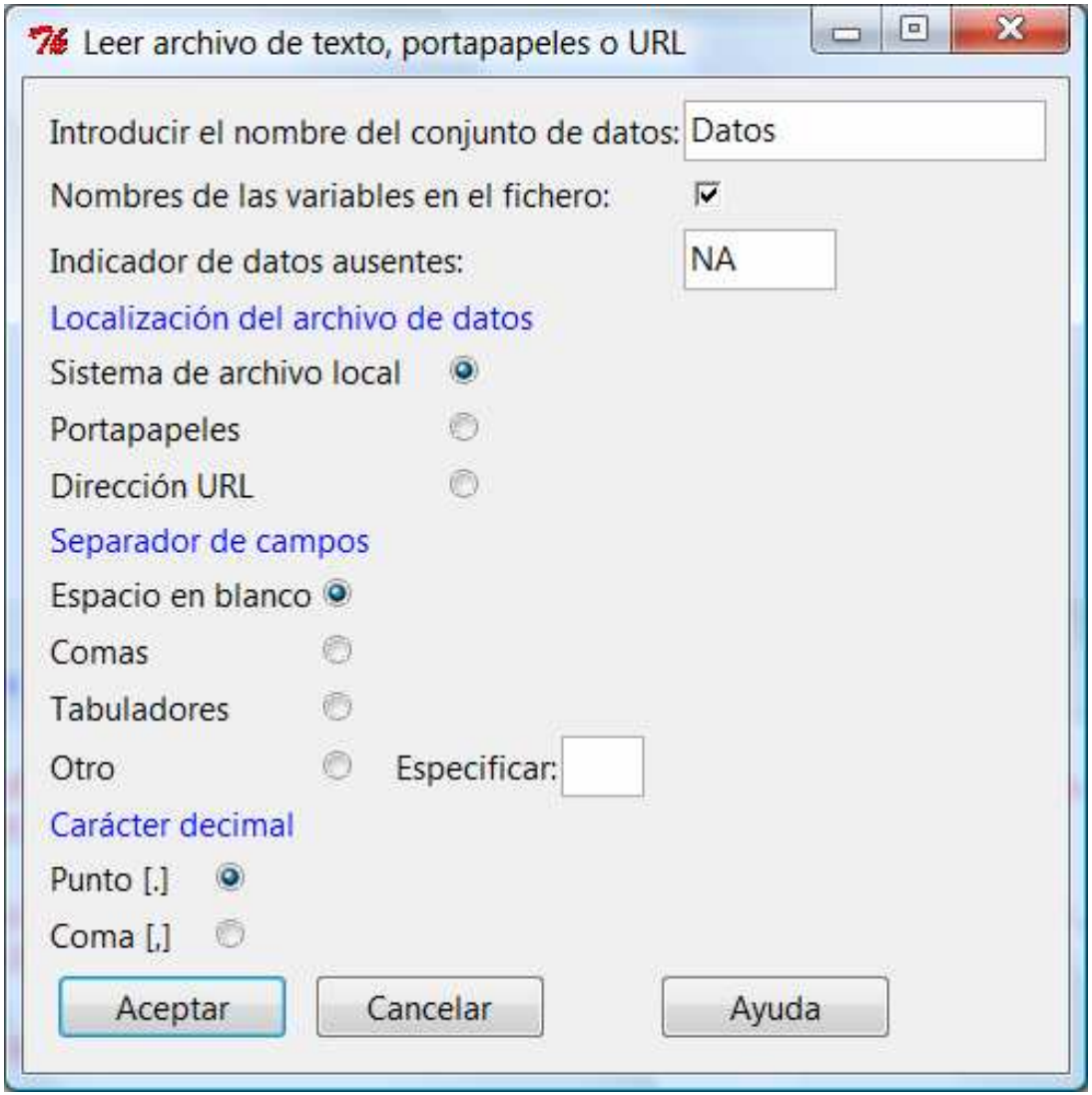
- Importando los datos desde un fichero `ASCII` o desde el *portapapeles*.
- Importando los datos desde un fichero de otros programas: `Minitab`, `SPSS`, `Stata`, `Excel`, `Access` ó `dBase`.
- Importar los datos incluidos en algún paquete o librería propia de `R`.

Ejemplo

Leer, por ejemplo, los datos del fichero `Nations.txt`.

