



Estadística Básica
con
R y R–Commander

2ª Edición Revisada
(Versión Marzo 2013)

Autores:

A. J. Arriaza Gómez
F. Fernández Palacín
M. A. López Sánchez
M. Muñoz Márquez
S. Pérez Plaza
A. Sánchez Navas



UCA | Universidad
de Cádiz

Servicio de Publicaciones

Copyright ©2008, 2013 Universidad de Cádiz. Se concede permiso para copiar, distribuir y/o modificar este documento bajo los términos de la Licencia de Documentación Libre de GNU, Versión 1.3 o cualquier otra versión posterior publicada por la Free Software Foundation. Una traducción de la licencia está incluida en la sección titulada "Licencia de Documentación Libre de GNU".

Copyright ©2008, 2013 Universidad de Cádiz. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation. A copy of the license is included in the section entitled "GNU Free Documentation License".

Edita: Servicio de Publicaciones de la Universidad de Cádiz
C/ Dr. Marañón, 3
11002 Cádiz

<http://www.uca.es/publicaciones>

ISBN:

Depósito legal:



Capítulo 2

Organización y manipulación de datos con R

En este capítulo se aprenderá a organizar los datos, reconocer su naturaleza, cargarlos en Rcmdr y editarlos al objeto de adaptarlos a diferentes necesidades. En primer lugar se verán algunas cuestiones conceptuales y de organización, a continuación se describirá como introducir datos en un editor, cargarlos, e importarlos y exportarlos desde o hacia distintos formatos, y por último, se verá como manipularlos para adaptarlos a otras necesidades, realizando operaciones sobre el archivo de datos o editando algunas variables del mismo.

1. La matriz de datos

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.

Los datos constituyen pues la materia prima de la estadística, pudiéndose establecer distintas clasificaciones en función de la forma en que

éstos vengan dados. Se obtienen datos al realizar cualquier tipo de prueba, experimento, valoración, medición, observación, . . . , dependiendo de la naturaleza de los mismos y del método empleado para su obtención. Una vez obtenidos los datos por los procedimientos que se consideren pertinentes, pueden generarse nuevos datos mediante transformación y/o combinación de las variables originales. Al conjunto de datos convenientemente organizados se le llamará *modelo 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, pudiendo tratarse de caracteres nominales, dicotómicos o politómicos, presencias–ausencias, ordenaciones, conteos, escalas de intervalo, razones, . . . ; también se podrían tener variables compuestas como ratios, densidades, . . . En ocasiones se añade una columna que se suele colocar en primer lugar y que asigna un nombre a cada individuo; dicha columna recibe el nombre de *variable etiqueta*.

De forma general, se puede considerar que una matriz de datos tiene dos dimensiones fundamentales, la **dimensión física**, constituida por los individuos, y la **dimensión de estado**, que contiene a las variables y que proporciona información de los individuos en relación a los aspectos que se consideren relevantes para dar respuesta a los objetivos del estudio.

Físicamente, la estructura de una matriz de datos se corresponde con el esquema de una base de datos o una hoja de cálculo. Al igual que pasa con los editores de los programas de tratamiento de datos, las dos dimensiones de una pantalla se acomodan perfectamente al tanden individuo–variable. Si se consideran los individuos identifica-

dos por los términos $I_1, I_2, \dots, I_n$ y los caracteres por $C_1, C_2, \dots, C_k$, la casilla x_{ij} representa el comportamiento del individuo I_i respecto al carácter C_j . En la figura se muestra la matriz de datos del fichero *Iris* del paquete **datasets** de R.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3	1.4	0.1	setosa
14	4.3	3	1.1	0.1	setosa
15	5.8	4	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa

En R este tipo de estructura de datos constituye un objeto de tipo *data.frame*. La mayoría de los procedimientos estadísticos se aplican sobre objetos del tipo *data.frame*.

