Práctica 1

Introducción a R-Commander. Almacenamiento y manejo de datos.

1.1 Introducción

R es un entorno y lenguaje de programación enfocado, principalmente, al análisis estadístico de datos. Es una implementación de software libre que se encuentra disponible para los sistemas operativos Windows, Linux, Unix y Macintosh.

R proporciona una gran variedad de herramientas estadísticas para el análisis de datos y para la construcción de gráficas que permitan resumir y entender fácilmente la información contenida en los mismos. Al ser un lenguaje de programación, permite a los usuarios el desarrollo de sus propias funciones.

Sin embargo el manejo de R no es sencillo: se realiza mediante la introducción de comandos en una consola y estos comandos pueden llegar a ser muy complicados. Por este motivo, existen distintas interfaces gráficas (GUI) que simplifican su manejo. Una de estas interfaces es R-Commander, que será la interface que usaremos en las prácticas de este curso.

Para simplificar la instalación de R-Commander, así como de los diferentes paquetes de R que usaremos en las prácticas, instalaremos el paquete R-UCA. Es un proyecto de la Universidad de Cádiz que permite instalar en un único paso R, R-Commander y algunos de los paquetes de uso frecuente. Para la realización de las prácticas usaremos la versión 3.3.1, que puede descargarse haciendo uso del enlace `http://knuth.uca.es/R/R-UCA-3.3.1.exe`

Cuando ejecutemos R, veremos dos ventanas. Una de ellas, más pequeña, es la consola de R. La otra ventana corresponde a R-Commander y es con la que trabajaremos. En ésta podemos distinguir cuatro partes:

- Un menú desplegable en el que aparecen los menús `Fichero, Editar, Datos, Estadísticos,...`.
- La ventana *R-Script*. En esta ventana se mostrarán las instrucciones que

se ejecutan al seleccionar cada una de las opciones de los menús: cuando seleccionamos una de estas opciones R-Commander escribe en esta ventana el código necesario para ejecutar las instrucciones correspondientes y se lo envía a R para que lo ejecute.

Podemos también escribir instrucciones en esta ventana y ejecutarlas. Para ello basta con seleccionarlas, arrastrando con el cursor sobre las instrucciones al mismo tiempo que se mantiene pulsado el botón izquierdo del ratón, y pulsar el boton ejecutar.

- La ventana de salida: es el espacio en el que se muestran los resultados proporcionados por el programa.
- La ventana de Mensajes, utilizada por R-Commander para transmitirnos mensajes de información así como los errores.

1.2 Introducción de datos

A la hora de introducir datos usaremos dos de las posibles alternativas que nos ofrece el programa: cargarlos desde un archivo externo de R y la introducción de los mismos de forma manual mediante el editor de datos.

Para cargar datos, cuando se encuentran en un archivo externo, usaremos el menú `Datos`→`Cargar conjunto de datos...`.

Para la introducción manual de datos seleccionaremos el menú `Datos`→`Nuevo conjunto de datos`. Se abre entonces una ventana emergente en la que tenemos que introducir el nombre que queremos dar al conjunto de datos. Tras hacer esto y pusar el botón `Aceptar` se abrirá la ventana del editor de datos. En dicha ventana podemos encontrar los menús:

- `Fichero:` permite salir y guardar los datos introducidos.
- `Editar:` permite borrar y añadir filas y/o columnas, así como copiar y pegar celdas.
- Ayuda: proporciona ayuda sobre el editor de datos.

Cada fila representa el caso o individuo sobre el que se observan los valores. Las columnas corresponden a las distintas magnitudes (variables) que se miden en cada individuo. Pinchando con el cursor en cada celda podemos cambiar su contenido. Nótese que, una vez que hayamos introducido algún dato en alguna de las celdas, podremos movernos por las restantes celdas del editor usando los cursores. Es importante no presionar la tecla 'enter' hasta que terminemos de introducir los datos porque, en tal caso, se cerraría el editor. También **es importante cerrar el editor antes de volver a trabajar con R-Commander** porque en caso contrario puede

producirse un mal funcionamiento del mismo.

Ejemplo: Introducir los siguientes datos, con el nombre `Datos_personas`, usando el editor de datos.

Nombre	Color_de_pelo	Altura
Luis	Moreno	1.75
María	Rubio	1.64
Pedro	Rubio	1.72
Alberto	Castaño	1.78
Isabel	Castaño	1.67
Manuel	Moreno	1.71
María	Castaño	1.73

Para ello seleccionamos el menú `Datos`→`Nuevo conjunto de datos`. En la ventana emergente escribimos el nombre del conjunto de datos y pulsamos el botón `Aceptar`. Nótese que son siete casos en los que se miden tres magnitudes: nombre, color de pelo y altura. Por tanto, una vez abierto el editor, utilizando los botones `Añadir fila` y `Añadir columna` añadimos 6 filas y 2 columnas más. A continuación introducimos los datos y, cuando hayamos terminado, pulsamos el botón `Aceptar`. `Nombre`, `Color_de_pelo` y `Altura` son los nombres de las variables bajo estudio y debemos escribirlos en la *fila 0*, sustituyendo a los nombres por defecto de las variables (`V1, V2 y V3`).

