Bioestadística Aplicada con R y RKTeaching

Santiago Angulo Díaz-Parreño (sangulo@ceu.es)
Euardo López Ramírez (elopez@ceu.es)
José Rojo Montijano (jrojo.eps@ceu.es)
Anselmo Romero Limón (arlimon@ceu.es)
Alfredo Sánchez Alberca (asalber@ceu.es)

Septiembre de 2014

Índice general

Práctica 1

Introducción a R y RKWard

1 Introducción

La gran potencia de cálculo alcanzada por los ordenadores ha convertido a los mismos en poderosas herramientas al servicio de todas aquellas disciplinas que, como la estadística, requieren manejar un gran volumen de datos. Actualmente, prácticamente nadie se plantea hacer un estudio estadístico serio sin la ayuda de un buen programa de análisis estadístico.

R es un potente lenguaje de programación que incluye multitud de funciones para la representación el análisis de datos. Fue desarrollado por Robert Gentleman y Ross Ihaka en la Universidad de Auckland en Nueva Zelanda, aunque actualmente es mantenido por una enorme comunidad científica en todo el mundo.

Las ventajas de R frente a otros programas habituales de análisis de datos, como pueden ser SPSS, SAS, SPlus, Matlab o Minitab, son múltiples:

- Es software libre y por tanto gratuito. Puede descargarse desde la web `http://www.r-project.org/`.
- Es multiplataforma. Existen versiones para Windows, Macintosh, Linux y otras plataformas.
- Está avalado y en constante desarrollo por una amplia comunidad científica que lo utiliza como estándar para el análisis de datos.
- Cuenta con multitud de paquetes para todo tipo de análisis estadísticos y representaciones gráficas, desde los más habituales, hasta los más novedosos y sofisticados que no incluyen otros programas. Los paquetes están organizados y documentados en un repositorio CRAN (Comprehensive R Archive Network) desde donde pueden descargarse libremente. En España hay una copia de este repositorio en la web `http://cran.es.r-project.org/`.
- Es programable, lo que permite que el usuario pueda crear fácilmente sus propias funciones o paquetes para análisis de datos específicos.
- Existen multitud de libros, manuales y tutoriales libres que permiten su aprendizaje e ilustran el análisis estadístico de datos en distintas disciplinas científicas como las matemáticas, la física, la biología, la psicología, la medicina, etc.

Por defecto el entorno de trabajo de R es en línea de comandos, lo que significa que los cálculos y los análisis se relizan mediante comandos o instrucciones que el usuario teclea en una ventana de texto. No

obstante, existen distintas interfaces gráficas de usuario que facilitan su uso, sobre todo para usuarios noveles. La interfaz gráfica que se utilizará para realizar estas prácticas será *RKWard*, desarrollada por Thomas Friedrichsmeier, junto al paquete rkTeaching especialmente desarrollado por el departamento de Matemáticas de la Universidad San Pablo CEU para la docencia de estadística.

El objetivo de esta práctica es introducir al alumno en la utilización de este programa, enseñándole a realizar las operaciones básicas más habituales de carga y manipulación de datos.

2 Instalación

2.1 Instalación de R

Linux En la distribución Debian y cualquiera de sus derivadas (Ubuntu, Kubuntu, etc.) basta con teclear en la línea de comandos

```
> sudo apt-get install r-base-html r-cran-rcmdr r-cran-rodbc r-doc-html r
  -recommended
```

Windows Descargar de http://cran.es.r-project.org/bin/windows/base/release.htm el programa de instalación de R, ejecutarlo y seguir las instrucciones de instalación.

2.2 Instalación de la interfaz gráfica RKWard y el paquete rkTeaching

La interfaz gráfica de usuario RKWard puede descargarse desde la web http://rkward.sourceforge.net/ donde se indican las instrucciones para instalarlo en cada plataforma.

Para Windows se recomienda seleccionar el paquete de instalación completa que incorpora R, las librerías gráficas de KDE y el propio RKWard.

R dispone de una gran librería de paquetes que incorporan nuevas funciones y procedimientos. En la instalación base de R vienen ya cargados los procedimientos y funciones para los análisis más comunes, pero en ocasiones, para otros análisis será necesario cargar algún paquete adicional como por ejemplo el paquete rkTeaching que incorpora un nuevo menú a RKWard con la mayoría de los análisis que se realizarán en estas prácticas.

Para instalar el paquete rk.Teaching, basta con descargarlo desde la dirección http://asalber.github.io/rkTeaching_es/, arrancar R o RKWard y, en la consola de comandos, teclear el comando

```
> setwd("ruta_a_descargas")
> install.packages("rk.Teaching",repos=NULL,dep=True)
```

La instalación de cualquier otro paquete se realiza con el mismo comando, cambiando el nombre del paquete por el deseado.

En RKWard, también puede instalarse desde la ventana de R mediante el menú `Preferencias`▸ `Configurar paquetes`. Con esto aparecerá una ventana donde se muestran los paquetes instalados localmente. Para cargar un paquete instalado localmente basta son seleccionarlo y hacer clic sobre el botón Cargar. En esa misma ventana aparece una solapa `Install/Update/Remove` que permite instalar nuevos paquetes desde un repositorio de R. Al hacer clic sobre esta solapa se abrirá una conexión a internet y aparecerá una ventana con los distintos repositorios disponibles. Normalmente seleccionaremos en más cercano geográficamente, en nuestro caso Spain(Madrid). Después aparecerá un lista de paquetes instalados y nuevos. Para instalar un paquete nuevo basta con seleccionarlo y hacer clic en el botón Aceptar. Una vez instalado localmente, podrá cargarse como se ha indicado antes.

3 Arranque

Como cualquier otra aplicación de Windows, para arrancar el programa hay que hacer clic sobre la opción correspondiente del menú `Inicio▸Programas▸RKWard`, o bien sobre el icono de escritorio

Al arrancar, aparece la ventana de bienvenida de RKWard (figura 1.1).

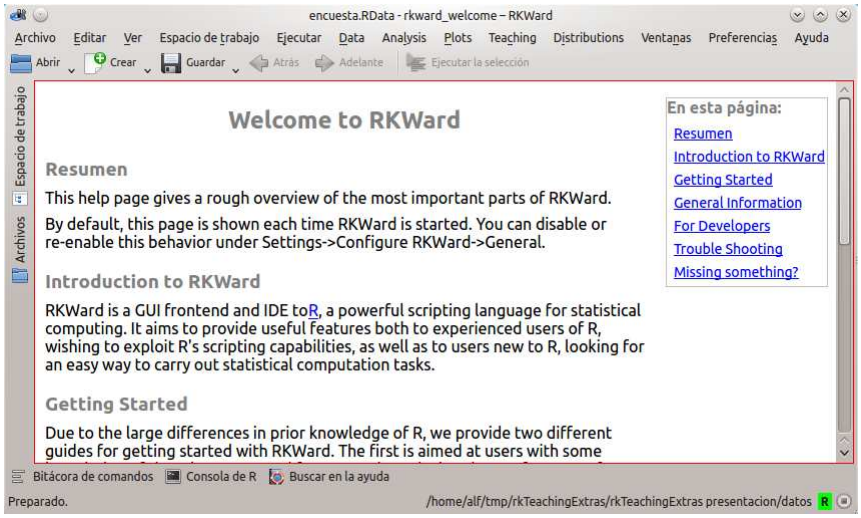


Figura 1.1 – Interfaz gráfica de usuario de RKWard.

La interfaz gráfica de usuario RKWard consta de los siguientes elementos:

- **Barra de menús**. Contiene distintos menús con operaciones que pueden realizarse con R. Si se ha instalado el paquete rkTeaching debe de aparecer el menú Teaching.
- **Barra de botones**. Contiene botones para abrir, crear y guardar conjuntos de datos, espacios de trabajo y guiones de comandos.
- **Ventana principal**. Es la ventana central donde apareceran la ventana de introducción de datos, los resultados de los comandos ejecutados o de las búsquedas realizadas.
- **Espacio de trabajo**. Es una ventana desplegable al hacer clic sobre la solapa situada en el lado izquierdo que contiene todos los elementos del espacio de trabajo de R. Entre estos elementos aparecen los paquetes cargados, los conjuntos de datos y las variables que contienen los datos de la sesión actual.
- **Bitácora de comandos** Es una solapa desplegable situada en la parte inferior donde aparece un registro de todas las acciones realizadas o comandos ejecutados en la sesión de trabajo actual. Cada vez que se seleccione un menú que lleve asociado la ejecución de algún comando, dicho comando aparecerá en esta ventana. Esto permite modificar fácilmente los parámetros del comando y volver a ejecutarlo rápidamente sin necesidad de volver al menú.
- **Consola de R** Es una solapa desplegable situada también en la parte inferior que da acceso al intérprete de comandos de R. En esta ventana pueden teclearse y ejecutarse directamente los comandos de R.

- **Buscar en la ayuda** Es una solapa desplegable situada en la parte inferior que permite hacer búsquedas sobre comandos de R o de algún paquete.
- **Mensajes**. Es la línea de texto que aparece en la parte inferior, donde se muestra información adicional sobre errores, advertencias u otra información auxiliar al ejecutar un comando, así como la ruta del espacio de trabajo activo.

4 Tipos de datos y operadores aritméticos y lógicos

En R existen distintos tipos de datos. Los más básicos son:

Numeric : Es cualquier número decimal. Se utiliza el punto como separador de decimales. Por defecto, cualquier número que se teclee tomará este tipo.

Integer : Es cualquier número entero. Para convertir un número de tipo Numeric en un entero se utiliza el comando `as.integer()`

Logical : Puede tomar cualquiera de los dos valores lógicos `TRUE` (verdadero) o `FALSE` (falso).

Character : Es cualquier cadena de caracteres alfanuméricos. Deben introducirse entre comillas. Para convertir cualquier número en una cadena de caracteres se utiliza el comando `as.character()`.

Los valores de estos tipos de datos pueden operarse utilizando distintos operadores o funciones predefinidas para cada tipo de datos. Los más habituales son:

Operadores aritméticos : + (suma), - (resta), * (producto), / (cociente), ^ (potencia).

Operadores de comparación : > (mayor), < (menor), >= (mayor o igual), <= (menor o igual), == (igual), != (distinto).

Operadores lógicos : & (conjunción y), | (disyunción o), ! (negación no).

Funciones predefinidas : `sqrt()` (raíz cuadrada), `abs()` (valor absoluto), `log()` (logarítmo neperiano), `exp()` (exponencial), `sin()` (seno), `cos()` (coseno), `tan()` (tangente).

Al evaluar las expresiones aritméticas existe un orden de prioridad entre los operadores de manera que primero se evaluan las funciones predefinidas, luego las potencias, luego los productos y cocientes, luego las sumas y restas, luego los operadores de comparación, luego las negaciones, luego las conjunciones y finalmente las disyunciones. Para forzar un orden de evaluación distinto del predefinido se pueden usar paréntesis. Por ejemplo

```
> 2^2+4/2
[1] 6
> (2^2+4)/2
[1] 4
> 2^(2+4/2)
[1] 16
> 2^(2+4)/2
[1] 32
> 2^((2+4)/2)
[1] 8
```

También es posible asignar valores a variables mediante el operador de asignación =. Una vez definidas, las variables pueden usarse en cualquier expresión aritmética o lógica. Por ejemplo,

```
> x=2
> y=x+2
> y
[1] 4
> y>x
```

```
[1] TRUE
> x>=y
[1] FALSE
> x==y-2
[1] TRUE
> x!=0 & !y<x
[1] TRUE
```

5 Introducción y manipulación de datos

Antes de realizar cualquier análisis de datos hay que introducir los datos que se quieren analizar.

5.1 Introducción de datos en línea de comandos

Existen muchas formas de introducir datos en R pero aquí sólo veremos las más habituales. La forma más rápida de introducir datos es usar la consola de R para crear un vector de datos mediante el comando `c()`. Por ejemplo, para introducir las notas de 5 alumnos se debe teclear en la consola de R

```
> nota = c(5.6,7.2,3.5,8.1,6.4)
```

Esto crea el vector nota con el que posteriormente se pueden realizar cálculos como por ejemplo la media

```
> mean(nota)
[1] 6.16
```

Otra forma habitual de introducir los datos de una muestra es crear un conjunto de datos mediante el comando `data.frame()`. Por ejemplo, para crear un conjunto de datos a partir de las notas anteriores, hay que teclear

```
> curso = data.frame(nota)
```

Esto crea una matriz de datos en la que cada columna se corresponde con una variable y cada fila con un individuo de la muestra. En el ejemplo la matriz curso sólo tendría una columna que se correspondería con las notas y 5 filas, cada una de ellas correspondiente a un alumno de la muestra. Es posible acceder a las variables de un conjunto de datos con el operador dolar `$`. Por ejemplo, para acceder a las notas hay que teclear

```
> curso$nota
[1] 5.6 7.2 3.5 8.1 6.4
```

Es fácil añadir nuevas variables a un conjunto de datos, pero siempre deben tener el mismo tamaño muestral. Por ejemplo, para añadir una nueva variable con el grupo (mañana o tarde) de los alumnos, hay que teclear

```
> curso$grupo = c("m","t","t","m","m")
```

Ahora el conjunto de datos curso tendría dos columnas, una para la nota y otra para el grupo de los alumnos. Tecleando el nombre de cualquier objeto, se muestra su información:

```
> curso
   nota   grupo
1   5.6      m
2   7.2      t
3   3.5      t
4   8.1      m
5   6.4      m
```

Cuando se introducen datos se puede utilizar el código `NA` (not available), para indicar la ausencia del dato.

Las variables definidas en cada sesión de trabajo quedan almacenas en la memoria interna de R en lo que se conoce como *espacio de trabajo*. Es posible obtener un listado de todos los objetos almacenados en el espacio de trabajo mediante los comandos `ls()`. Si se desea más información, el comando `ls.str()` además de mostrar los objetos de la memoria indica sus tipos y sus valores.

```
> ls()
[1] "curso" "nota"  "x"     "y"
> ls.str()
curso : 'data.frame':   5 obs. of  2 variables:
 $ nota : num  5.6 7.2 3.5 8.1 6.4
 $ grupo: chr  " m " " t " " t " " m " ...
nota :  num [1:5] 5.6 7.2 3.5 8.1 6.4
x :  num 2
y :  num 4
```

Para eliminar un objeto de la memoria se utiliza el comando `rm()`.

```
> ls()
[1] "curso" "nota"  "x"     "y"
> rm(x,y)
> ls()
[1] "curso" "nota"
```

5.2 Introducción de datos en RKWard

RKWard dispone de una interfaz gráfica para introducir los datos sin necesidad de saberse los comandos anteriores. Para ello hay que ir al menu `Archivo` ▸ `Nuevo` ▸ `Conjunto de datos`. Con esto aparecerá una ventana donde hay que darle un nombre al conjunto de datos y tras esto aparece la ventana de la figura 1.2 con una tabla en la que se pueden introducir los datos de la muestra. Al igual que antes, cada variable debe introducirse en una columna y cada individuo en una fila.

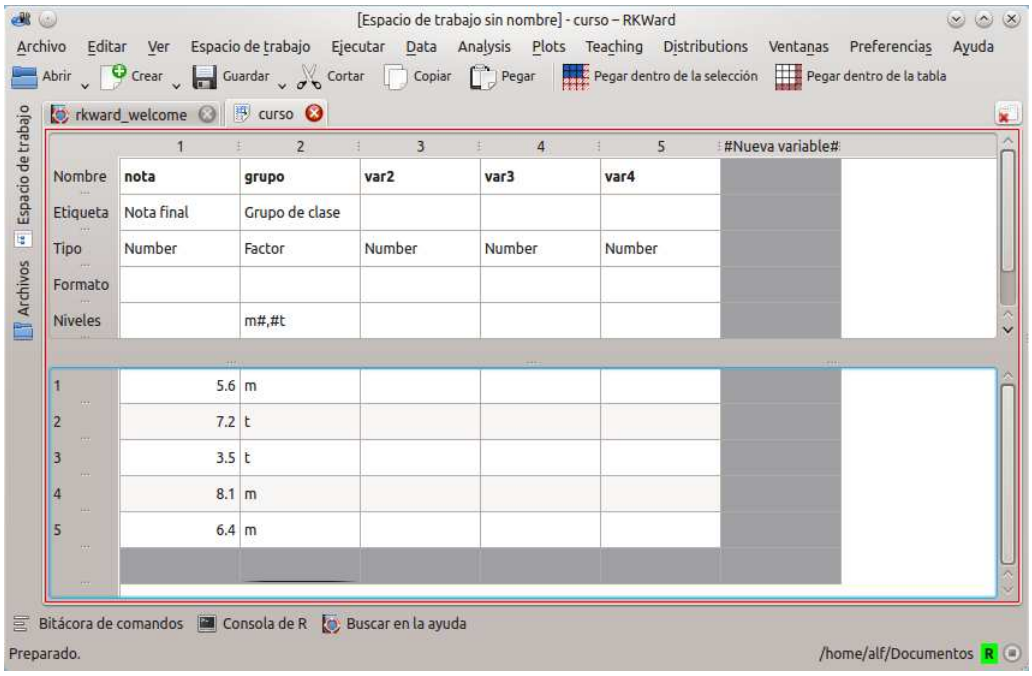


Figura 1.2 – Ventana de introducción de datos

Haciendo clic en las casillas de la cabecera cada fila es posible cambiar el nombre de la variable, ponerle una etiqueta, su tipo, su formato y los niveles en caso de tratarse de un factor o variable categórica. Los nombres de variables deben comenzar con una letra o un punto y pueden contener cualquier letra, punto, subrayado (_) o número. En particular, no se pueden utilizar espacios en blanco. Además, R es distingue entre mayúsculas y minúsculas.