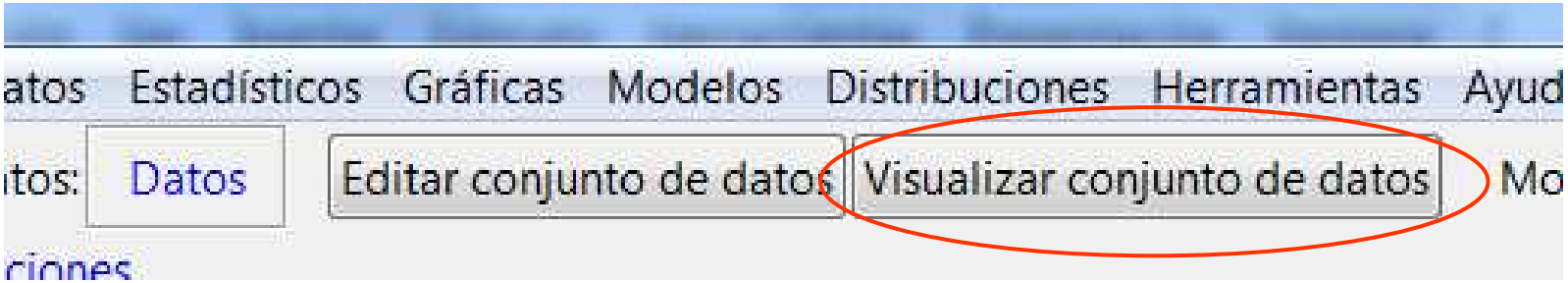
Datos → **Importar Datos** → **Desde archivo de texto o portapapeles** → **Directorio** (donde esté guardado el fichero `Nations.txt`)

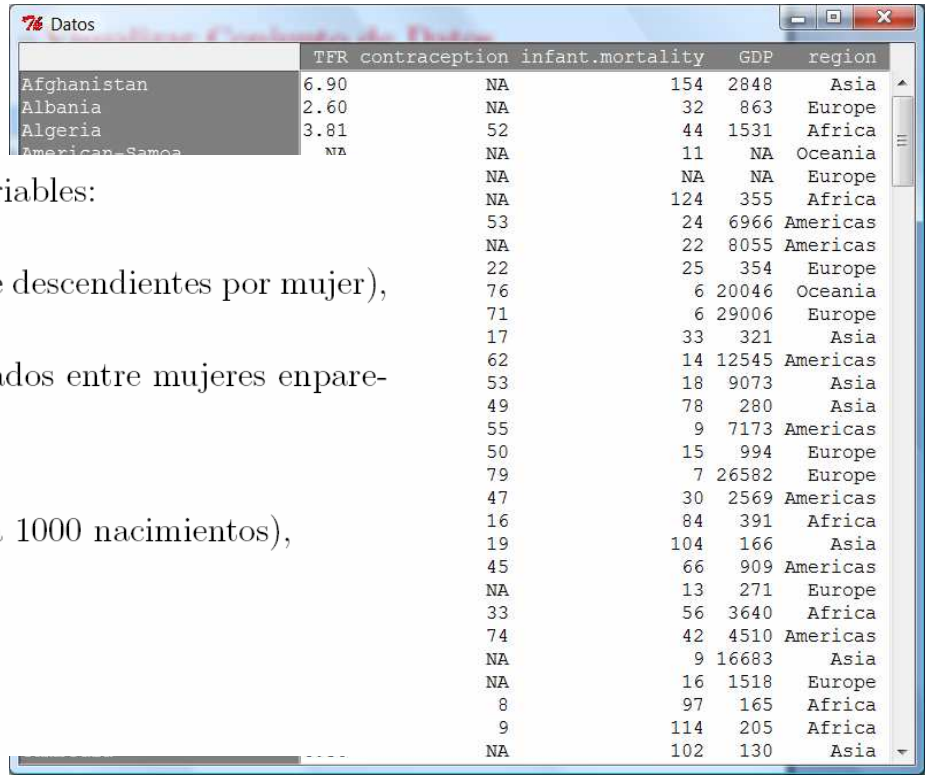



Leer archivo de texto, portapapeles o URL
Introducir el nombre del conjunto de datos:
Datos
Nombres de las variables en el fichero:
Indicador de datos ausentes:
NA
Localización del archivo de datos
Sistema de archivo local
Portapapeles
Dirección URL
Separador de campos
Espacio en blanco
Comas
Tabuladores
Otro
Especificar:
Carácter decimal
Punto [.]
Coma [,]
Aceptar
Cancelar
Ayuda

Para ver los datos que se acaban de importar, se pueden usar los botones

Editar Conjunto de Datos o **Visualizar Conjunto de Datos**.