1.1. Anomalías de la matriz de datos

Hay veces en que por distintos motivos la matriz de datos presenta casillas vacías, ello se debe a que no se ha podido medir un dato o a que se ha perdido la observación. En otras ocasiones un dato presente en la matriz ha sido *depurado* por presentar algún tipo de anomalía, como haber sido mal medido, mal transcrito a la matriz de datos, pertenecer a un colectivo distinto del que se está analizando, etc... La identificación de estos elementos anómalos se realiza mediante un proceso de detección de inconsistencias o de evaluación de valores extremos, muy grandes o muy pequeños, que determinará si razonablemente pueden pertenecer al colectivo bajo estudio. A veces se sustituye el valor depurado de un individuo por uno que sea congruente con el resto de caracteres del mismo, mediante técnicas que se conocen como de *imputación*. Los huecos que definitivamente queden en la matriz se referirán como *valores omitidos* o, más comunmente, como *valores missing*. En R estos valores deben ser identificados con los caracteres NA (Not Available). En función del tipo de análisis que se esté realizando, el procedimiento desestimaré sólo el dato faltante o todo el registro (fila) completo.



Figura 2.1: Esquema de cantidad de información

1.2. Naturaleza de los caracteres: Atributos y Variables

Respecto a la *cantidad de información* que porta cada tipo de carácter, se puede considerar que los caracteres nominales son los más “pobres”, puesto que ni siquiera poseen orden, mientras que los más ricos serían las escalas de intervalos y las razones, que tienen orden, son cuantitativas y en el caso de las razones el cero lo es en términos absolutos, es decir, el 0 representa la ausencia de la característica. En posiciones intermedias se situarían el resto, en el orden en que se han introducido en la figura 2.1.

Ejemplo 2.1

El caso más evidente para apreciar las diferencias entre las escalas de intervalo y las razones o escalas de cociente, lo ofrece el termómetro. Un termómetro genera una variable de escala de intervalo, porque la diferencia real entre 2 y 3 grados es la misma que entre 40 y 41 grados, pero no se puede decir que cuando el termómetro marca 30 grados hace el doble de calor que cuando marca 15.

Por otra parte, muchas magnitudes físicas, como el peso, la longitud o la intensidad de corriente, son razones porque, por ejemplo en el caso del peso, un objeto de 20 kilogramos pesa el doble que otro de 10 kilogramos. Es decir existe el cero absoluto.

Como ya se ha comentado, la naturaleza del carácter condicionará su tratamiento, aunque en ningún caso hay que confundir la cantidad de información que porta con su valor intrínseco en el análisis.

En una primera instancia, se distinguirá entre los caracteres que no están ordenados y los que sí lo están, los primeros jugarán en general un rol de *atributos* mientras que los segundos habitualmente actuarán como *variables*. Los atributos tendrán la misión de establecer clases, dividiendo el colectivo global en subgrupos o categorías; por su parte, las variables caracterizarán a dichos subgrupos e intentarán establecer diferencias entre unos y otros, para lo que necesariamente se debe considerar algún tipo de métrica. Pero ello es una regla general que tiene muchas excepciones y así, en ocasiones, un carácter llamado a adoptar el papel de variable podría, mediante una operación de *punto de corte*, actuar como atributo, mientras que es factible definir una medida de asociación sobre caracteres intrínsecamente de clase que permita caracterizar a los individuos del colectivo en base a una serie de atributos.

Ejemplo 2.2

Es habitual que la edad, que es intrínsecamente una variable –medida en un soporte temporal– se emplee para dividir la población en clases dando cortes en el intervalo de tiempo, obteniéndose por ejemplo grupos de alevines, adultos y maduros de una comunidad de peces y adoptando por tanto la variable un rol de atributo.

En el extremo opuesto, hay investigaciones médicas que relacionan el tipo de patología con el sexo del paciente y con el desenlace de la enfermedad, caracteres todos ellos intrínsecamente atributos.

Las variables pueden clasificarse según su conjunto soporte. El soporte de una variable es el conjunto de todos los posibles valores que toma. Cuando el conjunto soporte es finito o numerable se habla de variable discreta. Por el contrario, cuando el conjunto soporte es no numerable, se habla de variable continua. Si la variable continua no toma valores en puntos aislados se dice absolutamente continua. Esta diferencia tendrá relevancia cuando se planteen, más adelante, estructuras de probabilidad para modelizar la población bajo estudio.