Una vez definida la variable, para introducir los datos basta con teclearlos en las casillas que aparecen más abajo en la misma columna.

R permite definir más de un conjunto de datos en un mismo espacio de trabajo.

Los objetos definidos en el espacio de trabajo pueden verse haciendo clic en la solapa `Espacio de trabajo`. Para editar una variable o un conjunto de datos basta con hacer doble clic sobre él. También puede obtenerse un resumen como el que se muestra en la figura 1.3 haciendo clic en el botón derecho y seleccionando `ver` en el menú contextual que aparece.

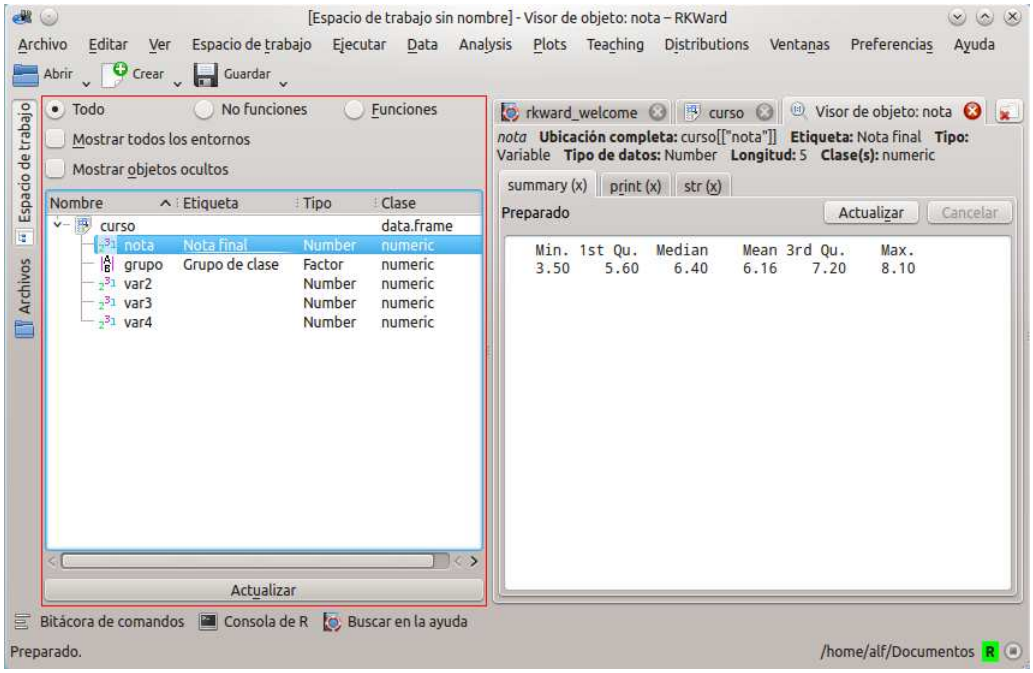


Figura 1.3 – Ventana de resumen descriptivo de un conjunto de datos

5.3 Ponderación de datos

Cuando una variable o un conjunto de datos tiene unos pocos valores que se repiten mucho, en lugar de teclearlos es más rápido indicar los valores y ponderarlos por sus frecuencias. Para ello se utiliza el menú `Teaching`▸`Datos`▸`Ponerar datos`. Al seleccionarlo aparece una ventana donde hay que seleccionar el conjunto de datos a ponderar, la variable numérica de dicho conjunto de datos que contiene las frecuencias de ponderación, e indicar un nombre para el nuevo conjunto de datos. Por ejemplo, si en una clase hay 20 chicas y 30 chicos, se puede crear un conjunto de datos con la variables sexo y frequencia, tal y como se muestra en la figura 1.4, y después llamar al menú de ponderación con los datos que aparencen la figura 1.5.

5.4 Guardar datos

Una vez introducidos los datos, conviene guardarlos en un fichero para no tener que volver a introducirlos en futuras sesiones. Para guardar los conjunto de datos definidos en el espacio de trabajo, se utiliza el menú `Espacio de trabajo`▸`Guardar espacio de trabajo`. Con esto aparece una ventana

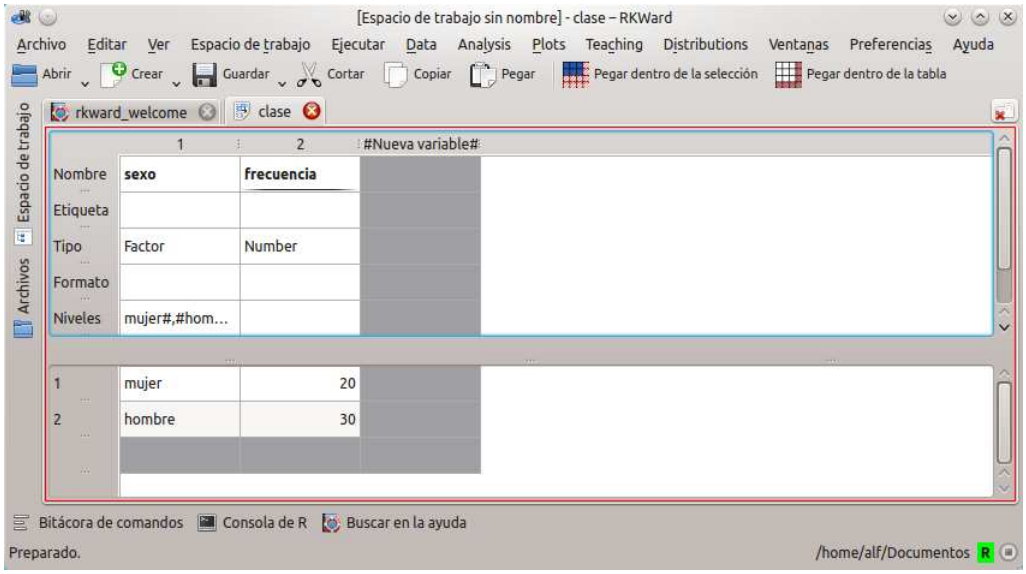


Figura 1.4 – Conjunto de datos preparado para ser ponderado

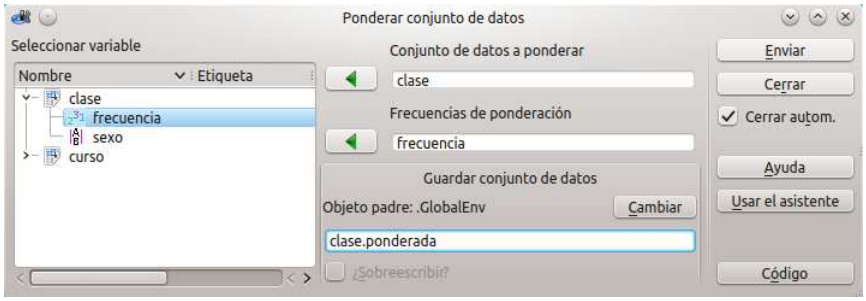


Figura 1.5 – Ventana de ponderación de datos

donde hay que darle un nombre al fichero y seleccionar la carpeta donde se guardará. Los conjuntos de datos se guardan siempre en ficheros de R con extensión `rda` o `rData`.

También es posible guardar los datos en un fichero de texto plano mediante el menú `Archivo▸ ExportarExport tabular data`. Tras esto aparece una ventana donde hay que seleccionar el conjunto de datos a exportar, darle un nombre al fichero de texto y seleccionar la carpeta donde se guardará. Esta ventana contiene también solapas donde se puede indicar entre otras cosas si incluir los nombres de las variables o no, el separador de decimales o el separador de los datos, que puede ser un espacio, tabuladores, comas u otro caracter.

5.5 Abrir datos

Si los datos con los que se pretende trabajar ya están guardados en un fichero de R, entonces tendremos que abrir dicho fichero. Para ello se utiliza el `Espacio de trabajo▸Abrir espacio de trabajo` y en la ventana que aparece se selecciona el fichero que se desea abrir. Automáticamente se cargará el conjunto de datos del fichero y pasará a ser el conjunto de datos activo.

También es posible cargar datos de ficheros con otros formatos, como por ejemplo un fichero de texto. Para ello se utiliza el menú `Archivo▸Importar▸Importar datos` y en la ventana que aparece se selecciona el fichero de texto que se desea abrir y en el cuadro desplegable del formato de archivo se debes seleccionar Text. Después aparecerá una ventana donde habrá que darle un nombre al conjunto de datos y seleccionar el tipo de separador y si los nombres de las variables aparecen en la primera línea del fichero.

5.6 Eliminación de datos

Para eliminar una variable del conjunto de datos primero hay que editar el conjunto de datos, y después, en la ventana de edición de datos, hay que hacer clic con el botón derecho del ratón sobre la cabecera de la columna correspondiente y seleccionar en el menú contextual que aparece `Borrar esta variable`.

Para eliminar individuos del conjunto de datos que hacer clic con el botón derecho del ratón sobre la cabecera de la fila correspondiente y seleccionar en el menú contextual que aparece `Borrar esta fila`.

En la ventana del espacio de trabajo también es posible borrar cualquier objeto del espacio de trabajo de R haciendo clic con el botón derecho del ratón sobre él y seleccionando el menú `Eliminar`.

6 Transformación de datos

A menudo en los análisis hay que realizar transformaciones en los datos originales. A continuación se presentan las transformaciones más habituales.

6.1 Filtrado de datos

Cuando se desea realizar un análisis con un subconjunto de individuos del conjunto de datos activo que cumplen una determinada condición es posible filtrar el conjunto de datos para quedarse con esos individuos. Para ello se utiliza el menú `Teaching`▸`Datos`▸`Filtrar`. Con esto aparece un cuadro de diálogo en el que hay que seleccionar el conjunto de datos que se desea filtrar, y en el cuadro de texto Condición de selección indicar la condición lógica que tienen que cumplir los individuos seleccionados. También hay que indicar el nombre del nuevo conjunto de datos. Por ejemplo, para seleccionar los alumnos del grupo de la mañana habría que indicar la condición `grupo=="m"` tal y como se muestra en la figura 1.6.

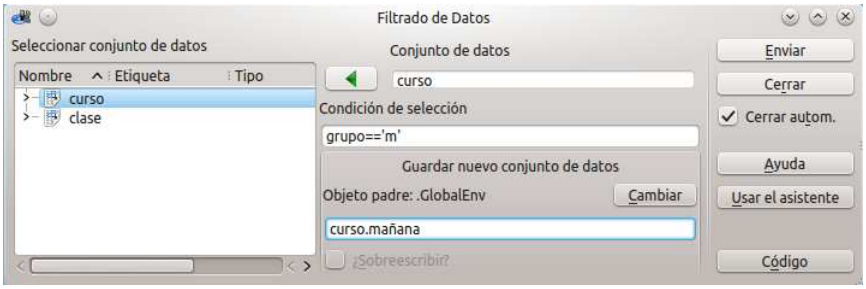


Figura 1.6 – Ventana de filtrado de datos.

6.2 Cálculo de variables

Para calcular una nueva variable a partir de otras ya existentes en el espacio de trabajo de R se utiliza el menú `Teaching`▸`Datos`▸`Calcular variable`. Con esto aparece un cuadro de diálogo en el que hay que introducir la expresión a partir de la que se calculará la nueva variable en el cuadro de texto Expresión de cálculo, e indicar el nombre de la nueva variable. La expresión de cálculo puede ser cualquier expresión aritmética o lógica de R, en las que pueden utilizarse cualquiera de las variables del espacio de trabajo de R. Por ejemplo, para eliminar los decimales de la variable nota podría crearse una nueva variable puntuacion multiplicando por 10 las notas, tal y como se muestra en la figura 1.7.

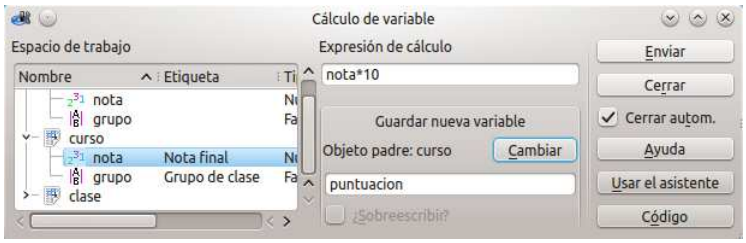


Figura 1.7 – Ventana de cálculo de nuevas variables.

6.3 Recodificación de variables

Otra transformación habitual es la recodificación de variables que permite transformar los valores de una variable de acuerdo a un conjunto de reglas de reescritura. Normalmente se utiliza para convertir una variable numérica en una variable categórica que pueda usarse como un factor.

Para recodificar una variable se utiliza el menú `Teaching▸Datos▸Recodificar variable`. Con esto aparece una ventana en la que hay que seleccionar la variable que se desea recodificar, indicar el nombre de la nueva variable recodificada e introducir las reglas de recodificación en el cuadro de texto Reglas de recodificación. Las reglas de recodificación siempre siguen la sintaxis `valor o rango de valores = nuevo valor` y pueden introducirse tantas reglas como se desee, cada una en una línea. Al lado izquierdo de la igualdad puede introducirse un único valor, varios valores separados por comas, o un rango de valores indicando el límite inferior y el límite superior del intervalo separados por el operador `:`. A la hora de definir el límite inferior puede utilizarse la palabra clave `lo` para referirse al menor de los valores de la muestra y `hi` para referirse al mayor de los valores. Por ejemplo, para recodificar la variable nota en categorías correspondientes a las calificaciones ([0-5) Suspenso, [5,7) Aprobado, [7,9) Notable y [9,10] Sobresaliente), habría que introducir las reglas que se muestran en la figura 1.8. Después, en la ventana de introducción de datos, se pueden renombrar los niveles del factor introduciendo el valor suspenso para la categoría 1, aprobado para la categoría 2, notable para la categoría 3 y sobresaliente para la categoría 4.

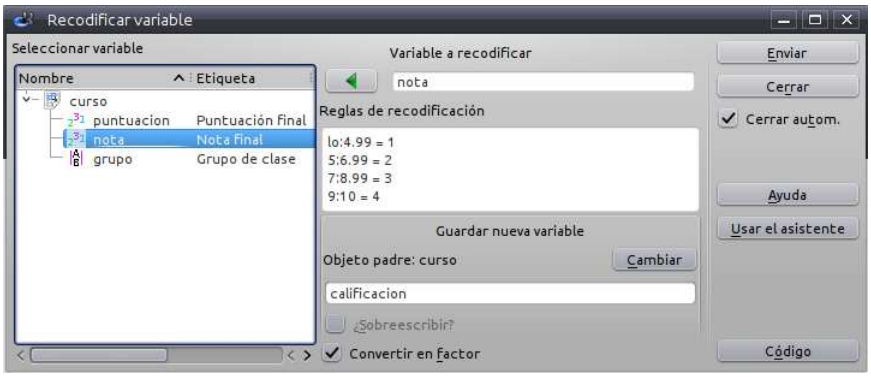


Figura 1.8 – Ventana de recodificación de variables

7 Manipulación de ficheros de resultados

7.1 Guardar los resultados

Cada vez que se ejecuta un comando de R, bien en la consola de comandos o a través de un menú, el comando ejecutado y su salida quedan registrados en la bitácora de comandos. Sin embargo, esta salida es en texto plano sin formato por lo que muchos de los procedimientos recogidos en los menús producen además una salida mucho más comprensible en formato HTML en la ventana de resultados.

Para guardar el contenido de la ventana de resultados en un fichero se utiliza el menú `Archivo ▸ Exportar página como HTML`. Con esto aparece un cuadro de diálogo en el que hay que indicar el nombre del fichero y la carpeta donde se desea guardar. El fichero resultante está en formato HTML por lo que se podrá visualizar con cualquier navegador web.

7.2 Limpiar la ventana de resultados

La vetana de resultados va acumulando todas las salidas de los análisis realizados en cada sesión de trabajo. Para no mezclar los resultados de estudios distintos, conviene limpiar la ventana de resultados cada vez que se empiece un estudio nuevo. Para ello hay que seleccionar el menú `Edición ▸ Limpiar salida`.

8 Manipulación de guiones de comandos

8.1 Creación de un guión de comandos

RKWard también incorpora un entorno de desarrollo para programadores de R que permite crear guiones de comandos que pueden ejecutarse todos seguidos. Esta opción es muy interesante para repetir análisis o automatizar tareas repetitivas. Para crear un guión de comandos hay que seleccionar el menú `Archivo ▸ Nuevo ▸ Archivo de guiones`. Con esto aparecerá una venta como la que aparece en la figura 1.9 donde se podrán teclecar los comandos de R para después ejecutarlos uno a uno o en bloque.

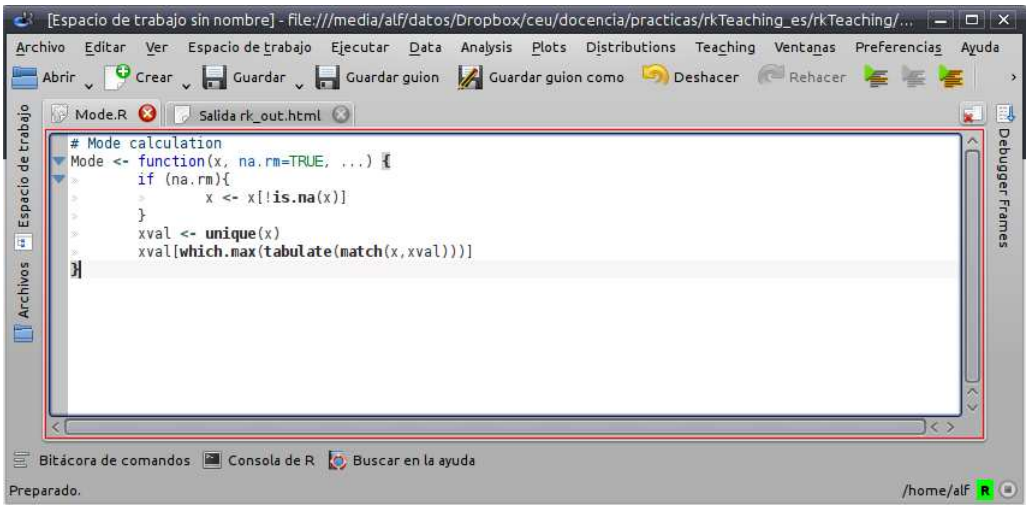


Figura 1.9 – Ventana de edición de guiones de comandos

8.2 Guardar un guión de comandos

Los guiones de comandos también pueden guardarse en un fichero de texto plano mediante el menú `Archivo ▸ Guardar guión` e indicando el nombre del fichero y la carpeta donde se guardará en el cuadro de diálo que aparece.