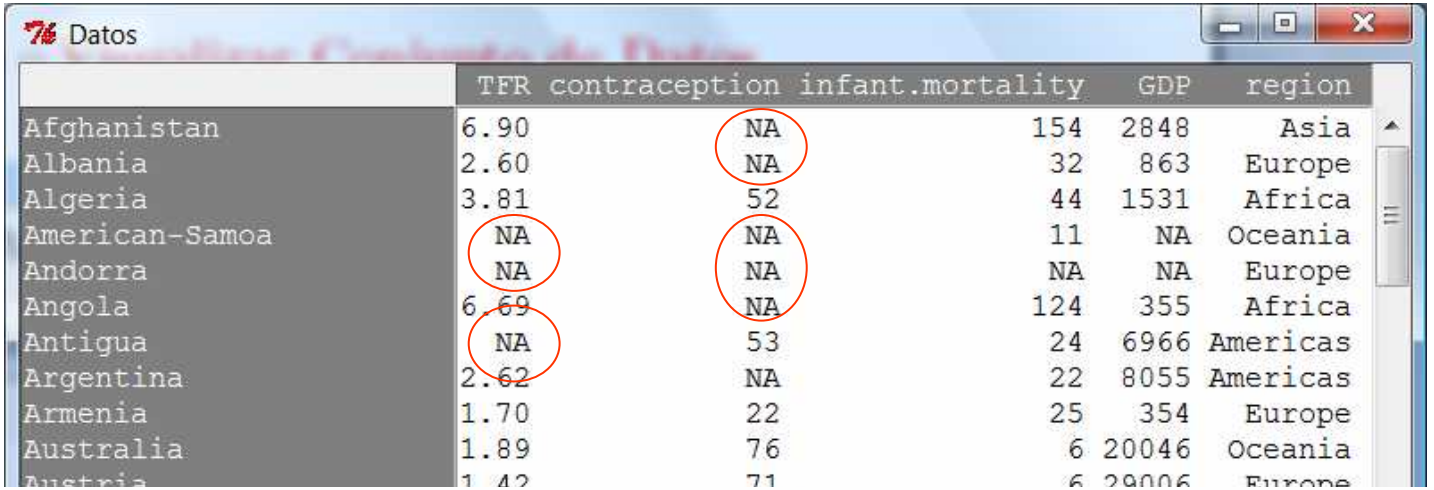




	TFR	contraception	infant.mortality	GDP	region
Afghanistan	6.90	NA	154	2848	Asia
Albania	2.60	NA	32	863	Europe
Algeria	3.81	52	44	1531	Africa
American Samoa	NA	NA	11	NA	Oceania
	NA	NA	NA	NA	Europe
	NA	NA	124	355	Africa
	53	24	6966	Americas	
	NA	22	8055	Americas	
	22	25	354	Europe	
	76	6	20046	Oceania	
	71	6	29006	Europe	
	17	33	321	Asia	
	62	14	12545	Americas	
	53	18	9073	Asia	
	49	78	280	Asia	
	55	9	7173	Americas	
	50	15	994	Europe	
	79	7	26582	Europe	
	47	30	2569	Americas	
	16	84	391	Africa	
	19	104	166	Asia	
	45	66	909	Americas	
	NA	13	271	Europe	
	33	56	3640	Africa	
	74	42	4510	Americas	
	NA	9	16683	Asia	
	NA	16	1518	Europe	
	8	97	165	Africa	
	9	114	205	Africa	
	NA	102	130	Asia	

La primera línea del fichero contiene los nombres de las variables:

- `TFR` (tasa de fertilidad total, expresada como el número de descendientes por mujer),
- `contraception` (porcentaje de uso de contraceptivos usados entre mujeres enparejadas),
- `infant.mortality` (tasa de mortalidad infantil por cada 1000 nacimientos),
- `GDP` (PIB, en dólares) y
- `region`.



	TFR	contraception	infant.mortality	GDP	region
Afghanistan	6.90	NA	154	2848	Asia
Albania	2.60	NA	32	863	Europe
Algeria	3.81	52	44	1531	Africa
American-Samoa	NA	NA	11	NA	Oceania
Andorra	NA	NA	NA	NA	Europe
Angola	6.69	NA	124	355	Africa
Antigua	NA	53	24	6966	Americas
Argentina	2.62	NA	22	8055	Americas
Armenia	1.70	22	25	354	Europe
Australia	1.89	76	6	20046	Oceania
Austria	1.42	71	6	29006	Europe

Algunos de los datos son *missing*. En R, es conveniente usar el código NA (representando *not available*) para codificar los datos faltantes.

Las variables TFR, contraception, infant.mortality y GDP son numéricas (variables cuantitativas); a su vez, region es nominal y contiene los nombres de las regiones.