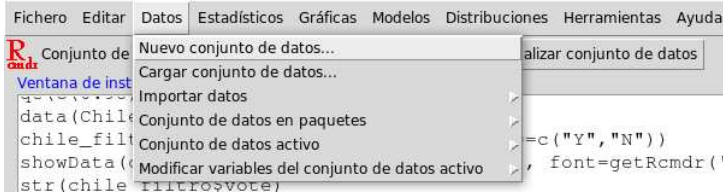


Figura 2.2: Menú Datos

Ejemplo 2.3

El número de lunares en la piel de pacientes aquejados de una cierta patología, el número de hijos de las familias de una comunidad o el número de meteoritos que surcan una cierta región estelar en periodos de tiempo determinados son variables discretas. La distancia por carretera entre las capitales de provincia peninsulares españolas, el tiempo de reacción de los corredores de una carrera de 100 metros o las longitudes de los cabellos de una persona son variables continuas.

2. Activación de datos en Rcmdr

Los ficheros de datos de R tienen extensión `rda`, sin embargo es habitual que la introducción y edición de datos se haga desde una herramienta más versátil como puede ser una hoja de cálculo, una base de datos o simplemente un editor de textos. Si el fichero tiene formato `rda` bastará con cargarlo para que esté operativo, sin embargo si el formato es cualquier otro, será necesario proceder a su importación. Estas operaciones de carga e importación de datos están disponibles en la opción **Datos** dentro del menú principal de Rcmdr (figura 2.2). La opción **Datos** permite realizar operaciones de carga, importación y edición de ficheros de datos. Muy interesante es la opción **Conjuntos de datos en paquetes**, en la que Rcmdr ofrece una serie de carpetas conteniendo ficheros de datos, muchos de ellos de carácter histórico y que se usarán frecuentemente para documentar los procedimientos en este manual.

Cuando se desea cargar uno de los ficheros incluidos en los pa-

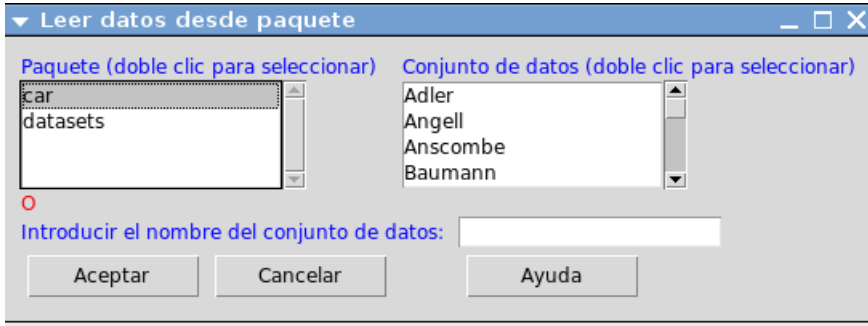


Figura 2.3: Ventana de selección de datos en paquetes adjuntos

quetes, se selecciona: **Datos**→**Conjunto de datos en paquetes**→**Leer conjunto de datos desde paquete adjunto...** Para ver el listado de ficheros empaquetados, se elige en el menú: **Datos**→**Conjunto de datos en paquete**→**Lista de conjuntos de datos en paquetes...** (figura 2.3). Estos ficheros empaquetados tienen la extensión de los ficheros de datos de R. Si se quisiera cargar un fichero de datos `rda` desde un directorio del ordenador bastaría seleccionar: **Datos**→**Cargar conjunto de datos...**; en el navegador de archivos que se abrirá sólo aparecerán los ficheros con extensión `rda`.

Para importar un fichero de datos con formato: “`dat`”, “`txt`”, “`xls`”, “`sav`”,..., se usa la secuencia **Datos**→**Importar datos...** y se elige el formato que corresponda. Una de las operaciones más habituales es importar desde una hoja de cálculo o un archivo de texto. En estos casos, cuando el fichero no es muy grande, lo más fácil es marcar el conjunto de datos y copiarlo al portapapeles, después se marca la opción del menú: **Datos**→**Importar datos**→ **desde archivo de texto, portapapeles o URL...**, y se seleccionan las opciones que interesen: el nombre del fichero, si los nombres de las variables están en la primera fila del mismo, el origen de los datos, el separador de campos y el decimal, como se puede ver en el diálogo de la figura 2.4.