8.3 Abrir un guión de comandos

Para abrir un fichero con un guión de comandos se utiliza el menú `Archivo‣Abrir archivo de guiones de R` y después seleccionar el fichero que se desea abrir en el cuadro de diálogo que aparece.

9 Ayuda

Otra de las ventajas de R es que tiene un sistema de ayuda muy documentado. Es posible conseguir ayuda sobre cualquier función, prodecimiento o paquete simplemente tecleando el comando `help()`. Por ejemplo, para obtener ayuda sobre el comando `mean` se teclearía

```
> help("mean")
```

y con esto aparecerá una ventana de ayuda donde se describe la función y también aparecen ejemplos que ilustran su uso. Si no se conoce exactamente el nombre de la función o comando, se puede hacer una búsqueda aproximada con el comando `help.search()`. Por emplo, si no se recuerda el nombre de la función logarítmica, se podría teclear

```
> help("logarithm")
```

y con esto aparecerá una ventana con todos los ficheros de ayuda que contienen la palabra logarithm.

Finalmente, también es posible invocar la ayuda general de R en RKWard con el menú `Ayuda‣Ayuda de R` con lo que aparecerá una página web desde donde podremos navegar a la información deseada. También es posible buscar ayuda sobre un comando concreto en el menú `Ayuda‣Buscar en la ayuda de R`.

Para más información sobre R se recomienda visitar la página `http://www.r-project.org/`, y para más información sobre RKWard se recomienda visitar la página `http://rkward.sourceforge.net/`.

10 Ejercicios resueltos

1. Crear un conjunto de datos con los datos de la siguiente muestra y guardarlo con el nombre `colesterol.rda`

Nombre	Sexo	Peso	Altura	Colesterol
José Luis Martínez Izquierdo	H	85	179	182
Rosa Díaz Díaz	M	65	173	232
Javier García Sánchez	H	71	181	191
Carmen López Pinzón	M	65	170	200
Marisa López Collado	M	51	158	148
Antonio Ruiz Cruz	H	66	174	249

Para crear el conjunto de datos:

a) Seleccionar el menú `Archivo▸Nuevo▸Conjunto de datos`.

b) En el cuadro de diálogo que aparece introducir el nombre del conjunto de datos colesterol y hacer clic en el botón Aceptar.

c) En la ventana del editor de datos hay que definir una variable en cada columna introduciendo su nombre y tipo en las casillas de la cabecera de cada columna.

d) Una vez definidas las variables hay que introducir los datos de cada variable en la columna correspondiente.

Para guardar los datos:

a) Selecionar el menú `Espacio de trabajo▸Guardar espacio de trabajo`.

b) En el cuadro de diálogo que aparece hay que darle un nombre al fichero, seleccionar la carpeta donde guardarlo y hacer clic en el botón Aceptar.

2. Abrir el fichero creado en el ejercicio anterior y realizar las siguientes operaciones:

a) Insertar una nueva variable Edad con las edades de todos los individuos de la muestra.

Nombre	Edad
José Luis Martínez Izquierdo	18
Rosa Díaz Díaz	32
Javier García Sánchez	24
Carmen López Pinzón	35
Marisa López Collado	46
Antonio Ruiz Cruz	68

Para abrir el conjunto de datos del ejercicio anterior:

1) Seleccionar el menú `Espacio de trabajo▸Abrir espacio de trabajo`.
2) En el cuadro de diálogo que aparece seleccionar la carpeta donde se encuentra el fichero con los datos del ejercicio anterior, seleccionar el fichero y hacer clic en el botón Aceptar.

Para insertar la variable Edad:

1) Hacer clic en la solapa Espacio de trabajo.
2) En la ventana del espacio de trabajo doble clic sobre el conjunto de datos colesterol.
3) En la ventana del editor de datos introducir el nombre de la variable edad y su tipo en las casillas de la cabecera de una nueva columna vacía, e introducir los datos de las edades en las celdas de maś abajo.

b) Insertar un nuevo individuo con siguientes datos

Nombre: Cristóbal Campos Ruiz.
Edad: 44 años.
Sexo: Hombre.
Peso: 70 Kg.
Altura: 178 cm.
Colesterol: 220 mg/dl.

1) En la ventana del editor de datos introducir los datos de del nuevo individuo en la primera fila vacía.

c) Crear una nueva variable donde se calcule el índice de masa corporal de cada paciente mediante la formula:

$$\text{imc} = \frac{\text{Peso (en Kg)}}{\text{Altura (en mt)}^2}$$

1) Seleccionar el menú `Teaching▸Datos▸Calcular variable`.
2) En el cuadro de diálogo que aparece introducir la fórmula para calcular el índice de masa corporal en el campo Expresión de cálculo.
3) En el cuadro Guardar nueva variable hacer clic sobre el botón Cambiar.
4) En el cuadro de diálogo que aparece seleccionar como objeto padre la el conjunto de datos colesterol y hacer clic sobre el botón Aceptar.
5) Introducir el nombre de la nueva variable imc y hacer clic sobre el botón Aceptar.

d) Recodificar el índice de masa corporal en una nueva variable de acuerdo a las siguientes categorías:

Menor de 18,5	Bajo peso
De 18,5 a 24,5	Saludable
De 24,5 a 30	Sobrepeso
Mayor de 30	Obeso

1) Selecionar el menú `Teaching▸Datos▸Recodificar variable`.
2) En el cuadro de diálogo que aparece seleccionar como variable a recodificar la variable imc.
3) Introducir las reglas de recodificación en el campo Reglas de recodificación:

```
lo:18.5 = 1
18.5:24.5 = 2
24.5:30 = 3
30:hi = 4
```

4) En el cuadro Guardar nueva variable hacer clic sobre el botón Cambiar.
5) En el cuadro de diálogo que aparece seleccionar como objeto padre la el conjunto de datos colesterol y hacer clic sobre el botón Aceptar.
6) Introducir el nombre de la nueva variable obesidad y hacer clic sobre el botón Aceptar.
7) En la ventada de edición de datos introducir los niveles del factor, asignando Bajo peso a la categoría 1, Saludable a la categoría 2, Sobrepeso a la categoría 3 y Obeso a la categoría 4.

e) Filtrar el conjunto de datos para obtener un nuevo conjunto de datos con los datos de los hombres

1) Selecionar el menú `Teaching‣Datos‣Filtrar`.
2) En el cuadro de diálogo que aparece seleccionar como conjunto de datos colesterol.
3) En el campo Condición de selección introducir la condición `sexo=="H"`.
4) Introducir el nombre del nuevo conjunto de datos colesterol.hombres y hacer clic sobre el botón Aceptar.

11 Ejercicios propuestos

1. El conjunto de datos neonatos del paquete rk.Teaching, contiene información sobre una muestra de 320 recién nacidos en un hospital durante un año que cumplieron el tiempo normal de gestación. Se pide:

   a) Cargar el conjunto de datos.

1) Hacer clic en la solapa Espacio de trabajo para desplegarla y ver los paquetes del espacio de trabajo.
2) Hacer doble clic sobre el paquete rk.Teaching para ver todos los conjuntos de datos que contiene.
3) Hacer clic con el botón derecho sobre el conjunto de datos nenonatos y en el menú contextual que aparece selecconar `Copiar a .GlobalEnv` para hacer una copia del conjunto de datos en nuestro entorno de trabajo.

   b) Calcular la variable apgar.medio como la media de las variables apgar1 y apgar5.

   c) Recodificar la varible peso en el factor categoria.peso con dos categorias que se correspondan con los pesos menores y mayores de 2,5 Kg.

   d) Recodificar la variable apgar1 en el factor estado.apgar1 con tres categorías: deprimido (Apgar≤ 3), moderadamente deprimido (3 <Apgar≤ 6) y normal (Apgar> 6).

   e) Filtrar el conjunto de datos para quedarse con los hijos de las madres no fumadoras con una puntuación Apgar al minuto de nacer menor o igual que 3. ¿Cuántos niños hay?

Práctica 2

Distribuciones de Frecuencias y Representaciones Gráficas

1 Fundamentos teóricos

Uno de los primeros pasos en cualquier estudio estadístico es el resumen y la descripción de la información contenida en una muestra. Para ello se van a aplicar algunos métodos de análisis descriptivo, que nos permitirán clasificar y estructurar la información al igual que representarla gráficamente.

Las características que estudiamos pueden ser o no susceptibles de medida; en este sentido definiremos una *variable* como un carácter susceptible de ser medido, es decir, cuantitativo y cuantificable mediante la observación, (por ejemplo el peso de las personas, la edad, etc...), y definiremos un *atributo* como un carácter no susceptible de ser medido, y en consecuencia observable tan sólo cualitativamente (por ejemplo el color de ojos, estado de un paciente, etc...). Se llaman modalidades a las posibles observaciones de un atributo.

Dentro de los atributos, podemos hablar de *atributos ordinales*, los que presentan algún tipo de orden entre las distintas modalidades, y de *atributos nominales*, en los que no existe ningún orden entre ellas.

Dentro de las variables podemos diferenciar entre *discretas*, si sus valores posibles son valores aislados, y *continuas*, si pueden tomar cualquier valor dentro de un intervalo.

En algunos textos no se emplea el término *atributo* y se denominan a todos los caracteres *variables*. En ese caso se distinguen *variables cuantitativas* para designar las que aquí hemos definido como *variables*, y *variables cualitativas* para las que aquí se han llamado *atributos*. En lo sucesivo se aplicará este criterio para simplificar la exposición.

1.1 Cálculo de Frecuencias

Para estudiar cualquier característica, lo primero que deberemos hacer es un recuento de las observaciones, y el número de repeticiones de éstas. Para cada valor x_i de la muestra se define:

Frecuencia absoluta Es el número de veces que aparece cada uno de los valores x_i y se denota por n_i.

Frecuencia relativa Es el número de veces que aparece cada valor x_i dividido entre el tamaño muestral y se denota por f_i

$$f_i = \frac{n_i}{n}$$

Generalmente las frecuencias relativas se multiplican por 100 para que representen el tanto por ciento.

En el caso de que exista un orden entre los valores de la variable, a veces nos interesa no sólo conocer el número de veces que se repite un determinado valor, sino también el número de veces que aparece dicho valor y todos los menores. A este tipo de frecuencias se le denomina *frecuencias acumuladas*.

Frecuencia absoluta acumulada Es la suma de las frecuencias absolutas de los valores menores que x_i más la frecuencia absoluta de x_i, y se denota por N_i

$$N_i = n_1 + n_2 + \ldots + n_i$$

Frecuencia relativa acumulada Es la suma de las frecuencias relativas de los valores menores que x_i más la frecuencia relativa de x_i, y se denota por F_i

$$F_i = f_1 + f_2 + \ldots + f_i$$

Los resultados de las observaciones de los valores de una variable estadística en una muestra suelen representarse en forma de tabla. En la primera columna se representan los valores x_i de la variable colocados en orden creciente, y en la siguiente columna los valores de las frecuencias absolutas correspondientes n_i.

Podemos completar la tabla con otras columnas, correspondientes a las frecuencias relativas, f_i, y a las frecuencias acumuladas, N_i y F_i. Al conjunto de los valores de la variable observados en la muestra junto con sus frecuencias se le conoce como *distribución de frecuencias muestral*.

■ **Ejemplo 2.1** En una encuesta a 25 matrimonios, sobre el número de hijos que tienen, se obtienen los siguientes datos:

$$1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2.$$

Los valores distintos de la variable son: 0, 1, 2, 3 y 4. Así la tabla será:

x_i	Recuento	n_i
0	II	2
1	IIII I	6
2	IIII IIII IIII	14
3	II	2
4	I	1

La distribución de las frecuencias quedaría:

x_i	n_i	f_i	N_i	F_i
0	2	0,08	2	0,08
1	6	0,24	8	0,32
2	14	0,56	22	0,88
3	2	0,08	24	0,96
4	1	0,04	25	1
Suma	25	1		

Cuando el tamaño de la muestra es grande en el caso de variables discretas con muchos valores distintos de la variable, y en cualquier caso si se trata de variables continuas, se agrupan las observaciones en *clases*, que son intervalos contiguos, preferiblemente de la misma amplitud.

Para decidir el número de clases a considerar, una regla frecuentemente utilizada es tomar el entero más próximo a $\sqrt{n}$ donde n es el número de observaciones en la muestra. Pero conviene probar con distintos números de clases y escoger el que proporcione una descripción más clara. Así se prefijan los intervalos $(a_{i-1}, a_i], i = 1, 2, \ldots, l$ siendo $a = a_0 < a_1 < .... < a_l = b$ de tal modo que todos los valores observados estén dentro del intervalo $(a, b]$, y sin que exista ambigüedad a la hora de decidir a qué intervalo pertenece cada dato.

Llamaremos *marca de clase* al punto medio de cada intervalo. Así la *marca de la clase* $(a_{i-1}, a_i]$ es el punto medio x_i de dicha clase, es decir

$$x_i = \frac{a_{i-1} + a_i}{2}$$

En el tratamiento estadístico de los datos agrupados, todos los valores que están en una misma clase se consideran iguales a la marca de la clase. De esta manera si en la clase $(a_{i-1}, a_i]$ hay n_i valores observados, se puede asociar la marca de la clase x_i con esta frecuencia n_i.

1.2 Representaciones Gráficas

Hemos visto que la tabla estadística resume los datos de una muestra, de forma que ésta se puede analizar de una manera más sistemática y resumida. Para conseguir una percepción visual de las características de la población resulta muy útil el uso de gráficas y diagramas. Dependiendo del tipo de variable y de si trabajamos con datos agrupados o no, se utilizarán distintos tipos.

Diagrama de barras y polígono de frecuencias

Consiste en representar sobre el eje de abscisas de un sistema de ejes coordenados los distintos valores de la variable X, y levantar sobre cada uno de esos puntos una barra cuya altura sea igual a la frecuencia absoluta o relativa correspondiente a ese valor, tal y como se muestra en la figura 2.1(a). Esta representación se utiliza para distribuciones de frecuencias con pocos valores distintos de la variable, tanto cuantitativas como cualitativas, y en este último caso se suele representar con rectángulos de altura igual a la frecuencia de cada modalidad.

En el caso de variables cuantitativas se puede representar también el diagrama de barras de las frecuencias acumuladas, tal y como se muestra en la figura 2.1(b).

Otra representación habitual es el *polígono de frecuencias* que consiste en la línea poligonal cuyos vertices son los puntos (x_i, n_i), tal y como se ve en la figura 2.1(c), y si en vez de considerar las frecuencias absolutas o relativas se consideran las absolutas o relativas acumuladas, se obtiene el *polígono de frecuencias acumuladas*, como se ve en la figura 2.1(d).

Histogramas

Este tipo de representaciones se utiliza en variables continuas y en variables discretas en que se ha realizado una agrupación de las observaciones en clases. Un *histograma* es un conjunto de rectángulos, cuyas bases son los intervalos de clase $(a_{i-1}, a_i]$ sobre el eje OX y su altura la correspondiente frecuencia absoluta , relativa, absoluta acumulada, o relativa acumulada, tal y como se muestra en la figuras 2.2(a) y 2.2(b).

Si unimos los puntos medios de las bases superiores de los rectángulos del histograma, se obtiene el *polígono de frecuencias* correspondiente a datos agrupados (figura 2.2(c)).

El polígono de frecuencias también se puede utilizar para representar las frecuencias acumuladas, tanto absolutas como relativas. En este caso la línea poligonal se traza uniendo los extremos derechos de las bases superiores de los rectángulos del histograma de frecuencias acumuladas, en lugar de los puntos centrales (figura 2.2(d)).

Para variables cualitativas y cuantitativas discretas también se pueden usar las superficies representativas; de éstas, las más empleadas son los *sectores circulares*.

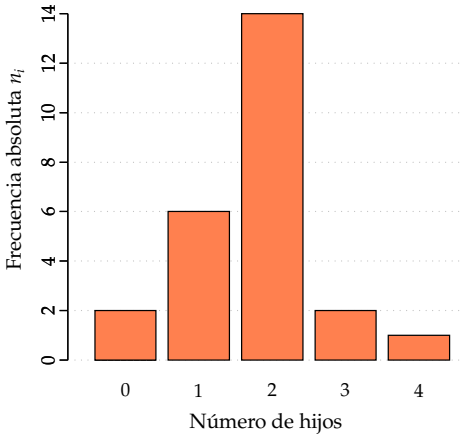
Sectores circulares o diagrama de sectores

Es una representación en la que un círculo se divide en sectores, de forma que los ángulos, y por tanto las áreas respectivas, sean proporcionales a la frecuencia.