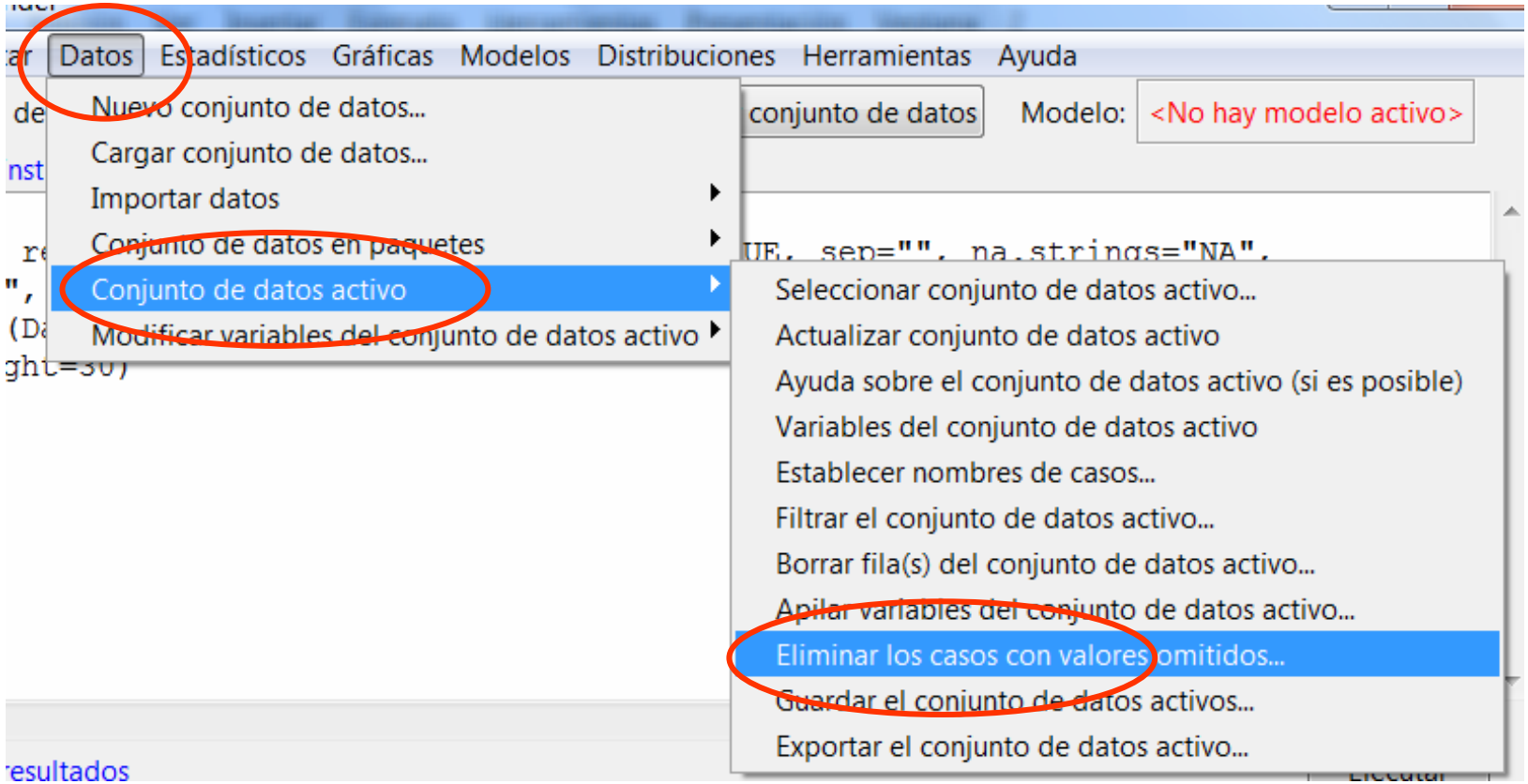
Cuando se leen los datos, R trata la variable region como un *factor*, es decir, como una variable categórica.

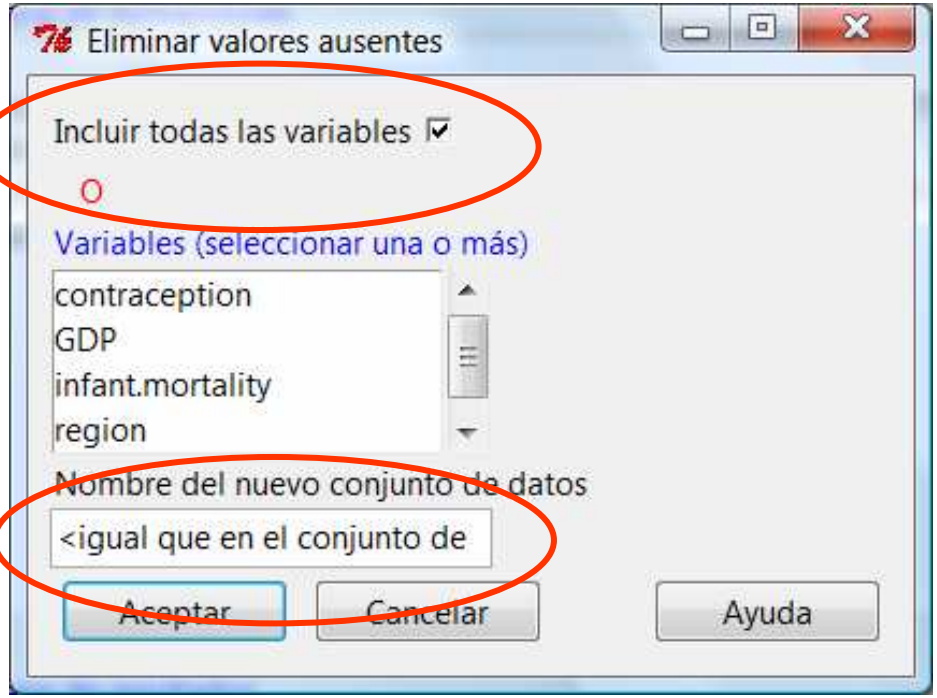
Brunei	2.70	NA	9	10003	Asia
Bulgaria	1.45	NA	16	1518	Europe
Burkina-Faso	6.57	8	97	165	Africa
Burundi	6.28	9	114	205	Africa
Cambodia	4.50	NA	102	130	Asia