Nótese que las variables en las que introducimos valores numericos son consideradas por el programa como variables numéricas mientras que, si introducimos cadenas de caracteres, la variable es considerada como una variable de tipo factor, esto es, una variable que contiene caractéres alfanuméricos. El tipo que se dé a una variable es importante porque, en función del tipo de variable, el programa permitirá o no realizar ciertas acciones. Por ejemplo, si una variable contiene caractéres numéricos pero es una variable de tipo factor, no podremos calcular su valor medio. En nuestro ejemplo, la variable `Color_de_pelo` es una variable de tipo factor y la variable `Altura` es una variable de tipo numérico. En prácticas posteriores veremos cómo convertir una variable de tipo numérico en tipo factor y viceversa.

Los nombres de las variables pueden contener letras y números pero siempre deben comenzar por una letra. No pueden contener caracteres reservados, como &,# o espacios. Además **R es sensible a mayúsculas** tanto en lo que se refiere a los datos como a los nombres de variables. De este modo, una variable de nombre *origen*, será distinta de una variable de nombre *Origen*.

Si una vez introducidos los datos queremos hacer alguna modificación en los mismos podemos abrir de nuevo el editor de datos pulsando el botón `Editar conjunto de datos`, en la ventana de R-Commander. Si simplemente queremos visualizarlos, podemos utilizar el botón `Visualizar conjunto de datos`.

Si ahora queremos guardar el archivo de datos creado, podemos hacerlo desde el menú `Datos`→`Conjunto de datos activo`→`Guardar el conjunto de datos activo`. En el menú desplegable `Tipo` (de archivo) seleccionaremos la opción `Archivos de datos de R`.

Aunque no trabajaremos este aspecto, utilizando las opciones del correspondiente menú de R-Commander, tambien es posible importar y exportar el conjunto de datos desde y a otros formatos como texto, Excel, y formatos de archivo de otros paquetes de software estadístico como SPSS o MINITAB.

1.3 Fusionar conjuntos de datos

Veremos en este apartado cómo fusionar datos que se encuentren guardados en dos conjuntos de datos distintos. Esto permitirá ampliar el número de observaciones que tengamos sobre una variable utilizando observaciones sobre esa variable, que se encuentren en otros conjuntos de datos, o bien añadir, a un conjunto de datos, otros datos relativos a variables que se hayan medido sobre los mismos casos.

Para ello disponemos del menú `Datos`→`Fusionar conjuntos de datos`. En la ventana emergente tendremos que indicar el nombre del nuevo conjunto de datos en el que se almacenarán los datos. A continuación, en las ventanas `Primer conjunto de datos` y `Segundo conjunto de datos` elegiremos los conjuntos de datos que se van a fusionar. En las opciones `Dirección de la fusión` podemos elegir `Fusión de filas` o `Fusionar columnas`. La primera de las opciones nos permitirá fusionar los casos (filas) de los conjuntos de datos seleccionados. La segunda opción nos permitirá fusionar las variables (columnas) de ambos conjuntos de datos.

Ejemplo: En un estudio sobre las características físicas de un grupo de personas se recogen los siguientes datos (conjunto 1):

Nombre	Peso	Altura
Pedro	78	1.75
María	53	1.62
Luisa	56	1.64
Miguel	80	1.70
Ana	59	1.70

posteriormente, el estudio se completa añadiendo también el color de pelo de las personas anteriores (conjunto 2):

Nombre	Color_pelo
Pedro	Moreno
María	Rubio
Luisa	Castaño
Miguel	Rubio
Ana	Castaño

Finalmente, se amplia el estudio con datos de cuatro personas más (conjunto 3):

Nombre	Peso	Altura	Color_pelo
Andrés	90	1.85	Rubio
María	52	1.57	Castaño
Rosa	63	1.72	Moreno
Sofía	57	1.65	Moreno

Crearemos tres conjunto de datos con el editor de R-Commander. El primer conjunto de datos, de nombre *Personas1*, con los datos del conjunto 1; el segundo conjunto de datos debe contener los datos del conjunto 2 y le pondremos como nombre *Personas1_pelo*; finalmente, al tercer conjunto de datos le pondremos como nombre *Personas2* y debe contener los datos del conjunto 3.