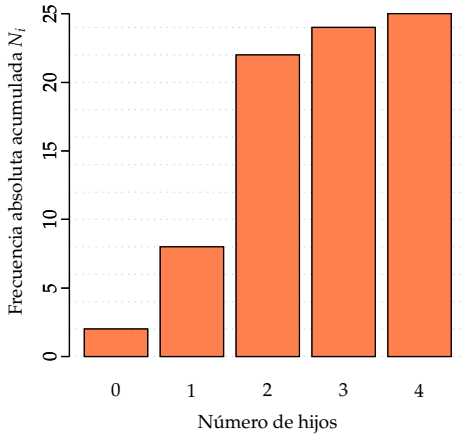
▪ **Ejemplo 2.2** Se está haciendo un estudio en una población del grupo sanguíneo de sus ciudadanos. Para ello disponemos de una muestra de 30 personas, con los siguientes resultados: 5 personas con grupo 0, 14 con grupo A, 8 con grupo B y 3 con grupo AB. El el diagrama de sectores de frecuencias relativas correspondiente aparece en la figura 2.3.

Diagrama de cajas y datos atípicos

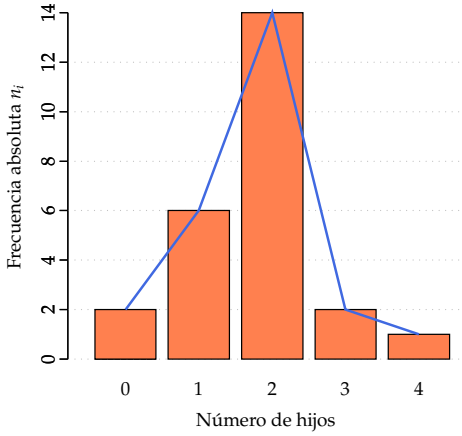
Los datos extremadamente altos o bajos, en comparación con los del resto de la muestra, reciben el nombre de datos influyentes o *datos atípicos*. Tales datos que, como su propio nombre indica, pueden modificar las conclusiones de un estudio, deben ser considerados atentamente antes de aceptarlos, pues no pocas veces podrán ser, simplemente, datos erróneos. La representación gráfica más apropiada para detectar estos datos es el *diagrama de cajas*. Este diagrama está formado por una caja que contiene el 50 % de los datos centrales de la distribución, y unos segmentos que salen de la caja, que indican los límites



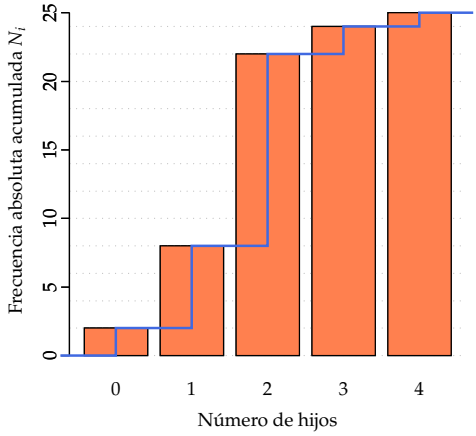
(a) Diagrama de barras de frecuencias absolutas.



(b) Diagrama de barras de frecuencias absolutas acumuladas.



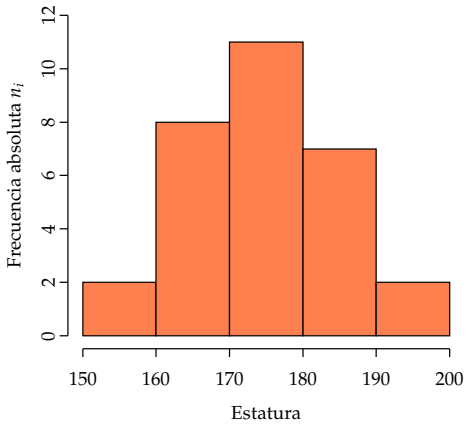
(c) Polígono de frecuencias absolutas.



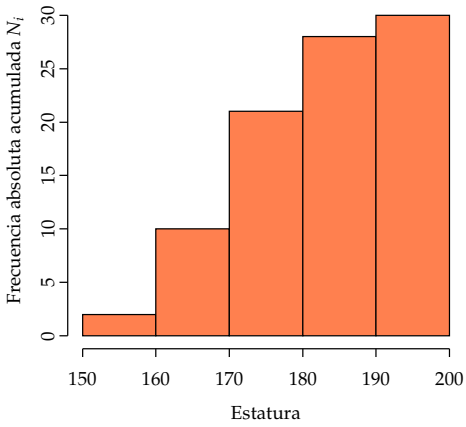
(d) Polígono de frecuencias absolutas acumuladas

Figura 2.1 – Diagramas de barras y polígonos asociados para datos no agrupados.

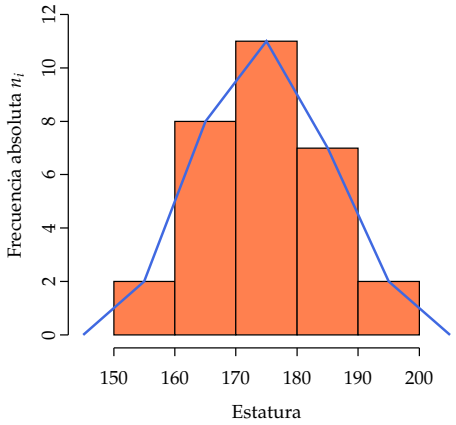
a partir de los cuales los datos se consideran atípicos. En la figura 2.4 se puede observar un ejemplo en el que aparecen dos datos atípicos.



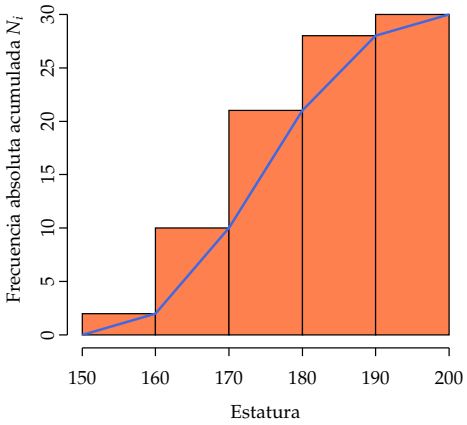
(a) Histograma de frecuencias absolutas.



(b) Histograma de frecuencias absolutas acumuladas.



(c) Polígono de frecuencias absolutas.



(d) Polígono de frecuencias absolutas acumuladas

Figura 2.2 – Histograma y polígonos asociados para datos agrupados.

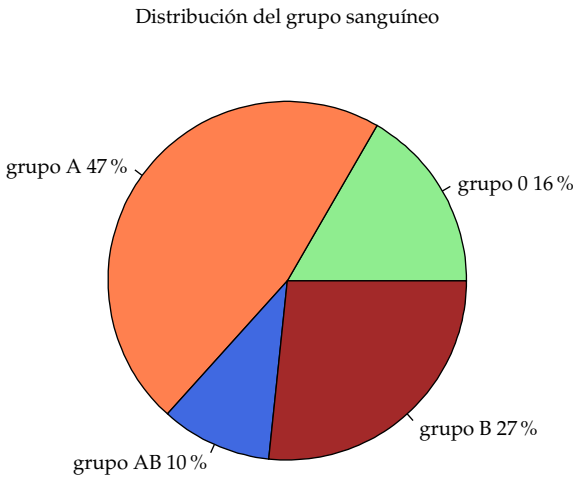


Figura 2.3 – Diagrama de sectores de frecuencias relativas del grupo sanguíneo.

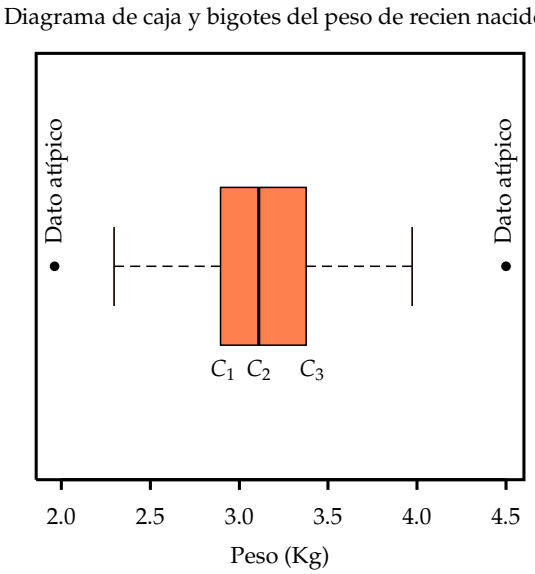


Figura 2.4 – Diagrama de cajas para una muestra de recién nacidos. Existen dos niños con pesos atípicos, uno con peso extremadamente bajo 1,9 kg, y otro con peso extremadamente alto 4,3 kg.

2 Ejercicios resueltos

1. En una encuesta a 25 matrimonios sobre el número de hijos que tenían se obtuvieron los siguientes datos:

   1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2

   Se pide:

   a) Crear un conjunto de datos con la variable hijos e introducir los datos.

   b) Construir la tabla de frecuencias.

   1) Seleccionar el menú `Teaching`▸`Distribución de frecuencias`▸`Tabla de frecuencias`.
   2) En el cuadro de diálogo que aparece, seleccionar la variable hijos en el campo Variable a tabular y hacer clic en el botón Enviar.

   c) Dibujar el diagrama de barras de las frecuencias absolutas.

   1) Seleccionar el menú `Teaching`▸`Gráficos`▸`Diagrama de barras`.
   2) En el cuadro de diálogo que aparece, seleccionar la variable hijos en el campo Variable y hacer clic en el botón Enviar.

   d) Para la misma tabla de frecuencias anterior, dibujar también el diagrama de barras de las frecuencias relativas, el de absolutas acumuladas y el de relativas acumuladas, además de sus correspondientes polígonos.

   Repetir los pasos del apartado anterior activando, en la solapa de Opciones de las barras, la opción Frecuencias relativas si se desea el diagrama de barras de frecuencias relativas, activando la opción Frecuencias acumuladas si se desea el diagrama de barras de frecuencias acumuladas y activando la opción Polígono para obtener el polígono asociado.

2. En un hospital se realizó un estudio sobre el número de personas que ingresaron en urgencias cada día del mes de noviembre. Los datos observados fueron:

   15, 23, 12, 10, 28, 50, 12, 17, 20, 21, 18, 13, 11, 12, 26
   30, 6, 16, 19, 22, 14, 17, 21, 28, 9, 16, 13, 11, 16, 20

   Se pide:

   a) Crear un conjunto de datos con la variable urgencias e introducir los datos.

   b) Dibujar el diagrama de cajas. ¿Existe algún dato atípico? En el caso de que exista, eliminarlo y proceder con los siguientes apartados.

   1) Seleccionar el menú `Teaching`▸`Gráficos`▸`Diagrama de cajas`.
   2) En el cuadro de diálogo que aparece, seleccionar la variable urgencias en el campo Variables y hacer clic en el botón Enviar.
   3) En la ventana que aparece con el diagrama de cajas identificar el dato atípico.
   4) Ir a la ventana de edición de datos y eliminar la fila del dato atípico haciendo clic con el botón derecho del ratón en la cabecera de la fila y seleccionando `Borrar esta fila`.

   c) Construir la tabla de frecuencias agrupando en 5 clases.

1) Seleccionar el menú `Teaching`▸`Distribución de frecuencias`▸`Tabla de frecuencias`.
2) En el cuadro de diálogo que aparece seleccionar la variable urgencias.
3) En la solapa de Clases activar la casilla Agrupar en intervalos, marcar la opción Número de intervalos e introducir el número deseado de intervalos en el campo Intervalos sugeridos y hacer clic sobre el botón Enviar.

d) Dibujar el histograma de frecuencias absolutas correspondiente a la tabla anterior.

1) Seleccionar el menú `Teaching`▸`Gráficos`▸`Histograma`.
2) En el cuadro de diálogo que aparece seleccionar la variable urgencias en el campo Variable.
3) En la solapa de Clases activar la casilla Agrupar en intervalos, marcar la opción Número de intervalos e introducir el número deseado de intervalos en el campo Intervalos sugeridos y hacer clic sobre el botón Enviar.

e) Para la misma tabla de frecuencias anterior, dibujar también el histograma de las frecuencias relativas, el de absolutas acumuladas y el de relativas acumuladas, además de sus correspondientes polígonos.

Repetir los pasos del apartado anterior activando, en la solapa de Opciones del histograma, la opción Frecuencias relativas si se desea el histograma de frecuencias relativas, activando la opción Frecuencias acumuladas si se desea el histograma de frecuencias acumuladas y activando la opción Polígono para obtener el polígono asociado.

3. Los grupos sanguíneos de una muestra de 30 personas son:

A, B, B, A, AB, 0, 0, A, B, B, A, A, A, A, AB,
A, A, A, B, 0, B, B, B, A, A, A, 0, A, AB, 0.

Se pide:

a) Crear un conjunto de datos con la variable grupo.sanguineo e introducir los datos.

b) Construir la tabla de frecuencias.

1) Seleccionar el menú `Teaching`▸`Distribución de frecuencias`▸`Tabla de frecuencias`.
2) En el cuadro de diálogo que aparece, seleccionar la variable grupo.sanguineo en el campo Variable a tabular y hacer clic en el botón Enviar.

c) Dibujar el diagrama de sectores.

1) Seleccionar el menú `Teaching`▸`Gráficos`▸`Diagrama de sectores`.
2) En el cuadro de diálogo que aparece, seleccionar la variable grupo.sanguineo en el campo Variables y hacer clic sobre el botón Enviar.

4. En un estudio de población se tomó una muestra de 27 personas, y se les preguntó por su edad y estado civil, obteniendo los siguientes resultados:

Estado civil	Edad								
Soltero	31	45	35	65	21	38	62	22	31
Casado	62	39	62	59	21	62			
Viudo	80	68	65	40	78	69	75		
Divorciado	31	65	59	49	65				

Se pide:

a) Crear un conjunto de datos con la variables estado.civil y edad e introducir los datos.

b) Construir la tabla de frecuencias de la variable edad para cada categoría de la variable estado.civil.

1) Seleccionar el menú `Teaching`▸`Distribución de frecuencias`▸`Tabla de frecuencias`.
2) En el cuadro de diálogo que aparece, seleccionar la variable edad en el campo Variable a tabular, activar la casilla Tabular por grupos, seleccionar la variable estado.civil en el campo Variable de agrupación y hacer clic en el botón Enviar.

c) Dibujar los diagramas de cajas de la edad según el estado civil. ¿Existen datos atípicos? ¿En qué grupo hay mayor dispersión?

1) Seleccionar el menú `Teaching`▸`Gráficos`▸`Diagrama de cajas`.
2) En el cuadro de diálogo que aparece, seleccionar la variable edad en el campo Variables, activar la casilla Dibujar por grupos, seleccionar la variable estado.civil en el campo Variable de agrupación y hacer clic en el botón Enviar.

3 Ejercicios propuestos

1. El número de lesiones padecidas durante una temporada por cada jugador de un equipo de fútbol fue el siguiente:

$$0, 1, 2, 1, 3, 0, 1, 0, 1, 2, 0, 1, 1, 1, 2, 0, 1, 3, 2, 1, 2, 1, 0, 1$$

Se pide:

a) Construir la tabla de frecuencias.

b) Dibujar el diagrama de barras de las frecuencias relativas y de frecuencias relativas acumuladas.

c) Dibujar el diagrama de sectores.

2. Para realizar un estudio sobre la estatura de los estudiantes universitarios, seleccionamos, mediante un proceso de muestreo aleatorio, una muestra de 30 estudiantes, obteniendo los siguientes resultados (medidos en centímetros):

$$179, 173, 181, 170, 158, 174, 172, 166, 194, 185,$$
$$162, 187, 198, 177, 178, 165, 154, 188, 166, 171,$$
$$175, 182, 167, 169, 172, 186, 172, 176, 168, 187.$$

Se pide:

a) Dibujar el histograma de las frecuencias absolutas agrupando desde 150 a 200 en clases de amplitud 10.

b) Dibujar el diagrama de cajas. ¿Existe algún dato atípico?.

3. El conjunto de datos neonatos del paquete rk.Teaching, contiene información sobre una muestra de 320 recién nacidos en un hospital durante un año que cumplieron el tiempo normal de gestación. Se pide:

a) Construir la tabla de frecuencias de la puntuación Apgar al minuto de nacer. Si se considera que una puntuación Apgar de 3 o menos indica que el neonato está deprimido, ¿qué porcentaje de niños está deprimido en la muestra?

b) Comparar las distribuciones de frecuencias de las puntuaciones Apgar al minuto de nacer según si la madre es mayor o menor de 20 años. ¿En qué grupo hay más neonatos deprimidos?

c) Construir la tabla de frecuencias para el peso de los neonatos, agrupando en clases de amplitud 0,5 desde el 2 hasta el 4,5. ¿En qué intervalo de peso hay más niños?

d) Comparar la distribución de frecuencias relativas del peso de los neonatos según si la madre fuma o no. Si se considera como peso bajo un peso menor de 2,5 kg, ¿En qué grupo hay un mayor porcentaje de niños con peso bajo?

e) Si en los recién nacidos se considera como peso bajo un peso menor de 2,5 kg, calcular la prevalencia del bajo peso de recién nacidos en el grupo de madres fumadoras y en el de no fumadoras.

f) Calcular el riesgo relativo de que un recién nacido tenga bajo peso cuando la madre fuma, frente a cuando la madre no fuma.

g) Construir el diagrama de barras de la puntuación Apgar al minuto. ¿Qué puntuación Apgar es la más frecuente?

h) Construir el diagrama de frecuencias relativas acumuladas de la puntuación Apgar al minuto. ¿Por debajo de que puntuación estarán la mitad de los niños?

i) Comparar mediante diagramas de barras de frecuencias relativas las distribuciones de las puntuaciones Apgar al minuto según si la madre ha fumado o no durante el embarazo. ¿Qué se puede concluir?

j) Construir el histograma de pesos, agrupando en clases de amplitud 0,5 desde el 2 hasta el 4,5. ¿En qué intervalo de peso hay más niños?

k) Comparar la distribución de frecuencias relativas del peso de los neonatos según si la madre fuma o no. ¿En qué grupo se aprecia menor peso de los niños de la muestra?

l) Comparar la distribución de frecuencias relativas del peso de los neonatos según si la madre fumaba o no antes del embarazo. ¿Qué se puede concluir?

m) Construir el diagrama de caja y bigotes del peso. ¿Entre qué valores se considera que el peso de un neonato es normal? ¿Existen datos atípicos?

n) Comparar el diagrama de cajas y bigotes del peso, según si la madre fumó o no durante el embarazo y si era mayor o no de 20 años. ¿En qué grupo el peso tiene más dispersión central? ¿En qué grupo pesan menos los niños de la muestra?