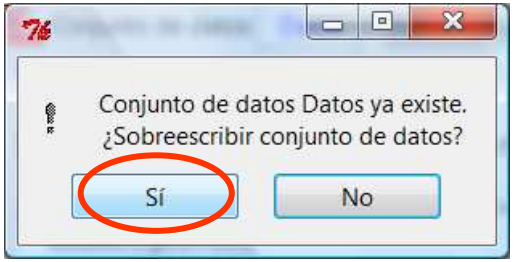
Ejemplo

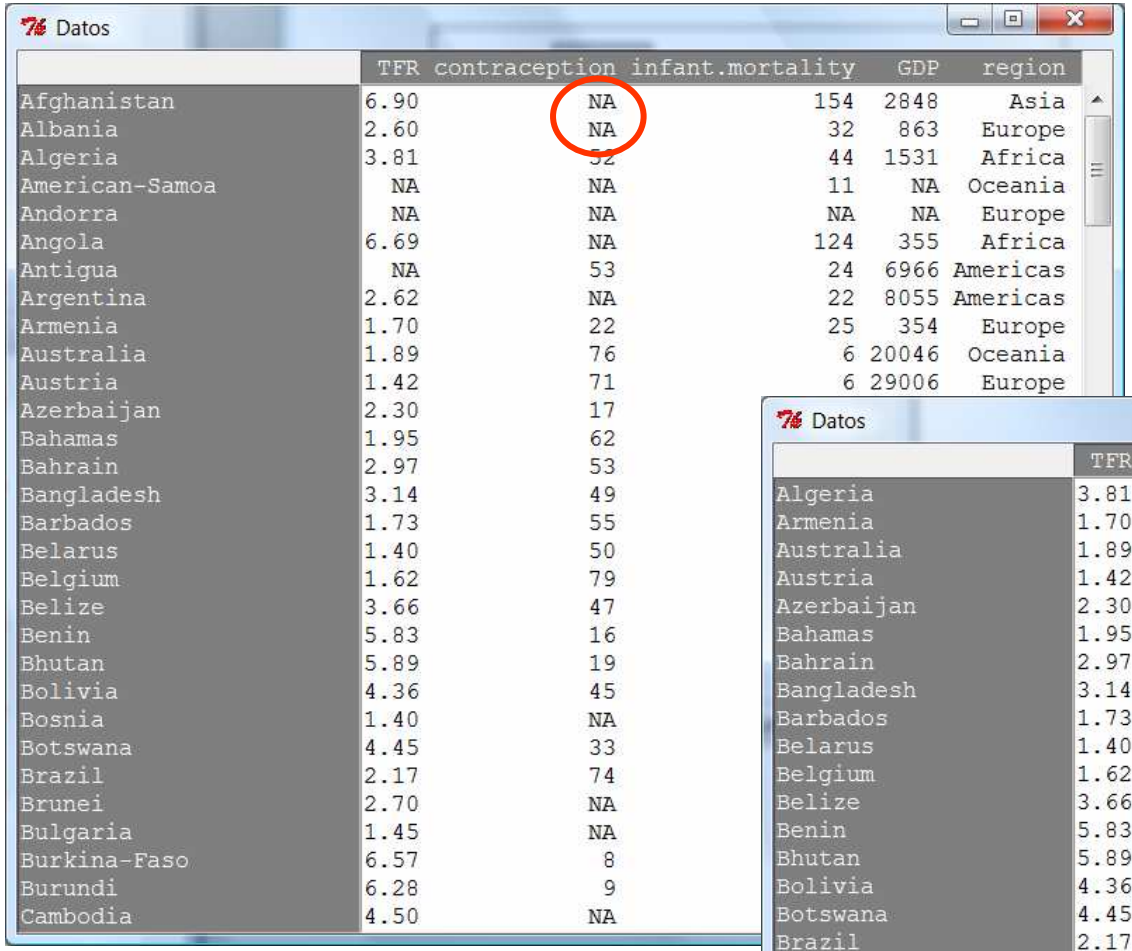
Para eliminar los datos missing del fichero de datos activo, se hace:

Datos → Conjunto de datos activo → Eliminar los casos con valores omitidos



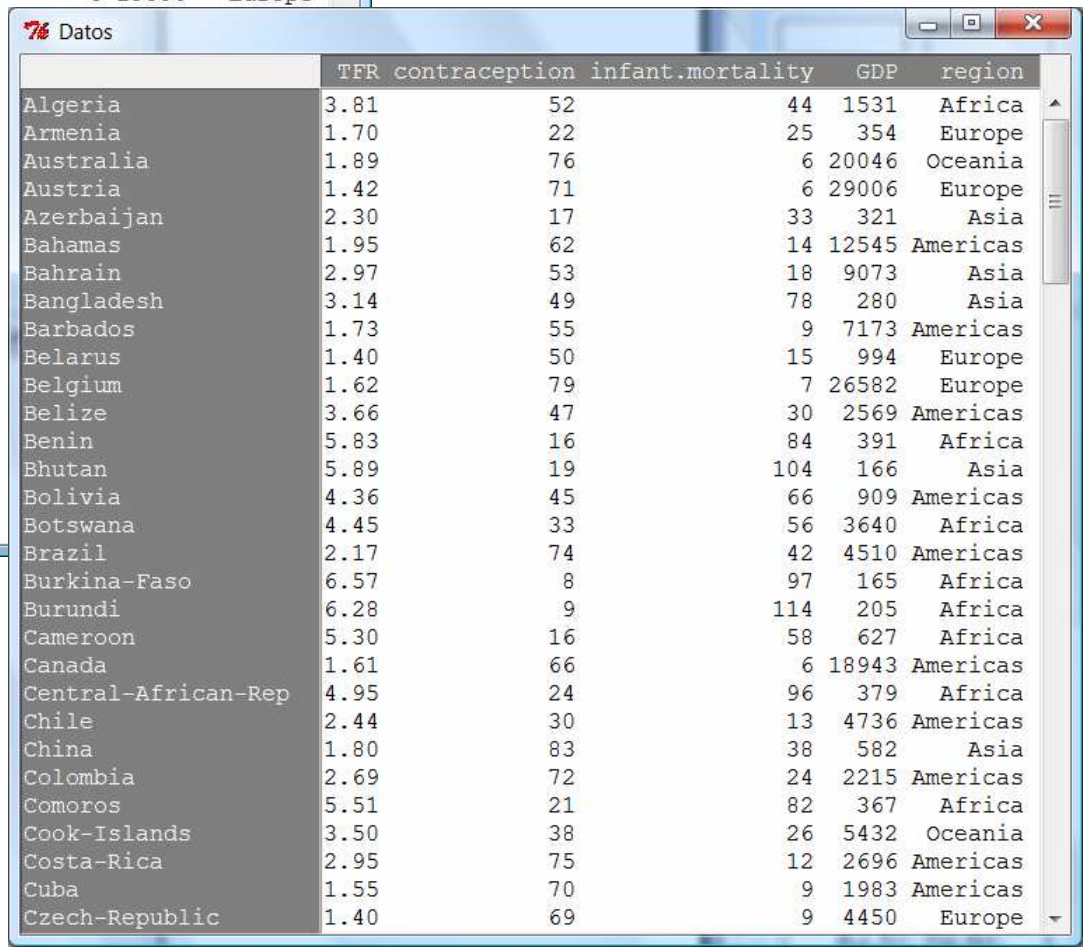

Eliminar valores ausentes
Incluir todas las variables
O
Variables (seleccionar una o más)
contraception
GDP
infant.mortality
region
Nombre del nuevo conjunto de datos
<igual que en el conjunto de
Aceptar
Cancelar
Ayuda