Una vez introducidos los conjuntos de datos, procederemos a fusionar los datos que contienen. En primer lugar fusionaremos los datos de los conjuntos 1 y 2. Para ello, seleccionamos el menú `Datos`→`Fusionar conjuntos de datos...`. En la ventana emergente indicamos el nombre del nuevo conjunto de datos que será, por ejemplo, *Datos1_completo*. A continuación, seleccionamos los conjuntos de datos a fusionar: *Personas1*, en la ventana izquierda, y *Personas1_pelo*, en la ventana derecha. Puesto que el conjunto 2 contiene los datos de una variable nueva (el color de pelo), que queremos añadir al fichero de datos, marcaremos la opción `Fusionar columnas`. Pulsamos aceptar y visualizamos el conjunto de datos.

Nótese que la columna correspondiente a los nombres aparece duplicada porque se encontraba en los dos conjuntos de datos. Eliminaremos una de las dos columnas, por ejemplo, la columna *Nombre.y*. Para ello, abrimos de nuevo el editor de datos, pinchamos con el cursor en una celda de esta columna y, en el menú editar, seleccionamos la opción `Borrar la columna actual`. Un vez hecho esto, cambiamos el nombre de la primera columna, *Nombre.x*, por *Nombre* y pulsamos en el botón `Aceptar`.

Añadiremos ahora los datos del conjunto 3. En este caso hay que añadir datos correspondientes a otros casos estudiados, por lo que repetimos los pasos que hemos dado antes para fusionar los conjuntos de datos, *Personas1* y *Personas1_pelo* pero, en este caso, los conjuntos de datos a fusionar son *Datos1_completo* y *Personas2*. Marcaremos la opción `Fusión de filas` y le daremos el nombre de *Datos_personas* al nuevo conjunto de datos. Visualizamos los datos para comprobar que todo se ha hecho correctamente.

1.4 Construcción de nuevas variables

En ocasiones en necesario construir variables nuevas a partir de otras variables de las que ya tenemos los datos introducidos en el programa. Para ello disponemos del menú `Datos`→`Modificar variables del conjunto de datos activo`→`Calcular`

`una nueva variable`. En la ventana emergente indicaremos el nombre de la nueva variable, que vamos a crear, dentro del espacio `Nombre de la nueva variable`. En la casilla `Expresión a calcular` indicaremos la fórmula de cálculo de la nueva variable, a partir de las variables existentes. Éstas pueden incorporarse a la fórmula pulsando dos veces sobre el nombre de la variable.

En la fórmula se pueden utilizar operadores y funciones matemáticas, operadores de comparación y operadores lógicos. Indicaremos, a continuación, algunos de los más habituales:

Funciones	
Raíz cuadrada: sqrt(·)	Exponencial: exp(·)
Logaritmo neperiano: log(·)	Seno: sin(·)
Coseno: cos(·)	Tangente: tan(·)

Operadores		
Aritméticos	De comparación	Logicos
Suma: +	Igualdad: ==	'Y' lógico: &
Diferencia: -	Distinto: !=	'No' lógico: !
Producto: *	Menor que: <	'O' lógico: |
División: /	Mayor que: >	
Potencia: ^	Menor o igual que: <=	
	Mayor o igual que: >=	

Seguiremos trabajando con el conjunto de datos obtenido en el apartado anterior. Vamos a construir una variable que contenga el índice de masa corporal (IMC) de cada uno de los casos. El IMC se calcula dividiendo el peso en Kg. por el cuadrado de la altura en metros. Para ello seleccionamos, en la ventana `Conjunto de datos`, el conjunto de datos deseado y, en el menú `Datos`→`Modificar variables del conjunto de datos activo`→`Calcular una nueva variable`, indicaremos *IMC* como nombre de la nueva variable . A continuación, en la `Expresión a calcular` indicaremos la misma: *Peso/Altura*^2 y pulsaremos en `Aceptar`. Finalmente, visualizamos el conjunto de datos para observar los valores que toma la nueva variable.

1.5 Filtrado de datos

En ocasiones es necesario seleccionar ciertas variables y/o casos de un conjunto de datos, de acuerdo a ciertas condiciones que deben cumplir, desechando el resto de variables y/o casos. Para llevar a cabo esta tarea, disponemos del menú `Datos`→`Conjunto de datos activo`→`Filtrar el conjunto de datos activo`. En la ventana emergente podemos indicar que, en el nuevo conjunto de datos filtrados, aparezcan todas las variables o únicamente aquellas que seleccionemos. En caso de que no tengan que aparecer todas las variables, podemos indicar qué variables vamos a necesitar. En la ventana `Expresión de selección` indicaremos la expresión o fórmula que se utilizará para seleccionar los casos. En dicha expresion

pueden aparecer nombres de variables y cualquier tipo de operador y/o fórmula. En la última ventana indicamos el nombre del nuevo conjunto de datos.