ñ) Comparar el diagrama de cajas de la puntuación Apgar al minuto y a los cinco minutos. ¿En qué variable hay más dispersión central?

Práctica 3

Estadísticos Muestrales

1 Fundamentos teóricos

Hemos visto cómo podemos presentar la información que obtenemos de la muestra, a través de tablas o bien a través de gráficas. La tabla de frecuencias contiene toda la información de la muestra pero resulta difícil sacar conclusiones sobre determinados aspectos de la distribución con sólo mirarla. Ahora veremos cómo a partir de esos mismos valores observados de la variable estadística, se calculan ciertos números que resumen la información muestral. Estos números, llamados *Estadísticos*, se utilizan para poner de manifiesto ciertos aspectos de la distribución, tales como la dispersión o concentración de los datos, la forma de su distribución, etc. Según sea la característica que pretenden reflejar se pueden clasificar en medidas de posición, medidas de dispersión y medidas de forma.

1.1 Medidas de posición

Son valores que indican cómo se sitúan los datos. Los más importantes son la Media aritmética, la Mediana y la Moda.

Media aritmética $\overline{x}$

Se llama *media aritmética* de una variable estadística X, y se representa por $\overline{x}$, a la suma de todos los resultados observados, dividida por el tamaño muestral. Es decir, la media de la variable estadística X, cuya distribución de frecuencias es (x_i, n_i), viene dada por

$$\overline{x} = \frac{x_1 + \ldots + x_1 + \ldots + x_k + \ldots + x_k}{n_1 + \ldots + n_k} = \frac{x_1 n_1 + \ldots + x_k n_k}{n} = \frac{1}{n} \sum_{i=1}^{k} x_i n_i$$

La media aritmética sólo tiene sentido en variables cuantitativas.

Mediana *Me*

Se llama *mediana* y lo denotamos por *Me*, a aquel valor de la muestra que, una vez ordenados todos los valores de la misma en orden creciente, tiene tantos términos inferiores a él como superiores. En consecuencia, divide la distribución en dos partes iguales.

La mediana sólo tiene sentido en atributos ordinales y en variables cuantitativas.

Moda *Mo*

La *moda* es el valor de la variable que presenta una mayor frecuencia en la muestra. Cuando haya más de un valor con frecuencia máxima diremos que hay más de una moda. En variables continuas o discretas agrupadas llamaremos clase modal a la que tenga la máxima frecuencia. Se puede calcular la moda tanto en variables cuantitativas como cualitativas.

Cuantiles

Si el conjunto total de valores observados se divide en r partes que contengan cada una $\frac{n}{r}$ observaciones, los puntos de separación de las mismas reciben el nombre genérico de *cuantiles*.

Según esto la mediana también es un cuantil con $r = 2$. Algunos cuantiles reciben determinados nombres como:

Cuartiles. Son los puntos que dividen la distribución en 4 partes iguales y se designan por C_1, C_2, C_3. Es claro que $C_2 = Me$.

Deciles. Son los puntos que dividen la distribución en 10 partes iguales y se designan por $D_1, D_2, \ldots, D_9$.

Percentiles. Son los puntos que dividen la distribución en 100 partes iguales y se designan por $P_1, P_2, \ldots, P_{99}$.

1.2 Medidas de dispersión

Miden la separación existente entre los valores de la muestra. Las más importantes son el Rango o Recorrido, el Rango Intercuartílico, la Varianza, la Desviación Típica y el Coeficiente de Variación.

Rango o Recorrido *Re*

La medida de dispersión más inmediata es el rango. Llamamos *recorrido* o *rango* y lo designaremos por *Re* a la diferencia entre los valores máximo y mínimo que toma la variable en la muestra, es decir

$$Re = max\{x_i, i = 1, 2, \ldots, n\} - min\{x_i, i = 1, 2, \ldots, n\}.$$

Este estadístico sirve para medir el campo de variación de la variable, aunque es la medida de dispersión que menos información proporciona sobre la mayor o menor agrupación de los valores de la variable alrededor de las medidas de tendencia central. Además tiene el inconveniente de que se ve muy afectado por los datos atípicos.

Rango Intercuartílico *RI*

El *rango intercuartílico RI* es la diferencia entre el tercer y el primer cuartil, y mide, por tanto, el campo de variación del 50 % de los datos centrales de la distribución. Por consiguiente

$$RI = C_3 - C_1.$$

La ventaja del rango intercuartílico frente al recorrido es que no se ve tan afectado por los datos atípicos.

Varianza s_x^2

Llamamos *varianza* de una variable estadística X, y la designaremos por s_x^2, a la media de los cuadrados de las desviaciones de los valores observados respecto de la media de la muestra, es decir,

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i.$$

Desviación Típica s_x

La raíz cuadrada positiva de la varianza se conoce como *desviación típica* de la variable X, y se representa por s,

$$s = +\sqrt{s_x^2}.$$

Coeficiente de Variación de Pearson Cv_x

Al cociente entre la desviación típica y el valor absoluto de la media se le conoce como *coeficiente de variación de Pearson* o simplemente *coeficiente de variación*:

$$Cv_x = \frac{s_x}{|\overline{x}|}.$$

El coeficiente de variación es adimensional, y por tanto permite hacer comparaciones entre variables expresadas en distintas unidades. Cuanto más próximo esté a 0, menor será la dispersión de la muestra en relación con la media, y más representativa será ésta última del conjunto de observaciones.

1.3 Medidas de forma

Indican la forma que tiene la distribución de valores en la muestra. Se pueden clasificar en dos grupos: Medidas de *asimetría* y medidas de *apuntamiento o curtosis*.

Coeficiente de asimetría de Fisher g_1

El *coeficiente de asimetría de Fisher*, que se representa por g_1, se define

$$g_1 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^3 f_i}{s_x^3}.$$

Dependiendo del valor que tome tendremos:

- $g_1 = 0$. Distribución simétrica.
- $g_1 < 0$. Distribución asimétrica hacia la izquierda.
- $g_1 > 0$. Distribución asimétrica hacia la derecha.

Coeficiente de apuntamiento o curtosis g_2

El grado de apuntamiento de las observaciones de la muestra, se caracteriza por el *coeficiente de apuntamiento o curtosis*, que se representa por g_2, y se define

$$g_2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^4 f_i}{s_x^4} - 3.$$

Dependiendo del valor que tome tendremos:

- $g_2 = 0$. La distribución tiene un apuntamiento igual que el de la distribución normal de la misma media y desviación típica. Se dice que es una distribución *mesocúrtica*.
- $g_2 < 0$. La distribución es menos apuntada que la distribución normal de la misma media y desviación típica. Se dice que es una distribución *platicúrtica*.
- $g_2 > 0$. La distribución es más apuntada que la distribución normal de la misma media y desviación típica. Se dice que es una distribución *leptocúrtica*.

Tanto g_1 como g_2 suelen utilizarse para comprobar si los datos muestrales provienen de una población no normal. Cuando g_1 está fuera del intervalo [-2,2] se dice que la distribución es demasiado asimétrica como para que los datos provengan de una población normal. Del mismo modo, cuando g_2 está fuera del intervalo [-2,2] se dice que la distribución es, o demasiado apuntada, o demasiado plana, como para que los datos provengan de una población normal.

1.4 Estadísticos de variables en las que se definen grupos

Ya sabemos cómo resumir la información contenida en una muestra utilizando una serie de estadísticos. Pero hasta ahora sólo hemos estudiado ejemplos con un único carácter objeto de estudio.

En la mayoría de las investigaciones no estudiaremos un único carácter, sino un conjunto de caracteres, y muchas veces será conveniente obtener información de un determinado carácter, en función de los grupos creados por otro de los caracteres estudiados en la investigación. A estas variables que se utilizan para formar grupos se les conoce como *variables clasificadoras* o *factores*.

Por ejemplo, si se realiza un estudio sobre un conjunto de niños recién nacidos, podemos estudiar su peso. Pero si además sabemos si la madre de cada niño es fumadora o no, podremos hacer un estudio del peso de los niños de las madres fumadoras por un lado y los de las no fumadoras por otro, para ver si existen diferencias entre ambos grupos.

2 Ejercicios resueltos

1. En una encuesta a 25 matrimonios sobre el número de hijos que tenían se obtuvieron los siguientes datos:

   1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2

   Se pide:

   a) Crear un conjunto de datos con la variable hijos e introducir los datos. Si ya se tienen los datos, simplemente recuperarlos.

   b) Calcular la media aritmética, varianza y desviación típica de dicha variable. Interpretar los estadísticos.

   1) Seleccionar el menú `Teaching`▸`Estadística descriptiva`▸`Estadísticos`.
   2) En el cuadro de diálogo que aparece seleccionar la variable hijos en el campo Variables.
   3) En la solapa `Estadísticos básicos` seleccionar Media y Desviación típica, y hacer click sobre el botón Enviar.

   c) Calcular los cuartiles, el recorrido, el rango intercuartílico, el tercer decil y el percentil 68.

   1) Seleccionar el menú `Teaching`▸`Estadística descriptiva`▸`Estadísticos`.
   2) En el cuadro de diálogo que aparece seleccionar la variable hijos en el campo Variables.
   3) En la solapa `Estadísticos básicos` seleccionar Cuartiles, Rango, Rango intercuartílico, introducir los valores 0,3 y 0,68 en el campo Percentiles, y hacer click sobre el botón Enviar.

2. En un hospital se realizó un estudio sobre el número de personas que ingresaron en urgencias cada día del mes de noviembre. Los datos observados fueron:

   15, 23, 12, 10, 28, 50, 12, 17, 20, 21, 18, 13, 11, 12, 26
   30, 6, 16, 19, 22, 14, 17, 21, 28, 9, 16, 13, 11, 16, 20

   Se pide:

   a) Crear un conjunto de datos con la variable urgencias e introducir los datos.

   b) Calcular la media aritmética, varianza, desviación típica y coeficiente de variación de dicha variable. Interpretar los estadísticos.

   1) Seleccionar el menú `Teaching`▸`Estadística descriptiva`▸`Estadísticos`.
   2) En el cuadro de diálogo que aparece seleccionar la variable urgencias en el campo Variables.
   3) En la solapa `Estadísticos básicos` seleccionar Media, Varianza, Desviación típica y Coeficiente de variación, y hacer click sobre el botón Enviar.

   c) Calcular el coeficiente de asimetría y el de curtosis e interpretar los resultados

   Seguir los mismos pasos del apartado anterior, seleccionando Cofeficiente de asimetría y Coeficiente de Curtosis en la solapa `Estadísticos básicos`.

3. En un grupo de 20 alumnos, las calificaciones obtenidas en Matemáticas fueron:

   SS, AP, SS, AP, AP, NT, NT, AP, SB, SS
   SB, SS, AP, AP, NT, AP, SS, NT, SS, NT

Se pide:

a) Crear un conjunto de datos curso con la variable calificaciones e introducir los datos.

b) Recodificar esta variable, asignando 2,5 al SS, 6 al AP, 8 al NT y 9,5 al SB.

1) Seleccionar el menú `Teaching▸Datos▸Recodificar variable`.
2) En el cuadro de diálogo que aparece seleccionar como variable a recodificar la variable calificaciones.
3) Introducir las reglas de recodificación en el campo Reglas de recodificación:

```
"SS" = 2.5
"AP" = 6
"NT" = 8
"SB" = 9.5
```

4) En el cuadro Guardar nueva variable hacer click sobre el botón Cambiar.
5) En el cuadro de diálogo que aparece seleccionar como objeto padre la el conjunto de datos curso y hacer click sobre el botón Enviar.
6) Introducir el nombre de la nueva variable nota, desmarcar la casilla Convertir en factor y hacer click sobre el botón Enviar.

c) La mediana y el rango intercuartílico.

1) Seleccionar el menú `Teaching▸Estadística descriptiva▸Estadísticos`.
2) En el cuadro de diálogo que aparece seleccionar la variable nota en el campo Variables.
3) En la solapa `Estadísticos básicos` seleccionar Mediana y Rango intercuartílico, y hacer click sobre el botón Enviar.

4. Para realizar un estudio sobre la estatura de los estudiantes universitarios se ha seleccionado mediante un proceso de muestreo aleatorio, una muestra de 30 estudiantes, obteniendo los siguientes resultados (medidos en centímetros):

Mujeres:	173, 158, 174, 166, 162, 177, 165, 154, 166, 182, 169, 172, 170, 168.
Hombres:	179, 181, 172, 194, 185, 187, 198, 178, 188, 171, 175, 167, 186, 172, 176, 187.

Se pide:

a) Crear un conjunto de datos con las variables estatura y sexo e introducir los datos.

b) Obtener un resumen de estadísticos en el que se muestren la media aritmética, mediana, varianza, desviación típica y cuartiles según el sexo. Interpretar los estadísticos.

1) Seleccionar el menú `Teaching▸Estadística descriptiva▸Estadísticos`.
2) En el cuadro de diálogo que aparece seleccionar la variable estatura en el campo Variables, marcar la casilla Estadística por grupos y seleccionar la variable sexo en el campo Variables de agrupación.
3) En la solapa `Estadísticos básicos` seleccionar Media, Mediana, Varianza, Desviación típica y Cuartiles, y hacer click sobre el botón Enviar.

3 Ejercicios propuestos

1. El número de lesiones padecidas durante una temporada por cada jugador de un equipo de fútbol fue el siguiente:

0, 1, 2, 1, 3, 0, 1, 0, 1, 2, 0, 1, 1, 1, 2, 0, 1, 3, 2, 1, 2, 1, 0, 1

Se pide:

a) Calcular la media aritmética, mediana, varianza y desviación típica de las lesiones e interpretarlas.

b) Calcular los coeficientes de asimetría y curtosis e interpretarlos.

c) Calcular el cuarto y el octavo decil e interpretarlos.

2. En un estudio de población se tomó una muestra de 27 personas, y se les preguntó por su edad y estado civil, obteniendo los siguientes resultados:

Estado civil	Edad								
Soltero	31	45	35	65	21	38	62	22	31
Casado	62	39	62	59	21	62			
Viudo	80	68	65	40	78	69	75		
Divorciado	31	65	59	49	65				

Se pide:

a) Calcular la media y la desviación típica de la edad según el estado civil e interpretarlas.

b) ¿En qué grupo es más representativa la media?

3. En un estudio se ha medido la tensión arterial de 25 individuos. Además se les ha preguntado si fuman y beben:

Fumador	si	no	si	si	si	no	no	si	no	si	no	si	no
Bebedor	no	no	si	si	no	no	si	si	no	si	no	si	si
Tensión arterial	80	92	75	56	89	93	101	67	89	63	98	58	91

Fumador	si	no	no	si	no	no	no	si	no	si	no	si
Bebedor	si	no	si	si	no	no	si	si	si	no	si	no
Tensión arterial	71	52	98	104	57	89	70	93	69	82	70	49

Calcular la media aritmética, desviación típica, coeficiente de asimetría y curtosis de la tensión arterial por grupos dependiendo de si beben o fuman e interpretarlos.

4. El conjunto de datos neonatos del paquete rk.Teaching, contiene información sobre una muestra de 320 recién nacidos en un hospital durante un año que cumplieron el tiempo normal de gestación. Se pide:

a) Calcular la media y la mediana muestral del peso de los nacidos e interpretarlos.

b) Calcular el peso medio de los recién nacidos de la muestra según si la madre ha fumado o no durante el embarazo. Calcular también el peso medio de los recién nacidos de madres que no han fumado durante el embarazo, según si la madre fumaba o no antes del embarazo. ¿Qué conclusiones se pueden sacar?

c) ¿Cuál es la puntuación Apgar al minuto de nacer más frecuente?

d) Calcular la media de la diferencia entre las puntuaciones Apgar a los 5 minutos y al minuto de nacer. ¿Cómo evolucionan los recién nacidos?

e) Calcular los cuartiles muestrales del peso de los recién nacidos e interpretarlos.

f) Comparar los cuartiles muestrales del peso de los recién nacidos según el sexo.

g) ¿Por encima de qué peso estarán el 10 % de los niños con mayor peso?

h) Si se considera que un niño es atípico por bajo peso si se encuentra entre el 5 % de los pesos más bajos, ¿por debajo de qué peso tiene que estar?

i) Calcular el recorrido y el rango intercuartílico muestrales del peso de los recién nacidos e interpretarlos.

j) Calcular la varianza y la desviación típica del peso de los recién nacidos e interpretarlos.

k) ¿En qué grupo hay más variabilidad del peso de los recién nacidos, en las madres fumadoras o en las madres no fumadoras durante el embarazo? ¿En qué grupo será más representativo el peso medio?

l) ¿Qué variable presenta más variabilidad relativa, el peso de los recién nacidos o el Apgar al minuto de nacer?

m) Calcular el coeficiente de asimetría y de apuntamiento muestrales del peso de los recién nacidos e interpretarlos.

n) ¿Qué distribución es más asimétrica, la de los pesos de recién nacidos en madres mayores de 20 años o en madres menores de 20 años?