Conjunto de datos Datos ya existe.
¿Sobreescribir conjunto de datos?
Sí
No



Antes

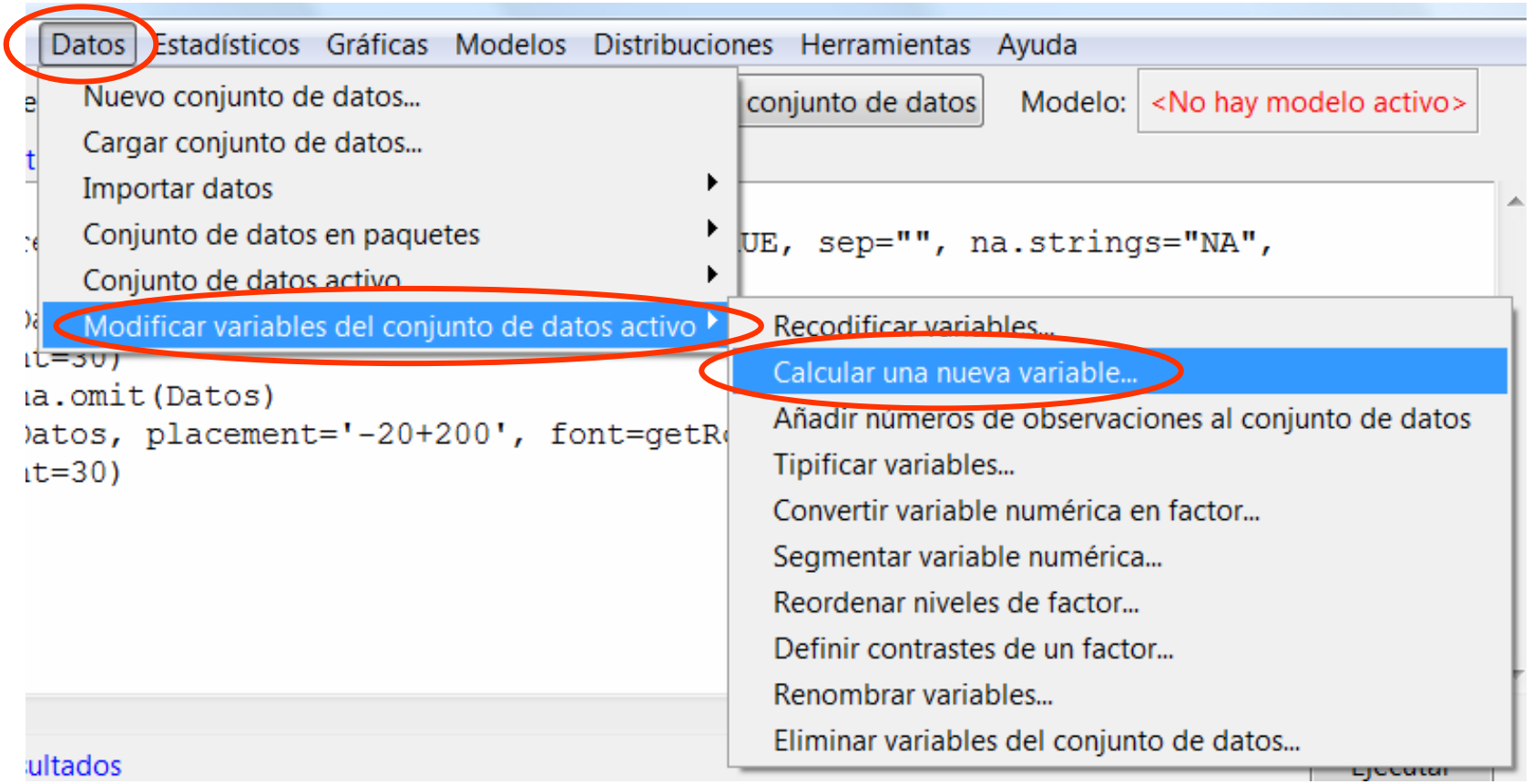
Después

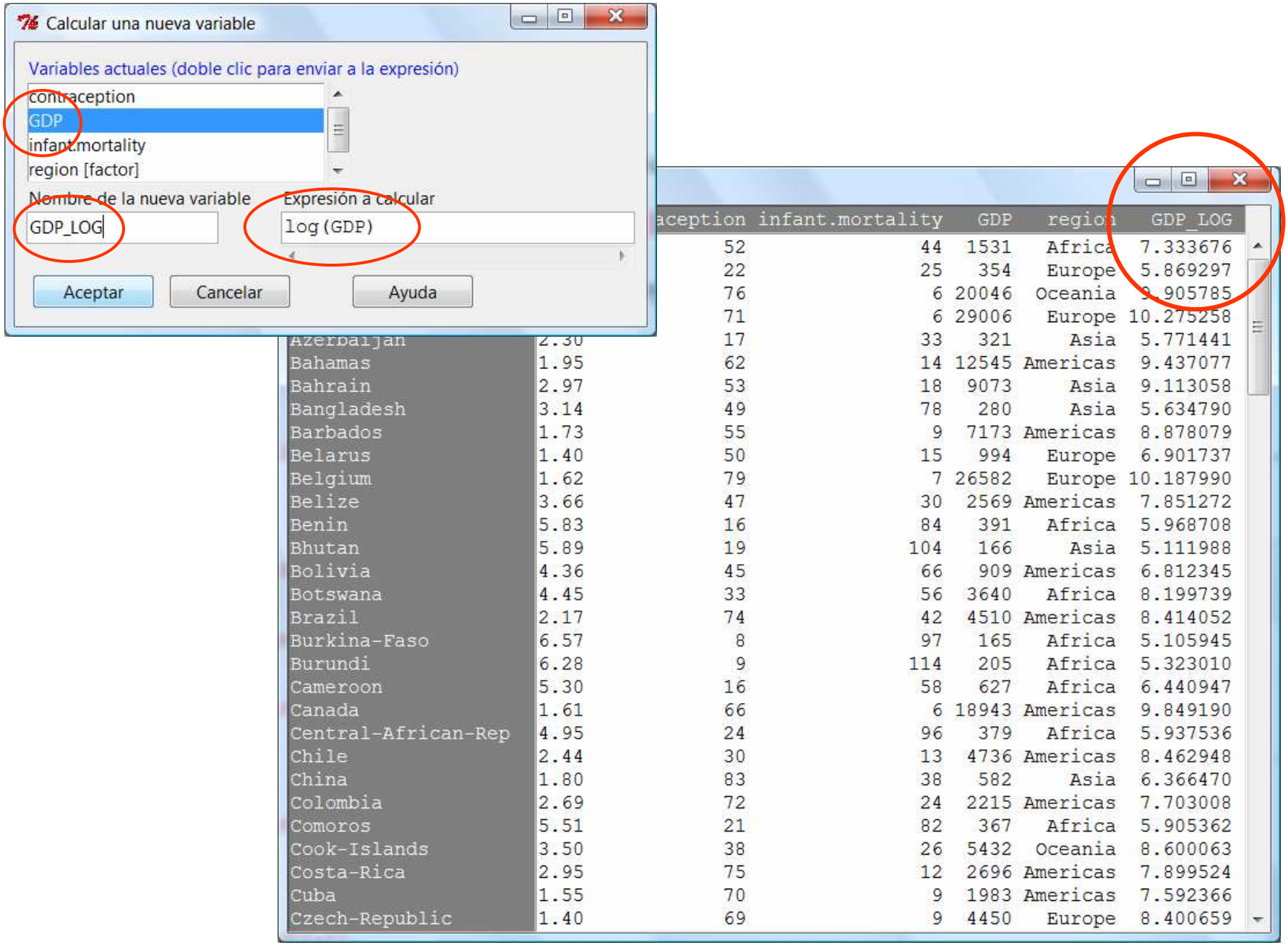


Calcular nuevas variables a partir de las existentes:

Datos → Modificar variables del conjunto activo → Calcular una nueva variable

Calculamos, por ejemplo, el logaritmo de la variable GPD:




Calcular una nueva variable
Variables actuales (doble clic para enviar a la expresión)
contraception
GDP
infant.mortality
region [factor]
Nombre de la nueva variable
GDP_LOG
Expresión a calcular
log(GDP)
Aceptar
Cancelar
Ayuda
aception infant.mortality GDP regio GDP_LOG
52 44 1531 Africa 7.333676
22 25 354 Europe 5.869297
76 6 20046 Oceania 9.905785
71 6 29006 Europe 10.275258
17 33 321 Asia 5.771441
Azerbaijan 2.30
Bahamas 1.95 62 14 12545 Americas 9.437077
Bahrain 2.97 53 18 9073 Asia 9.113058
Bangladesh 3.14 49 78 280 Asia 5.634790