Otra de las posibilidades que ofrece Rcmdr es la de importar datos directamente desde una URL. Ya se ha visto en el ítem anterior como



Figura 2.4: Opciones de importación de datos

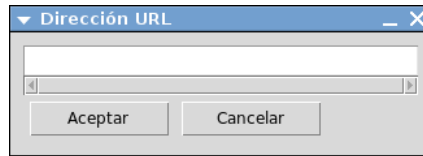


Figura 2.5: Ventana de importación de datos desde una URL

la ventana de importación daba la posibilidad de especificar, dentro de **Localización del archivo de datos**, una dirección URL. Al hacerlo, emerge la ventana de la figura 2.5 en la que habría que escribir la dirección.

Cuando se cargan o se importan datos, R crea un objeto del tipo *data.frame*, con una estructura que contiene además de la propia matriz de datos, el número de individuos (filas), el número de variables (columnas) y los nombres y tipos de las variables, entre otros. Como ya se indicó en un epígrafe anterior, a las casillas vacías de la matriz se les asigna el código NA.

Por último, un fichero con formato “rda” es exportable a un fichero de texto mediante la secuencia: Datos→Conjunto de datos activo→

Exportar el conjunto de datos activo...

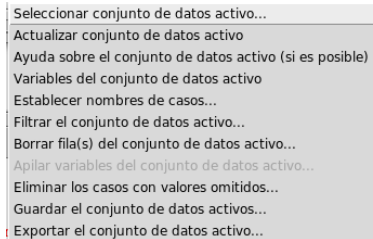
3. Manipulación de datos con Rcmdr

R tiene implementada una gran cantidad de funciones que actúan sobre la matriz de datos o sobre una parte de la misma. En Rcmdr estas funciones están incluidas en el menú **Datos** en los submenús **Conjunto de datos activo** y **Modificar variables del conjunto de datos activo**. Se verá en esta sección como operar con dichas funciones.

3.1. Edición del conjunto de datos activo

Una vez cargado el fichero de datos en Rcmdr es posible que haya que modificar, el fichero completo o algunas de las variables. En este epígrafe se aprenderá a manipular el fichero completo. Las opciones que ofrece Rcmdr se recogen en **Datos**→**Conjunto de datos activo**. Algunas de estas opciones son evidentes e incluso

como ocurre con **Seleccionar conjunto de datos activo...**, puede accederse a ellas pulsando el botón con el nombre del fichero debajo del menú principal. Se describirán someramente las más destacables.



3.1.1. Solicitar ayuda sobre el conjunto de datos activo

Esta opción despliega la ficha que documenta el fichero de datos activo, si éste está “empaquetado”. El resultado es que R aplicará la función **help** sobre el fichero, como se puede ver en la ventana de instrucciones. Esta función es accesible también a partir del símbolo “?”.

Ejemplo 2.4

Para solicitar ayuda sobre el fichero Chile se puede usar la opción de menú, ejecutar la instrucción `help(Chile)` o bien ejecutar `?Chile`.

3.1.2. Etiquetar los individuos

Se pueden seleccionar una de las variables como etiqueta del conjunto de datos en la opción **Establecer nombres de casos...**

3.1.3. Filtrado de casos

Una de las operaciones que suelen realizarse con el conjunto de datos es filtrarlo para obtener los individuos que cumplan una cierta condición, por ejemplo que sean mayores o menores de una cierta edad, que sean hombres, etc.

Ejemplo 2.5

Para filtrar el fichero **Chile** del paquete **car** y quedarse con los individuos que sean mayores o iguales a 50 años, se selecciona la opción **Filtrar el conjunto de datos activo...** con lo que se accede a la ventana de la figura 2.6.

En **Expresión de selección** se escribe **age>=50** y, si no se quiere que sobrescriba el fichero original, en **Nombre del nuevo conjunto de datos** se puede escribir **Chile50**, con lo que se generará un nuevo fichero y se activará. Si se deseara filtrar el fichero para quedarse solo con las mujeres en **Expresión de selección** se escribe **sex=="F"**, utilizándose un doble símbolo igual y entrecomillando el carácter "F".

3.1.4. Eliminación-selección de individuos

Si se desea seleccionar un conjunto de individuos o filas del conjunto de datos se puede usar la opción **Borrar fila(s) del conjunto de datos activo...** e indicar en el campo que se ofrece los índices que se desea borrar. Obsérvese que en R, para seleccionar un rango de números se escribe el primero y el último separados por el símbolo ":". Por ejemplo, para seleccionar los 100 primeros números naturales se escribiría "1:100".