ñ) ¿Qué distribución es más apuntada, la del peso de los recién nacidos en hombres o en mujeres?

o) De acuerdo a la forma de la distribución, ¿puede considerarse la puntuación Apgar al minuto de nacer como una variable normal? ¿Y el número de cigarros fumados al día durante el embarazo?

5. Se quiere comparar la precisión de dos tensiómetros, uno de brazo y otro de muñeca, y para ello se han realizado 8 medidas repetidas de la tensión arterial de una misma persona con cada uno de ellos, obteniendo los siguientes valores en mmHg:

   - tens.brazo: 111, 109, 112, 111, 113, 113, 114, 111.
   - tens.muñeca: 115, 113, 117, 116, 112, 112, 117, 112.

   ¿Qué tensiómetro es más preciso?

Práctica 4

Regresión Lineal Simple y Correlación

1 Fundamentos teóricos

1.1 Regresión

La *regresión* es la parte de la estadística que trata de determinar la posible relación entre una variable numérica Y, que suele llamarse *variable dependiente*, y otro conjunto de variables numéricas, $X_1, X_2, \ldots, X_n$, conocidas como *variables independientes*, de una misma población. Dicha relación se refleja mediante un modelo funcional $y = f(x_1, \ldots, x_n)$.

El caso más sencillo se da cuando sólo hay una variable independiente X, y entonces se habla de *regresión simple*. En este caso el modelo que explica la relación entre X e Y es una función de una variable $y = f(x)$.

Dependiendo de la forma de esta función, existen muchos tipos de regresión simple. Los más habituales son los que aparecen en la siguiente tabla:

Modelo	Ecuación genérica
Lineal	$y = a + bx$
Parabólico	$y = a + bx + cx^2$
Polinómico de grado n	$y = a_0 + a_1x + \cdots + a_nx^n$
Potencial	$y = ax^b$
Exponencial	$y = e^{a+bx}$
Logarítmico	$y = a + b\log x$
Inverso	$y = a + b/x$
Curva S	$y = e^{a+b/x}$

Para elegir un tipo de modelo u otro, se suele representar el *diagrama de dispersión*, que consiste en dibujar sobre unos ejes cartesianos correspondientes a las variables X e Y, los pares de valores (x_i, y_j) observados en cada individuo de la muestra.

▪ **Ejemplo 4.1** En la figura la figura 4.1 aparece el diagrama de dispersión correspondiente a una muestra de 30 individuos en los que se ha medido la estatura en cm (X) y el peso en kg (Y). En este caso la forma de la nube de puntos refleja una relación lineal entre la estatura y el peso.

Según la forma de la nube de puntos del diagrama, se elige el modelo más apropiado (figura 4.2), y se determinan los parámetros de dicho modelo para que la función resultante se ajuste lo mejor posible a la nube de puntos.

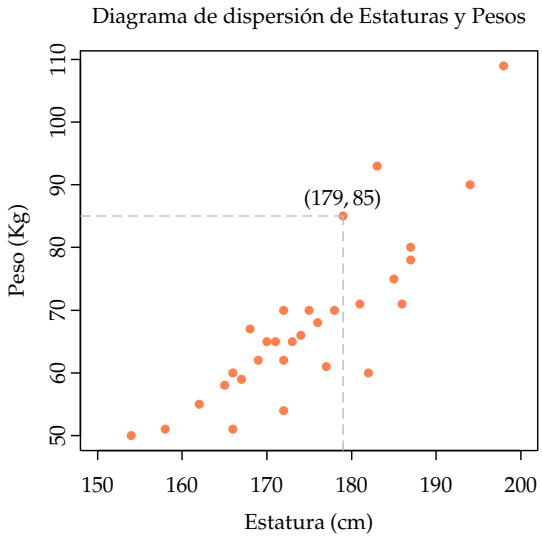
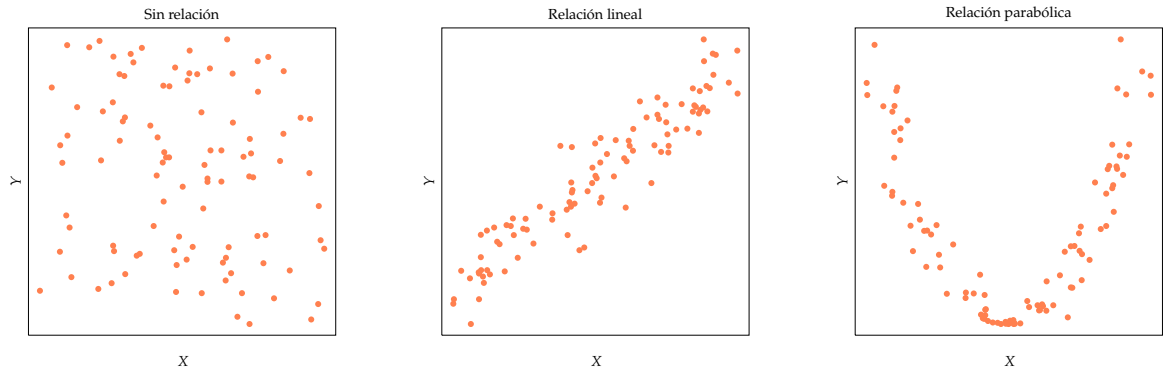


Figura 4.1 – Diagrama de dispersión. El punto (179,85) indicado corresponde a un individuo de la muestra que mide 179 cm y pesa 85 Kg.



(a) Sin relación. (b) Relación lineal. (c) Relación polinómica.

(d) Relación exponencial. (e) Relación logarítmica. (f) Relación inversa.

Figura 4.2 – Diagramas de dispersión correspondientes a distintos tipos de relaciones entre variables.

El criterio que suele utilizarse para obtener la función óptima, es que la distancia de cada punto a la curva, medida en el eje Y, sea lo menor posible. A estas distancias se les llama *residuos* o *errores* en Y (figura 4.3). La función que mejor se ajusta a la nube de puntos será, por tanto, aquella que hace mínima la suma de los cuadrados de los residuos.[1]

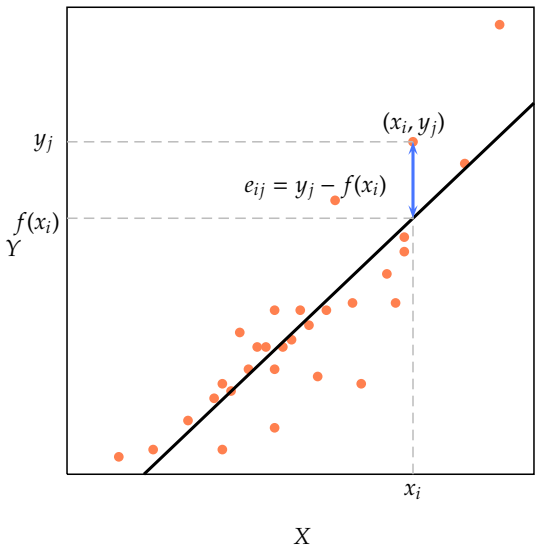


Figura 4.3 – Residuos o errores en Y. El residuo correspondiente a un punto (x_i, y_j) es la diferencia entre el valor y_j observado en la muestra, y el valor teórico del modelo $f(x_i)$, es decir, $e_{ij} = y_j - f(x_i)$.

Rectas de regresión

En el caso de que la nube de puntos tenga forma lineal y optemos por explicar la relación entre X e Y mediante una recta $y = a + bx$, los parámetros a determinar son a (punto de corte con el eje de ordenadas) y b (pendiente de la recta). Los valores de estos parámetros que hacen mínima la suma de residuos al cuadrado, determinan la recta óptima. Esta recta se conoce como *recta de regresión de Y sobre X* y explica la variable Y en función de la variable X. Su ecuación es

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}),$$

donde s_{xy} es un estadístico llamado *covarianza* que mide el grado de relación lineal, y cuya fórmula es

$$s_{xy} = \frac{1}{n}\sum_{i,j}(x_i - \bar{x})(y_j - \bar{y})n_{ij}.$$

▪ **Ejemplo 4.2** En la figura 4.4 aparecen las rectas de regresión de Estatura sobre Peso y de Peso sobre Estatura del ejemplo anterior.

La pendiente de la recta de regresión de Y sobre X se conoce como *coeficiente de regresión de Y sobre X*, y mide el incremento que sufrirá la variable Y por cada unidad que se incremente la variable X, según la recta.

Cuanto más pequeños sean los residuos, en valor absoluto, mejor se ajustará el modelo a la nube de puntos, y por tanto, mejor explicará la relación entre X e Y. Cuando todos los residuos son nulos, la recta pasa por todos los puntos de la nube, y la relación es perfecta. En este caso ambas rectas, la de Y sobre X y la de X sobre Y coinciden (figura 4.5(a)).

[1] Se elevan al cuadrado para evitar que en la suma se compensen los residuos positivos con los negativos.

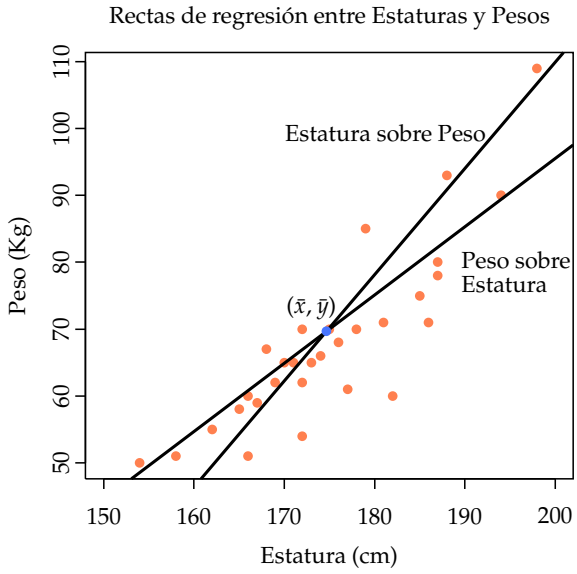
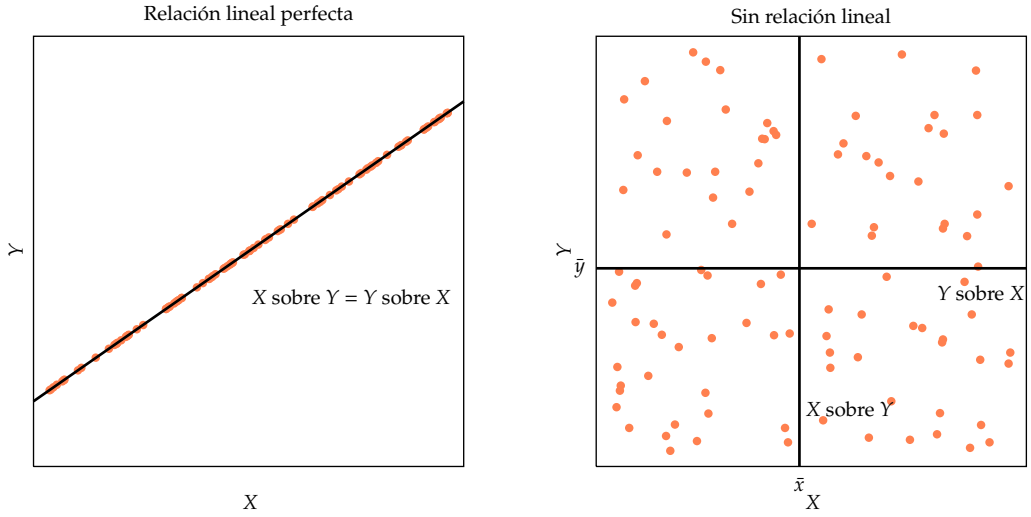


Figura 4.4 – Rectas de regresión de Estatura sobre Peso y de Peso sobre Estatura. Las rectas de regresión siempre se cortan en el punto de medias $(\bar{x}, \bar{y})$

Por contra, cuando no existe relación lineal entre las variables, la recta de regresión de Y sobre X tiene pendiente nula, y por tanto la ecuación es $y = \bar{y}$, en la que, efectivamente no aparece x, o $x = \bar{x}$ en el caso de la recta de regresión X sobre Y, de manera que ambas rectas se cortan perpendicularmente (figura 4.5(b)).



(a) Dependencia funcional lineal.

(b) Independencia lineal.

Figura 4.5 – Distintos grados de dependencia. En el primer caso, la relación es perfecta y los residuos son nulos. En el segundo caso no existe relación lineal y la pendiente de la recta es nula.

1.2 Correlación

El principal objetivo de la regresión simple es construir un modelo funcional $y = f(x)$ que explique lo mejor posible la relación entre dos variables X (variable independiente) e Y (variable dependiente) medidas en una misma muestra. Generalmente, el modelo construido se utiliza para realizar inferencias predictivas de Y en función de X en el resto de la población. Pero aunque la regresión garantiza que el modelo construido es el mejor posible, dentro del tipo de modelo elegido (lineal, polinómico, exponencial, logarítmico, etc.), puede que aún así, no sea un buen modelo para hacer predicciones, precisamente porque no haya relación de ese tipo entre X e Y. Así pues, con el fin de validar un modelo para realizar predicciones fiables, se necesitan medidas que nos hablen del grado de dependencia entre X e Y, con respecto a un modelo de regresión construido. Estas medidas se conocen como medidas de *correlación*.

Dependiendo del tipo de modelo ajustado, habrá distintos tipos de medidas de correlación. Así, si el modelo de regresión construido es una recta, hablaremos de correlación lineal; si es un polinomio, hablaremos de correlación polinómica; si es una función exponencial, hablaremos de correlación exponencial, etc. En cualquier caso, estas medidas nos hablarán de lo bueno que es el modelo construido, y como consecuencia, de si podemos fiarnos de las predicciones realizadas con dicho modelo.

La mayoría de las medidas de correlación surgen del estudio de los residuos o errores en Y, que son las distancias de los puntos del diagrama de dispersión a la curva de regresión construida, medidas en el eje Y, tal y como se muestra en la figura (4.3). Estas distancias, son en realidad, los errores predictivos del modelo sobre los propios valores de la muestra.

Cuanto más pequeños sean los residuos, mejor se ajustará el modelo a la nube de puntos, y por tanto, mejor explicará la relación entre X e Y. Cuando todos los residuos son nulos, la curva de regresión pasa por todos los puntos de la nube, y entonces se dice que la relación es perfecta, o bien que existe una dependencia funcional entre X e Y (figura 4.5(a)). Por contra, cuando los residuos sean grandes, el modelo no explicará bien la relación entre X e Y, y por tanto, sus predicciones no serán fiables (figura 4.5(b)).

Varianza residual

Una primera medida de correlación, construida a partir de los residuos es la *varianza residual*, que se define como el promedio de los residuos al cuadrado:

$$s_{ry}^2 = \frac{\sum_{i,j} e_{ij}^2 n_{ij}}{n} = \frac{\sum_{i,j} (y_j - f(x_i))^2 n_{ij}}{n}.$$

Cuando los residuos son nulos, entonces $s_{ry}^2 = 0$ y eso indica que hay dependencia funcional. Por otro lado, cuando las variables son independientes, con respecto al modelo de regresión ajustado, entonces los residuos se convierten en las desviaciones de los valores de Y con respecto a su media, y se cumple que $s_{ry}^2 = s_y^2$. Así pues, se cumple que

$$0 \leq s_{ry}^2 \leq s_y^2.$$

Según esto, cuanto menor sea la varianza residual, mayor será la dependencia entre X e Y, de acuerdo al modelo ajustado. No obstante, la varianza tiene como unidades las unidades de Y al cuadrado, y eso dificulta su interpretación.

Coeficiente de determinación

Puesto que el valor máximo que puede tomar la varianza residual es la varianza de Y, se puede definir fácilmente un coeficiente a partir de la comparación de ambas medidas. Surge así el *coeficiente de determinación* que se define como

$$R^2 = 1 - \frac{s_{ry}^2}{s_y^2}.$$

Se cumple que

$$0 \leq R^2 \leq 1,$$

y además no tiene unidades, por lo que es más fácil de interpretar que la varianza residual:

- $R^2 = 0$ indica que existe independencia según el tipo de relación planteada por el modelo de regresión.
- $R^2 = 1$ indica dependencia funcional.

Por tanto, cuanto mayor sea R^2, mejor será el modelo de regresión.

Si multiplicamos el coeficiente de determinación por 100, se obtiene el porcentaje de variabilidad de Y que explica el modelo de regresión. El porcentaje restante corresponde a la variabilidad que queda por explicar y se corresponde con el error predictivo del modelo. Así, por ejemplo, si tenemos un coeficiente de determinación $R^2 = 0{,}5$, el modelo de regresión explicaría la mitad de la variabilidad de Y, y en consecuencia, si se utiliza dicho modelo para hacer predicciones, estas tendrían la mitad de error que si no se utilizase, y se tomase como valor de la predicción el valor de la media de Y.

Coeficiente de determinación lineal

En el caso de que el modelo de regresión sea lineal, la fórmula del coeficiente de determinación se simplifica y se convierte en

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2},$$

que se conoce como *coeficiente de determinación lineal*.

Coeficiente de correlación

Otra medida de dependencia bastante habitual es el *coeficiente de correlación*, que se define como la raíz cuadrada del coeficiente de determinación:

$$R = \pm\sqrt{1 - \frac{s_{ry}^2}{s_y^2}},$$

tomando la raíz del mismo signo que la covarianza.