Barbados 1.73 55 9 7173 Americas 8.878079
Belarus 1.40 50 15 994 Europe 6.901737
Belgium 1.62 79 7 26582 Europe 10.187990
Belize 3.66 47 30 2569 Americas 7.851272
Benin 5.83 16 84 391 Africa 5.968708
Bhutan 5.89 19 104 166 Asia 5.111988
Bolivia 4.36 45 66 909 Americas 6.812345
Botswana 4.45 33 56 3640 Africa 8.199739
Brazil 2.17 74 42 4510 Americas 8.414052
Burkina-Faso 6.57 8 97 165 Africa 5.105945
Burundi 6.28 9 114 205 Africa 5.323010
Cameroon 5.30 16 58 627 Africa 6.440947
Canada 1.61 66 6 18943 Americas 9.849190
Central-African-Rep 4.95 24 96 379 Africa 5.937536
Chile 2.44 30 13 4736 Americas 8.462948
China 1.80 83 38 582 Asia 6.366470
Colombia 2.69 72 24 2215 Americas 7.703008
Comoros 5.51 21 82 367 Africa 5.905362
Cook-Islands 3.50 38 26 5432 Oceania 8.600063
Costa-Rica 2.95 75 12 2696 Americas 7.899524
Cuba 1.55 70 9 1983 Americas 7.592366
Czech-Republic 1.40 69 9 4450 Europe 8.400659