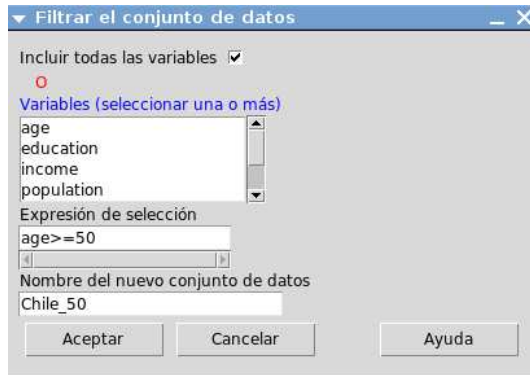


Figura 2.6: Ventana de selección de filtrado de datos

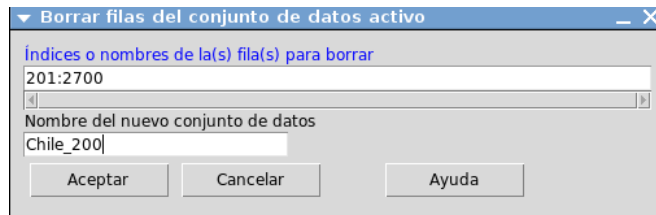


Figura 2.7: Ventana de selección de filtrado de datos

Ejemplo 2.6

*Para obtener los 200 primeros casos del fichero Chile, en **Borrar fila(s) del conjunto de datos activo...** se escribe lo que se ve en la ventana de la figura 2.7*

Es decir, se eliminan a partir del caso 201. Si se desea indicar varios subconjuntos para borrar, basta con separarlos por comas.

3.1.5. Apilar variables

Se trata de generar un nuevo fichero de datos que contendrá dos columnas, una variable unión o apilamiento de dos variables del fichero original y un factor que identifica los elementos de cada variable.

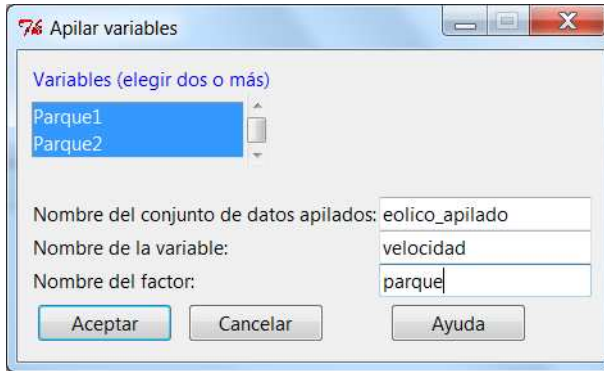


Figura 2.8: Ventana de apilamiento de datos

Ejemplo 2.7

Cargamos el fichero *Bfox* del paquete *car* que da información sobre los sueldos de hombres y mujeres en una serie de años. Para apilar las variables *menwage* y *womwages* se accede a la opción del menú **Apilar variables del conjunto de datos activo**. En la ventana de diálogo que emerge, figura 2.8, se especifican las variables, el nombre del nuevo fichero, el nombre de la variable apilada y el nombre del factor, cuyas dos categorías se corresponden con los nombres de las variables apiladas.

3.1.6. Selección de casos completos

En ocasiones se desean eliminar del conjunto de datos a aquellos casos–filas que no tengan todos sus valores o, lo que es lo mismo, que contengan valores *NA*. Para ello reurremos a la opción **Eliminar los casos con valores omitidos**.

3.1.7. Guardar conjunto de datos

Si se desea guardar el conjunto de datos se utiliza la opción del menú **Guardar conjunto de datos activo...**, en la ventana de diálogo se debe especificar el directorio y el nombre del fichero, si es que se

desea cambiar el que se le había dado en la sesión de trabajo.

3.1.8. Exportar fichero de datos

Para exportar el conjunto de datos activo se usa la opción del menú **Exportar el conjunto de datos activo**, en la ventana de diálogo emergente habrá que establecer las opciones que correspondan y en la siguiente ventana el directorio, nombre y la extensión. El archivo se exportará en formato texto, con extensión *dat*, *txt*, *csv*, ...