La única ventaja del coeficiente de correlación con respecto al coeficiente de determinación, es que tiene signo, y por tanto, además del grado de dependencia entre X e Y, también nos habla de si la relación es directa (signo +) o inversa (signo -). Su interpretación es:

- $R = 0$ indica independencia con respecto al tipo de relación planteada por el modelo de regresión.
- $R = -1$ indica dependencia funcional inversa.
- $R = 1$ indica dependencia funcional directa.

Por consiguiente, cuanto más próximo esté a -1 o a 1, mejor será el modelo de regresión.

Coeficiente de correlación lineal Al igual que ocurría con el coeficiente de determinación, cuando el modelo de regresión es lineal, la fórmula del coeficiente de correlación se convierte en

$$r = \frac{s_{xy}}{s_x s_y},$$

y se llama *coeficiente de correlación lineal*.

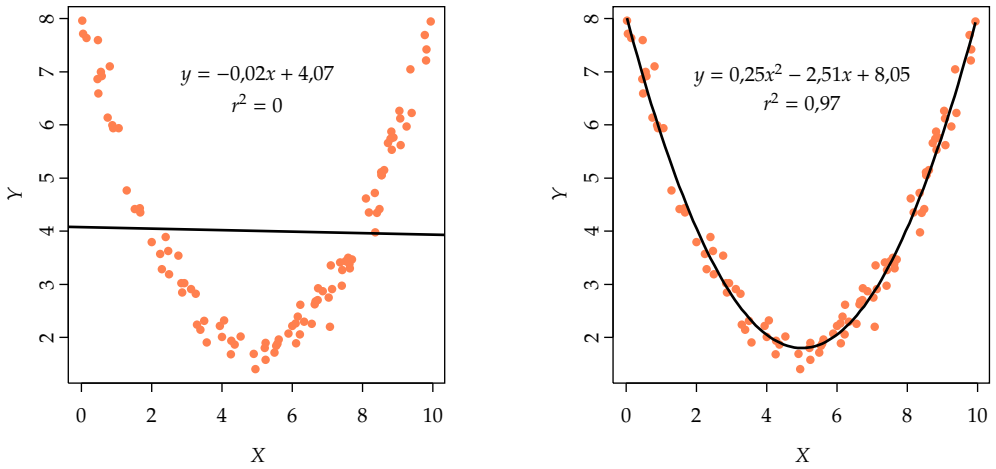
Por último, conviene remarcar que un coeficiente de determinación o de correlación nulo, indica que hay independencia según el modelo de regresión construido, pero puede haber dependencia de otro tipo. Esto se ve claramente en el ejemplo de la figura 4.6.

Fiabilidad de las predicciones

Aunque el coeficiente de determinación o de correlación nos hablan de la bondad de un modelo de regresión, no es el único dato que hay que tener en cuenta a la hora de hacer predicciones.

La fiabilidad de las predicciones que hagamos con un modelo de regresión depende de varias cosas:

- El coeficiente de determinación: Cuando mayor sea, menores serán los errores predictivos y mayor la fiabilidad de las predicciones.



(a) Dependencia lineal débil.

(b) Dependencia parabólica fuerte.

Figura 4.6 – En la figura de la izquierda se ha ajustado un modelo lineal y se ha obtenido un $R^2 = 0$, lo que indica que el modelo no explica nada de la relación entre X e Y, pero no podemos afirmar que X e Y son independientes. De hecho, en la figura de la derecha se observa que al ajustar un modelo parabólico, $R^2 = 0{,}97$, lo que indica que casi hay una dependencia funcional parabólica entre X e Y.

- La variablidad de la población: Cuanto más variable es una población, más difícil es predecir y por tanto menos fiables serán las predicciones del modelo.
- El tamaño muestral: Cuanto mayor sea, más información tendremos y, en consecuencia, más fiables serán las predicciones.

Además, hay que tener en cuenta que un modelo de regresión es válido para el rango de valores observados en la muestra, pero fuera de ese rango no tenemos información del tipo de relación entre las variables, por lo que no deberíamos hacer predicciones para valores que estén lejos de los observados en la muestra.

2 Ejercicios resueltos

1. Se han medido dos variables X e Y en 10 individuos obteniendo los siguientes resultados:

X	0	1	2	3	4	5	6	7	8	9
Y	2	5	8	11	14	17	20	23	26	29

Se pide:

a) Crear un conjunto de datos con las variables X y Y e introducir estos datos.

b) Dibujar el diagrama de dispersión correspondiente.

1) Seleccionar el menú `Teaching`▸`Gráficos`▸`Diagrama de Dispersión`.
2) En el cuadro de diálogo que aparece, seleccionar la variable Y en el campo Variable Y, la variable X en el campo Variable X, y hacer clic en el botón Enviar.

En vista del diagrama, ¿qué tipo de modelo crees que explicará mejor la relación entre X e Y?

c) Calcular la recta de regresión de Y sobre X.

1) Seleccionar el menú `Teaching`▸`Regresión`▸`Regresión lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable Y en el campo Variable dependiente y la variable X en el campo Variable independiente, y hacer clic sobre el botón Enviar.

d) Dibujar dicha recta sobre el diagrama de dispersión.

1) Seleccionar el menú `Teaching`▸`Gráficos`▸`Diagrama de Dispersión`.
2) En el cuadro de diálogo que aparece, seleccionar la variable Y en el campo Variable Y, la variable X en el campo Variable X, y hacer clic en el botón Enviar.
3) En la solapa `Línea de ajuste`, seleccionar Dibujar recta de regresión y hacer clic en el botón Enviar.

e) Calcular la recta de regresión de X sobre Y y dibujarla sobre el correspondiente diagrama de dispersión.

Repetir los pasos de los apartados anteriores pero escogiendo como Variable dependiente la variable X, y como Variable independiente la variable Y

f) ¿Son grandes los residuos? Comentar los resultados.

2. En una licenciatura se quiere estudiar la relación entre el número medio de horas de estudio diarias y el número de asignaturas suspensas. Para ello se obtuvo la siguiente muestra:

Horas	Suspensos	Horas	Suspensos	Horas	Suspensos
3,5	1	2,2	2	1,3	4
0,6	5	3,3	0	3,1	0
2,8	1	1,7	3	2,3	2
2,5	3	1,1	3	3,2	2
2,6	1	2,0	3	0,9	4
3,9	0	3,5	0	1,7	2
1,5	3	2,1	2	0,2	5
0,7	3	1,8	2	2,9	1
3,6	1	1,1	4	1,0	3
3,7	1	0,7	4	2,3	2

Se pide:

a) Crear un conjunto de datos con las variables horas.estudio y suspensos e introducir estos datos.

b) Construir la tabla de frecuencias bidimensional de las variables horas.estudio y suspensos.

1) Seleccionar el menú `Teaching`▸`Distribución de frecuencias`▸`Tabla de frecuencias bidimensional`.
2) En el cuadro de diálogo que aparece, seleccionar la variable horas.estudio en el campo Variable a tabular en filas, la variable suspensos en el campo Variable a tabular en columnas, y hacer clic sobre el botón Enviar.

c) Calcular la recta de regresión de suspensos sobre horas.estudio y dibujarla.

Para calcular la recta de regresión:

1) Seleccionar el menú `Teaching`▸`Regresión`▸`Regresión lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable suspensos en el campo Variable dependiente y la variable horas.estudio en el campo Variable independiente, seleccionar Guardar el modelo, introducir un nombre para el modelo y hacer clic sobre el botón Enviar.

Para dibujar la recta de regresión:

1) Seleccionar el menú `Teaching`▸`Gráficos`▸`Diagrama de Dispersión`.
2) En el cuadro de diálogo que aparece, seleccionar la variable suspensos en el campo Variable Y y la variable horas.estudio en el campo Variable X.
3) En la solapa `Línea de ajuste`, seleccionar Lineal y hacer clic en el botón Enviar.

d) Indicar el coeficiente de regresión de suspensos sobre horas.estudio. ¿Cómo lo interpretarías?

El coeficiente de regresión es la pendiente de la recta de regresión.

e) La relación lineal entre estas dos variables, ¿es mejor o peor que la del ejercicio anterior? Comentar los resultados a partir las gráficas de las rectas de regresión y sus residuos.

f) Calcular los coeficientes de correlación y de determinación lineal. ¿Es un buen modelo la recta de regresión? ¿Qué porcentaje de la variabilidad del número de suspensos está explicada por el modelo?

El coeficiente de determinación aparece en la ventana de resultados como R^2, y el coeficiente de correlación es su raíz cuadrada.

g) Utilizar la recta de regresión para predecir el número de suspensos correspondiente a 3 horas de estudio diarias. ¿Es fiable esta predicción?

1) Seleccionar el menú `Teaching`▸`Regresión`▸`Predicciones`.
2) En el cuadro de diálogo que aparece seleccionar como modelo de regresión la recta calculada en el segundo apartado, introducir los valores para los que se desea la predicción en el campo Predicciones para y hacer clic sobre el botón Enviar.

h) Según el modelo lineal, ¿cuántas horas diarias tendrá que estudiar como mínimo un alumno si quiere aprobarlo todo?

Seguir los mismos pasos de los apartados anteriores, pero escogiendo como variable dependiente horas.estudio, y como independiente suspensos, y haciendo la predicción para 0 suspensos.

3. Después de tomar un litro de vino se ha medido la concentración de alcohol en la sangre en distintos instantes, obteniendo:

Tiempo después (minutos)	30	60	90	120	150	180	210
Concentración (gramos/litro)	1,6	1,7	1,5	1,1	0,7	0,2	2,1

Se pide:

a) Crear las variables tiempo y alcohol e introducir estos datos.

b) Calcular el coeficiente de correlación lineal entre el alcohol y el tiempo e interpretarlo. ¿Es bueno el modelo lineal?

1) Seleccionar el menú `Teaching`▸`Regresión`▸`Regresión lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable alcohol en el campo Variable dependiente y la variable tiempo en el campo Variable independiente, y hacer clic sobre el botón Enviar.

c) Dibujar la recta de regresión del alcohol sobre el tiempo. ¿Existe algún individuo con un residuo demasiado grande? Si es así, eliminar dicho individuo de la muestra y volver a calcular el coeficiente de correlación. ¿Ha mejorado el modelo?

1) Seleccionar el menú `Teaching`▸`Gráficos`▸`Diagrama de Dispersión`.
2) En el cuadro de diálogo que aparece, seleccionar la variable alcohol en el campo Variable Y y la variable tiempo en el campo Variable X.
3) En la solapa `Línea de ajuste`, seleccionar Lineal y hacer clic en el botón Enviar.

Se observa que hay un residuo atípico para el punto que corresponde al los 210 minutos. Para eliminarlo: En la ventana de edición del conjunto de datos hacer clic con el botón derecho del ratón sobre la fila correspondiente al dato con el residuo atípico y seleccionar Borrar esta fila.

d) Si la concentración máxima de alcohol en la sangre que permite la ley para poder conducir es 0,3 g/l, ¿cuánto tiempo habrá que esperar después de tomarse un litro de vino para poder conducir sin infringir la ley? ¿Es fiable esta predicción?

Para construir la recta de regresión:

1) Seleccionar el menú `Teaching`▸`Regresión`▸`Regresión lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable tiempo en el campo Variable dependiente y la variable alcohol en el campo Variable independiente.
3) Seleccionar Guardar el modelo, introducir un nombre para el modelo y hacer clic sobre el botón Enviar.

Para hacer la predicción:

1) Seleccionar el menú `Teaching`▸`Regresión`▸`Predicciones`.
2) En el cuadro de diálogo que aparece seleccionar como modelo de regresión la recta calculada e introducir los valores para los que se desea la predicción en el campo Predicciones para y hacer clic sobre el botón Enviar.

4. El conjunto de datos edad.estatura del paquete rk.Teaching contine la edad y la estatura de 30 personas. Se pide:

a) Cargar datos del conjunto de datos edad.estatura desde el paquete rk.Teaching.

b) Calcular la recta de regresión de la estatura sobre la edad. ¿Es un buen modelo la recta de regresión?

1) Seleccionar el menú `Teaching▸Regresión▸Regresión lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable estatura en el campo Variable dependiente y la variable edad en el campo Variable independiente, y hacer clic en el botón Enviar.

c) Dibujar el diagrama de dispersión de la estatura sobre la edad. ¿Alrededor de qué edad se observa un cambio en la tendencia?

1) Seleccionar el menú `Teaching▸Gráficos▸Diagrama de Dispersión`.
2) En el cuadro de diálogo que aparece, seleccionar la variable estatura en el campo Variable Y, la variable edad en el campo Variable X, y hacer clic en el botón Enviar.

d) Recodificar la variable edad en dos grupos para mayores y menores de 20 años.

1) Seleccionar el menú `Teaching▸Datos▸Recodificar variable`.
2) En el cuadro de diálogo que aparece seleccionar en el campo Variable a recodificar la variable edad.
3) En el campo Reglas de recodificación introducir

```
lo:20 = "menores"
20:hi = "mayores"
```

4) En el cuadro Guardar nueva variable hacer clic sobre el botón Cambiar.
5) En el cuadro de diálogo que aparece seleccionar como objeto padre la el conjunto de datos edad_estatura y hacer clic sobre el botón Aceptar.
6) Introducir el nombre de la nueva variable grupo.edad y hacer clic sobre el botón Enviar.

e) Calcular la recta de regresión de la estatura sobre la edad para cada grupo de edad. ¿En qué grupo explica mejor la recta de regresión la relación entre la estatura y la edad? Justificar la respuesta.

1) Seleccionar el menú `Teaching▸Regresión▸Regresión lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable estatura en el campo Variable dependiente y la variable edad como Variable independiente.
3) Seleccionar la opición Ajuste por grupos, introducir la variable grupo.edad en el campo Variable de agrupación, y hacer clic en el Enviar.

f) Dibujar las rectas de regresión anteriores.

1) Seleccionar el menú `Teaching▸Gráficos▸Diagrama de Dispersión`.
2) En el cuadro de diálogo que aparece, seleccionar la variable estatura en el campo Variable Y y la variable edad en el campo Variable X.
3) Seleccionar la opción Dibujar por grupos e introducir la variable grupo.edad en el campo Variable de agrupación.
4) En la solapa `Línea de ajuste`, seleccionar Lineal y hacer clic en el botón Enviar.

g) ¿Qué estatura se espera que tenga una persona de 14 años? ¿Y una de 38?

Para predecir la estatura de la persona de 14 años:

1) Seleccionar el menú `Teaching`▸`Regresión`▸`Predicciones`.
2) En el cuadro de diálogo que aparece seleccionar como modelo de regresión la recta calculada para los menores e introducir 14 en el campo Predicciones para y hacer clic sobre el botón Enviar.

para predecir la estatura de la persona de 38 años, repetir lo mismo pero seleccionando la recta de regresión para los mayores e introducidento 38 en el campo Predicciones para.

5. La siguiente tabla recoge la información de las calificaciones obtenidas por un grupo de alumnos en dos asignaturas X e Y.

Alumno	1	2	3	4	5	6	7	8	9	10	11	12
X	NT	AP	SS	SS	AP	AP	SS	NT	SB	SS	AP	AP
Y	SB	SS	AP	SS	AP	NT	SS	NT	NT	AP	AP	NT

Se pide:

a) Crear un conjunto de datos con las variables X e Y e introducir los datos.

b) ¿Existe relación entre las calificaciones de X e Y? Justificar la respuesta.

1) Seleccionar el menú `Teaching`▸`Regresión`▸`Correlación`.
2) En el cuadro de diálogo que aparece seleccionar la variables X e Y en el campo Variables.
3) En la solapa `Opciones de correlación` seleccionar el método de Ro de Spearman y hacer clic sobre el botón Enviar.

3 Ejercicios propuestos

1. Se determina la pérdida de actividad que experimenta un medicamento desde el momento de su fabricación a lo largo del tiempo, obteniéndose el siguiente resultado:

Tiempo (en años)	1	2	3	4	5
Actividad restante (%)	96	84	70	58	52

Se desea calcular:

a) La relación fundamental (recta de regresión) entre actividad restante y tiempo transcurrido.

b) ¿En qué porcentaje disminuye la actividad cada año que pasa?

c) ¿Cuándo tiempo debe pasar para que el fármaco tenga una actividad del 80 %? ¿Cuándo será nula la actividad? ¿Son igualmente fiables estas predicciones?

2. Al realizar un estudio sobre la dosificación de un cierto medicamento, se trataron 6 pacientes con dosis diarias de 2 mg, 7 pacientes con 3 mg y otros 7 pacientes con 4 mg. De los pacientes tratados con 2 mg, 2 curaron al cabo de 5 días, y 4 al cabo de 6 días. De los pacientes tratados con 3 mg diarios, 2 curaron al cabo de 3 días, 4 al cabo de 5 días y 1 al cabo de 6 días. Y de los pacientes tratados con 4 mg diarios, 5 curaron al cabo de 3 días y 2 al cabo de 4 días. Se pide:

a) Calcular la recta de regresión del tiempo de curación con respecto a la dosis suministrada.

b) Calcular el coeficiente de regresión del tiempo de curación con respecto a la dosis e interpretarlo.

c) Calcular el coeficiente de correlación lineal e interpretarlo.

d) Determinar el tiempo esperado de curación para una dosis de 5 mg diarios. ¿Es fiable esta predicción?

e) ¿Qué dosis debe aplicarse si queremos que el paciente tarde 4 días en curarse? ¿Es fiable la predicción?