Veremos ahora algunos ejemplos, a partir del los datos en el archivo RCars. Dicho archivo contiene los datos de un estudio sobre 406 vehículos en el que se miden las siguientes variables:

- `consumo`: consumo de combustible en l/100Km.
- `motor`: cilindrada en c.c.
- `cv`: cilindrada en caballos.
- `peso`: peso del vehículo en Kg.
- `acel`: aceleración, medida como tiempo en segundos que tarda el vehículo en alcanzar los 100Km./h.
- `año`: año de fabricación.
- `origen`: lugar de fabricación del vehículo (EE.UU., Europa o Japón).
- `cilindr`: número de cilindros del vehículo.

En el archivo podemos observar también que algunos datos no aparecen y, en su lugar encontramos `NA`, que es la abreviatura de 'non available'. En esos casos lo que ha ocurrido es que no se ha podido estudiar el valor de alguna de las variables. Entonces, para no desechar la información que sí se ha podido recoger sobre el resto de las variables, se sustituye el valor no disponible por `NA`.

Cargamos el archivo *Rcars.Rdata* y lo visualizamos para conocer los valores que toman las variables. Si queremos trabajar, únicamente, con los consumos de los coches fabricados en Europa, tras seleccionar el menú `Datos`→`Conjunto de datos activos`→`Filtrar el conjunto de datos activo...` desmarcaremos la casilla `Incluir todas las variables` y seleccionamos la variable `consumo`. En la casilla `Expresión de selección` escribiremos *origen=='Europa'* y, finalmente, le daremos como nombre a nuestro nuevo conjunto de datos *Consumos_Europa*. Nótese que, al ser *Europa* una cadena de caracteres tenemos que ponerlo entre comillas. Por otra parte, dado que queremos seleccionar únicamente los datos de los casos en los que la variable *Origen* toma el valor *Europa*, hemos usado el operador de comparación '=='.

Si nuestra intención fuera generar un conjunto de datos que contuviera los datos de todas las variables, pero únicamente con los casos de los coches europeos, dejaríamos marcada la casilla `Incluir todas las variables` y no seleccionaremos ninguna variable en la ventana. En la casilla expresión de selección escribiríamos, nuevamente, `origen=='Europa'`.

En los siguientes ejemplos usaremos además el menú `Estadísticos`→`Resúmenes`→`Conjunto de datos activo`. La salida que nos proporciona esta opción incluye, para las variables numéricas del conjunto de datos activo, su valor mínimo (Min.), su valor máximo (Max.) y su valor medio (Mean). En la práctica 2 ampliaremos la información sobre esta herramienta.

Es importante, antes de realizar cada ejemplo, asegurarnos de que el conjunto de datos activo es el conjunto de datos Rcars y no cualquier otro que se encuentre en memoria. El conjunto de datos activo puede elegirse fácilmente, desde la ventana de R-Commander, pinchando con el botón izquierdo del ratón en la ventana `Conjunto de datos`, que se encuentra bajo los menús de R-Commander.

Ejemplos:

- Determinar el consumo medio de los coches japoneses que hay en la muestra.

  Seleccionamos *RCars* como conjunto de datos activo. A continuación realizamos un filtrado del conjunto de datos como sigue: desmarcamos la casilla `Incluir todas las variables` y seleccionamos la variable *consumo*. Dado que sólo queremos seleccionar los casos en los que el origen del coche es japonés, como expresión de selección ponemos *origen=='Japón '* (cuidado con las mayusculas y las tildes). Finalmente, pondremos como nombre al nuevo conjunto de datos *Consumos_Japón*. Pulsamos `Aceptar` y ya tenemos listo el nuevo conjunto de datos. Finalmente elegimos la opción `Estadísticos`→`Resúmenes`→`Conjunto de datos activo`; en la salida podemos ver que el consumo medio de los coches japoneses es de 8.051 l/100Km.

- Calcular la potencia máxima de los coches que no tienen 8 cilindros.

  Seleccionamos `RCars` como conjunto de datos activo y procedemos de manera similar al ejemplo anterior pero, en este caso, la variable a seleccionar es `cv` y la expresión de selección será `cilindr != 8`. Nótese que, al ser 8 un valor numérico no tenemos que usar comillas. Al nuevo conjunto de datos le pondremos como nombre *cv_no_8*. Finalmente, repitiendo el paso final del ejemplo anterior, encontramos que la potencia máxima de los coches que no tienen 8 cilindros es de 165cv.

Ejercicios:

- Determinar el valor máximo de la variable *acel* para los coches europeos que tienen cuatro cilindros (sol: 24.80 s.).

- Determinar la potencia media de los coches que tienen consumo no superior a 8 l/100Km (sol: 72.79cv).