3.2. Estructura y referenciación de los datos en R

Aunque muchas de las operaciones básicas con la matriz de datos o con algunas de las variables pueden realizarse con las opciones de menu de Rcmdr, en esta sección se aprenderá a identificar variables y, de forma general, a seleccionar subconjuntos del `data.frame` a partir del sistema de indexado de R.

3.2.1. Identificación de variables

Ya se ha visto que el formato de datos de R está determinado por un objeto de tipo `data.frame`. El `data.frame` es una estructura de datos rectangular con filas que son individuos y columnas que son variables. Cuando se quiera hacer referencia a una de las variables del `data.frame` desde el editor de entrada sólo se debe escribir el nombre del fichero, un símbolo de dólar y el nombre de la variable.

Ejemplo 2.8

Si se quisiera hacer referencia a la variable `Petal.Length` del fichero `iris` del paquete `datasets`, se escribe `iris$Petal.Length`.

3.2.2. Sistema de indexación de R

Como todo lenguaje de programación R tiene un sistema de indexación que permite identificar u operar con elementos simples o subcon-

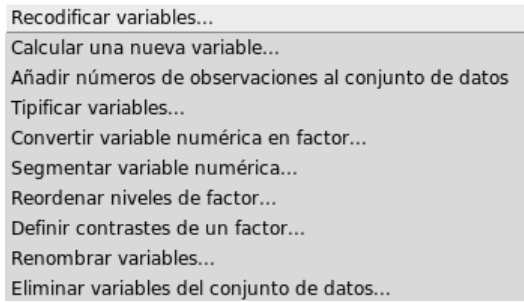


Figura 2.9: Menú para modificar variables

juntos de una estructura compleja, como por ejemplo un `data.frame`.

Ejemplo 2.9

Considérese de nuevo el fichero `iris` e imagínese que se quiere seleccionar el valor que toma el quinto individuo en la variable `Petal.Length`, que es la tercera columna del `data.frame`. En definitiva 5ª fila, 3ª columna. Para ello se introduce la instrucción `iris[5,3]`. A partir de aquí se pueden realizar distintas selecciones, como por ejemplo:

- Con `iris[5,]` se obtendría la 5ª fila completa
- Con `iris[,3]` se obtendría la tercera variable, es decir, la tercera columna
- Con `iris[1:100,]` se tendrían los 100 primeros individuos del fichero, o lo que es lo mismo las 100 primeras filas
- Con `iris[,1:4]` se tendrían las 4 primeras variables de todos los individuos
- Con `iris[-c(1:100),]` se obtienen todo el `data.frame` a excepción de las 100 primeras filas

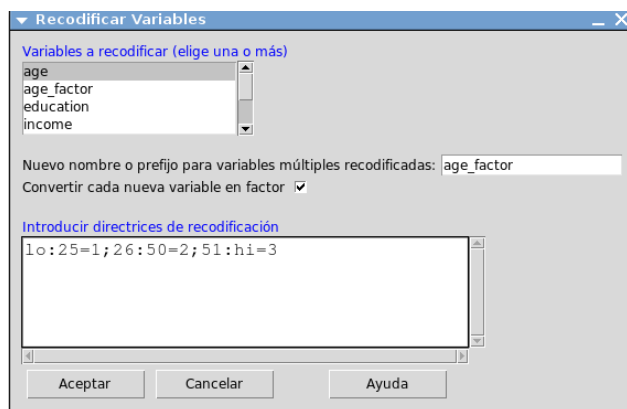


Figura 2.10: Ventana de recodificación

3.3. Modificación de las variables del conjunto de datos activo

En este epígrafe se aprenderá a modificar variables o a generar variables nuevas a partir de las originales. Para editar las variables se selecciona **Datos**→**Modificar variables del conjunto de datos activo**, con lo que se accede al menú de la figura 2.9

3.3.1. Recodificar variables

La recodificación se usa normalmente para construir un nuevo factor a partir de la combinación de valores de una variable numérica o de un factor del fichero de datos.

Ejemplo 2.10

A partir de la variable *age* del fichero *Chile* de *car* se va a obtener un factor que tendrá tres niveles: 1=hasta 25 años, 2=entre 26 y 50, 3=más de 50. Para ello en el cuadro de diálogo de *Recodificar variables*, figura 2.10, se selecciona la variable *age*, se especifican las condiciones y se da nombre al nuevo factor, dicho factor se añadirá al conjunto de datos, *age_factor*.