3. El fichero estaturas.pesos.alumnos del paquete rk.Teaching, contiene la estatura, el peso y el sexo de una muestra de alumnos universitarios. Se pide:

   a) Cargar el conjunto de datos estaturas.pesos.alumnos desde el paquete rk.Teaching.

   b) Calcular la recta de regresión del peso sobre la estatura y dibujarla.

   c) Calcular las rectas de regresión del peso sobre la estatura para cada sexo y dibujarlas.

   d) Calcular los coeficientes de determinación de ambas rectas. ¿Qué recta es mejor modelo? Justificar la respuesta.

   e) ¿Qué peso tendrá un hombre que mida 170 cm? ¿Y una mujer de la misma estatura?

4. El conjunto de datos neonatos del paquete rk.Teaching, contiene información sobre una muestra de 320 recién nacidos en un hospital durante un año que cumplieron el tiempo normal de gestación. Se pide:

   a) Construir la tabla de frecuencias bidimensional del Agpar al minuto de nacer frente a si la madre ha fumado o no durante el embarazo. ¿Qué conclusiones se pueden sacar?

   b) Construir la tabla de frecuencias bidimensional del peso de los recién nacidos frente a la edad de la madre. ¿Qué conclusiones se pueden sacar?

   c) Construir la recta de regresión del peso de los recién nacidos sobre el número de cigarros fumados al día por las madres. ¿Existe una relación lineal fuerte entre el peso y el número de cigarros?

   d) Dibujar la recta de regresión calculada en el apartado anterior. ¿Por qué la recta no se ajusta bien a la nube de puntos?

   e) Calcular y dibujar la recta de regresión del peso de los recién nacidos sobre el número de cigarros fumados al día por las madres en el grupo de las madres que si fumaron durante el embarazo. ¿Es este modelo mejor o pero que la recta de los apartados anteriores?

   Según este modelo, ¿cuánto disminuirá el peso del recién nacido por cada cigarro más diario que fume la madre?

   f) Según el modelo anterior, ¿qué peso tendrá un recién nacido de una madre que ha fumado 5 cigarros diarios durante el embarazo? ¿Y si la madre ha fumado 30 cigarros diarios durante el embarazo? ¿Son fiables estas predicciones?

   g) ¿Existe la misma relación lineal entre el peso de los recién nacidos y el número de cigarros fumados al día por las madres que fumaron durante el embarazo en el grupo de las madres menores de 20 y en el grupo de las madres mayores de 20? ¿Qué se puede concluir?

Práctica 5

Regresión no lineal

1 Fundamentos teóricos

La regresión simple tiene por objeto la construcción de un modelo funcional $y = f(x)$ que explique lo mejor posible la relación entre dos variables Y (variable dependiente) y X (variable independiente) medidas en una misma muestra.

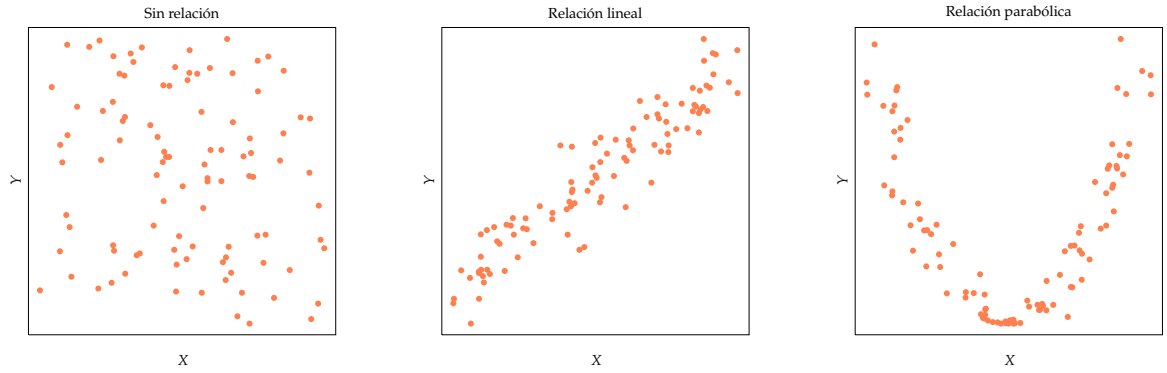
Ya vimos que, dependiendo de la forma de esta función, existen muchos tipos de regresión simple. Entre los más habituales están:

Modelo	Ecuación genérica
Lineal	$y = a + bx$
Parabólico	$y = a + bx + cx^2$
Polinómico de grado n	$y = a_0 + a_1x + \cdots + a_nx^n$
Potencial	$y = ax^b$
Exponencial	$y = e^{a+bx}$
Logarítmico	$y = a + b\log x$
Inverso	$y = a + b/x$
Curva S	$y = e^{a+b/x}$

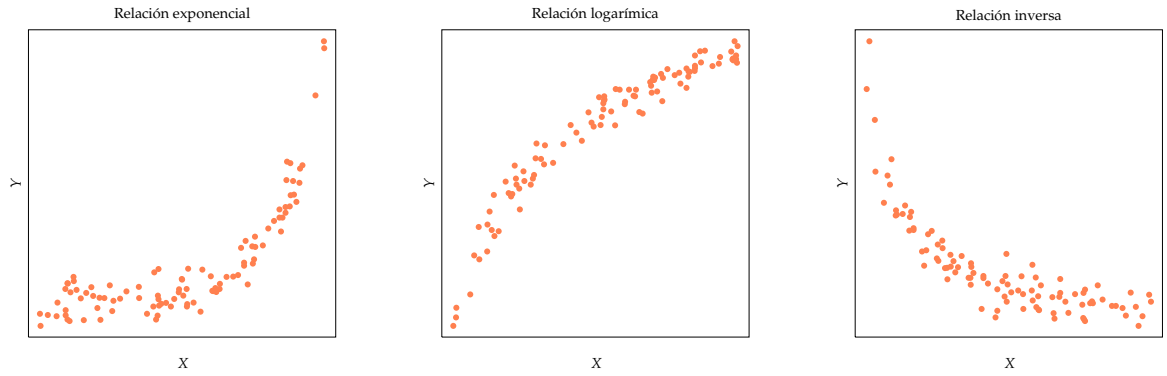
La elección de un tipo de modelo u otro suele hacerse según la forma de la nube de puntos del diagrama de dispersión. A veces estará claro qué tipo de modelo se debe construir, tal y como ocurre en los diagramas de dispersión de la figura 5.1. Pero otras veces no estará tan claro, y en estas ocasiones, lo normal es ajustar los dos o tres modelos que nos parezcan más convincentes, para luego quedarnos con el que mejor explique la relación entre Y y X, mirando el coeficiente de determinación[1] de cada modelo.

Ya vimos en la práctica sobre regresión lineal simple, cómo construir rectas de regresión. En el caso de que optemos por ajustar un modelo no lineal, la construcción del mismo puede realizarse siguiendo los mismos pasos que en el caso lineal. Básicamente se trata de determinar los parámetros del modelo que minimizan la suma de los cuadrados de los residuos en Y. En los modelos multiplicativo y exponencial, el sistema aplica transformaciones logarítmicas a las variables y después ajusta un modelo lineal a los datos transformados. En el modelo recíproco, el sistema sustituye la variable dependiente por su recíproco antes de estimar la ecuación de regresión.

[1] Ver la práctica de regresión lineal y correlación.



(a) Sin relación. (b) Relación lineal. (c) Relación polinómica.



(d) Relación exponencial. (e) Relación logarítmica. (f) Relación inversa.

Figura 5.1 – Diagramas de dispersión correspondientes a distintos tipos de relaciones entre variables.

2 Ejercicios resueltos

El procedimiento más sencillo para construir un modelo no lineal, siempre que sea posible, es transformar las variables para convertirlo en un modelo lineal. En el caso de los modelos de regresión simple más comunes las transformaciones que convierten cada modelo en un modelo lineal aparecen en la tabla siguiente:

Modelo	Modelo no lineal	Modelo lineal	Transformación
Potencial	$y = ax^b$	$\log(y) = \log(a) + b\log(x)$	Se toma el logaritmo de ambas variables
Exponencial	$y = e^{a+bx}$	$\log(y) = a + bx$	Se toma el logaritmo de la variable dependiente
Logarítmico	$y = a + b\log x$	$y = a + b\log x$	Se toma el logaritmo de la variable independiente
Inverso	$y = a + b/x$	$y = a + b\frac{1}{x}$	Se toma el inverso de la variable independiente
Curva S	$y = e^{a+b/x}$	$\log(y) = a + b\frac{1}{x}$	Se toma el logaritmo de la variable dependiente y el inverso de la independiente

1. En un experimento se ha medido el número de bacterias por unidad de volumen en un cultivo, cada hora transcurrida, obteniendo los siguientes resultados:

Horas	0	1	2	3	4	5	6	7	8
Nº Bacterias	25	28	47	65	86	121	190	290	362

Se pide:

a) Crear un conjunto de datos con las variables horas y bacterias e introducir estos datos.

b) Dibujar el diagrama de dispersión correspondiente. En vista del diagrama, ¿qué tipo de modelo crees que explicará mejor la relación entre el número de bacterias y el tiempo transcurrido?

1) Seleccionar el menú `Teaching▸Gráficos▸Diagrama de dispersión`.
2) En el cuadro de diálogo que aparece, seleccionar la variable bacterias en el campo Variable Y y la variable horas en el campo Variable X, y hacer clic en el botón Enviar.

c) Calcular los modelos exponencial y cuadrático de las bacterias sobre las horas. ¿Qué tipo de modelo es el mejor?

Para el modelo exponencial:

1) Seleccionar el menú `Teaching▸Regresión▸Regresión no lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable bacterias en el campo Variable dependiente y la variable horas en el campo Variable independiente.
3) En la solapa de Modelo de regresión seleccionar el modelo Exponencial.
4) Seleccionar Guardar modelo e introducir un nombre para el modelo y hacer clic sobre el botón Enviar.

Para el modelo cuadrático repetir los pasos pero seleccionando como modelo el Cuadrático. El modelo mejor será aquel que tenga un coeficiente de determinación mayor.

d) Dibujar la curva del mejor de los modelos anteriores.

1) Seleccionar el menú `Teaching▸Gráficos▸Diagrama de dispersión`.
2) En el cuadro de diálogo que aparece, seleccionar la variable bacterias en el campo Variable Y y la variable horas en el campo Variable X.
3) En la solapa `Línea de ajuste` seleccionar la opción Exponencial y hacer clic sobre el botón Enviar.

e) Según el modelo anterior, ¿cuántas bacterias habrá al cabo de 3 horas y media del inicio del cultivo? ¿Y al cabo de 10 horas? ¿Son fiables estas predicciones?

1) Seleccionar el menú `Teaching▸Regresión▸Predicciones`.
2) En el cuadro de diálogo que aparece seleccionar el modelo de regresión exponencial construido antes.
3) Introducir los valores 3,5, 10 en el campo Predicciones para y hacer clic sobre el botón Enviar.
4) Como se trata de un modelo exponencial, las predicciones obtenidas corresponden al logaritmo de bacterias. Para obtener la predicción de bacterias basta con aplicar la función exponencial a los valores obtenidos.

f) Dar una predicción lo más fiable posible del tiempo que tendría que transcurrir para que en el cultivo hubiese 100 bacterias.

Para construir el modelo logarítmico:

1) Seleccionar el menú `Teaching▸Regresión▸Regresión no lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable horas en el campo Variable dependiente y la variable bacterias en el campo Variable independiente.
3) Seleccionar como modelo el Logarítmico.
4) Seleccionar Guardar modelo e introducir un nombre para el modelo y hacer clic sobre el botón Enviar.

Para hacer la predicción:

1) Seleccionar el menú `Teaching▸Regresión▸Predicciones`.
2) En el cuadro de diálogo que aparece seleccionar el modelo de regresión logarítmico construido antes.
3) Introducir el valor 100 en el campo Predicciones para y hacer clic sobre el botón Enviar.

2. El conjunto de datos dieta del paquete rk.Teaching contiene los datos de un estudio llevado a cabo por un centro dietético para probar una nueva dieta de adelgazamiento. Para cada individuo se ha medido el número de días que lleva con la dieta, el número de kilos perdidos desde entonces y si realizó o no un programa de ejercicios. Se pide:

a) Cargar el conjunto de datos dieta desde el paquete rk.Teaching.

b) Dibujar el diagrama de dispersión. Según la nube de puntos, ¿qué tipo de modelo explicaría mejor la relación entre los kilos perdidos y los días de dieta?

1) Seleccionar el menú `Teaching▸Gráficos▸Diagrama de dispersión`.
2) En el cuadro de diálogo que aparece, seleccionar la variable kilos en el campo Variable Y, la variable dias en el campo Variable X, y hacer clic en el botón Enviar.

c) Construir el modelo de regresión que mejor explique la relación entre los kilos perdidos y los días de dieta.

1) Seleccionar el menú `Teaching`▸`Regresión`▸`Comparación de modelos`.
2) En el cuadro de diálogo que aparece, seleccionar la variable kilos en el campo Variable dependiente y la variable dias en el campo Variable independiente.
3) En la solapa `Modelos de regresión` seleccionar todos los modelos y hacer clic sobre el botón Enviar.
4) El mejor modelo aparece en primer lugar y es el que tenga el coeficiente de determinación mayor.

d) Dibujar el modelo del apartado anterior.

1) Seleccionar el menú `Teaching`▸`Gráficos`▸`Diagrama de Dispersión`.
2) En el cuadro de diálogo que aparece, seleccionar la variable kilos en el campo Variable Y y la variable dias en el campo Variable X.
3) En la solapa `Línea de ajuste` seleccionar la opción correspondiente al mejor modelo y hacer clic sobre el botón Enviar.

e) Construir el modelo de regresión que mejor explique la relación entre los kilos perdidos y los días de dieta para los que no hacen ejercicio.

Para ver qué modelo es mejor:

1) Seleccionar el menú `Teaching`▸`Regresión`▸`Comparación de modelos`.
2) En el cuadro de diálogo que aparece, seleccionar la variable kilos en el campo Variable dependiente y la variable dias en el campo Variable independiente.
3) Seleccionar la opción Filtro e introducir la condición `ejercicio=="no"` en el campo Condición de selección.
4) En la solapa `Modelos de regresión` seleccionar todos los modelos y hacer clic sobre el botón Enviar.
5) El mejor modelo aparece en primer lugar y es el que tenga el coeficiente de determinación mayor.

Para construir el modelo:

1) Seleccionar el menú `Teaching`▸`Regresión`▸`Regresión no lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable kilos en el campo Variable dependiente y la variable dias en el campo Variable independiente.
3) Seleccionar la opción Filtro e introducir la condición `ejercicio=="no"` en el campo Condición de selección.
4) Seleccionar Guardar modelo e introducir un nombre para el modelo y hacer clic sobre el botón Enviar.

f) Construir el modelo de regresión que mejor explique la relación entre los kilos perdidos y los días de dieta para los que si hacen ejercicio.

Para ver qué modelo es mejor:

1) Seleccionar el menú `Teaching`▸`Regresión`▸`Comparación de modelos`.
2) En el cuadro de diálogo que aparece, seleccionar la variable kilos en el campo Variable dependiente y la variable dias en el campo Variable independiente.
3) Seleccionar la opción Filtro e introducir la condición `ejercicio=="si"` en el campo Condición de selección.
4) En la solapa `Modelos de regresión` seleccionar todos los modelos y hacer clic sobre el botón Enviar.

5) El mejor modelo aparece en primer lugar y es el que tenga el coeficiente de determinación mayor.

Para construir el modelo:

1) Seleccionar el menú `Teaching‣Regresión‣Regresión no lineal`.
2) En el cuadro de diálogo que aparece, seleccionar la variable kilos en el campo Variable dependiente y la variable dias en el campo Variable independiente.
3) Seleccionar la opción Filtro e introducir la condición `ejercicio=="si"` en el campo Condición de selección.
4) Seleccionar Guardar modelo e introducir un nombre para el modelo y hacer clic sobre el botón Enviar.

g) Utilizar el modelo construido para predecir el número de kilos perdidos tras 40 y 500 días de dieta, tanto para los que hacen ejercicio como para los que no. ¿Son fiables estas predicciones?

1) Seleccionar el menú `Teaching‣Regresión‣Predicciones`.
2) En el cuadro de diálogo que aparece seleccionar el modelo de regresión construido antes para los que no hacen ejercicio.
3) Introducir los valores 40, 500 en el campo Predicciones para y hacer clic sobre el botón Enviar.

Repetir los pasos anteriores seleccionando el modelo de regresión construido antes para los que si hacen ejercicio.

3 Ejercicios propuestos

1. La concentración de un fármaco en sangre, C en mg/dl, es función del tiempo, t en horas, y viene dada por la siguiente tabla:

t	2	3	4	5	6	7	8
C	25	36	48	64	86	114	168

Se pide:

a) Según el modelo exponencial, ¿qué concentración de fármaco habría a las 4,8 horas? ¿Es fiable la predicción? Justificar adecuadamente la respuesta.

b) Según el modelo logarítmico, ¿qué tiempo debe pasar para que la concentración sea de 100 mg/dl?

2. El fichero `naciones.txt` contiene información sobre el desarrollo de distintos países (tasa de fertilidad, tasa de uso de anticonceptivos, tasa de mortalidad infantil, producto interior bruto per cápita y continente). Se pide:

a) Importar el fichero `naciones.txt` en un conjunto de datos.

b) Construir el mejor modelo de regresión de la tasa de fertilidad sobre el producto interior bruto. ¿Cómo explicarías esta relación?

c) Dibujar el modelo del apartado anterior.

d) ¿Qué tasa de fertilidad le corresponde a una mujer que viva en un país con un producto interior bruto per cápita de 10000 $? ¿Y si la mujer vive en Europa?