

Lección 1 Logística de R

R es un entorno de programación para el análisis estadístico y gráfico de datos muy popular, cada día más utilizado en empresas y universidades. Su uso tiene muchas ventajas. Para empezar, es *software* libre. La elección de *software* libre es, en general, acertada por varios motivos. Por un lado, transmite valores socialmente positivos, como por ejemplo la libertad individual, el conocimiento compartido, la solidaridad y la cooperación. Por otro, nos aproxima al método científico, porque permite el examen y mejora del código desarrollado por otros usuarios y la reproducibilidad de los resultados obtenidos. Finalmente, pero no menos importante desde un punto de vista práctico, podemos adquirir de manera legal y gratuita copias del programa, sin necesidad de licencias personales o académicas.

Aparte de su faceta de *software* libre, R tiene algunas ventajas específicas: por ejemplo, su sintaxis básica es sencilla e intuitiva, con la que es muy fácil familiarizarse, lo que se traduce en un aprendizaje rápido y cómodo. Además, tiene una enorme comunidad de usuarios, estructurada alrededor de la *Comprehensive R Archive Network*, o *CRAN*, que desarrolla cada día nuevos paquetes que extienden sus funcionalidades y cubren casi todas las necesidades computacionales y estadísticas de un científico o ingeniero. Para que os hagáis una idea, en el momento de revisar estas notas (septiembre de 2019) el número de paquetes en el repositorio de la CRAN acaba de superar los 15000.

1.1 Cómo instalar R y *RStudio*

Instalar R es muy sencillo; de hecho, seguramente ya lo tenéis instalado en vuestro ordenador, pero es conveniente que dispongáis de su versión más reciente y que regularmente lo pongáis al día. Los pasos a realizar en Windows o Mac OS X para instalar su última versión son los siguientes:

- Si sois usuarios de Windows, acceded a la página web de la [CRAN](#) y pulsad sobre el enlace *Download R for Windows*. A continuación, entrad en el enlace *base*, descargad R y seguid las instrucciones de instalación del documento *Installation and other*

instructions que encontraréis en esa misma página.

- Si sois usuarios de Mac OS X, acceded a la página web de la [CRAN](#) y pulsad sobre el enlace *Download R for Mac OS X*. A continuación, descargad el fichero `.pkg` correspondiente y, una vez descargado, abridlo y seguid las instrucciones del Asistente de Instalación.
- Si trabajáis con Ubuntu o Debian, para instalar la última versión de R basta que ejecutéis en una terminal, estando conectados a Internet, la siguiente instrucción:

```
sudo aptitude install r-base
```

Cuando instaláis R para Windows o Mac OS X, con él también se os instala una interfaz gráfica que se abrirá al abrir la aplicación y en la que podréis trabajar. La instalación para Linux no lleva una interfaz por defecto, así que sus usuarios tienen que trabajar con R en la terminal (ejecutando R para iniciar una sesión) o instalar aparte una interfaz.

Independientemente de todas estas posibilidades, en este curso usaremos *RStudio* como interfaz gráfica de usuario de R para todos los sistemas operativos.

Propiamente hablando, *RStudio* es mucho más que una interfaz de R: se trata de todo un entorno integrado para utilizar y programar con R, que dispone de un conjunto de herramientas que facilitan el trabajo con este lenguaje. Para instalarlo, se ha de descargar del url <http://www.rstudio.com/products/rstudio/download/> la versión correspondiente al sistema operativo en el que se trabaja; en cada caso, escoged la versión gratuita de *RStudio Desktop*. Una vez descargado, si usáis Windows o Mac OS X ya lo podéis abrir directamente. En el caso de Linux, hay que ejecutar en una terminal la siguiente instrucción para completar su instalación:

```
sudo dpkg -i rstudio-<version>-i386.deb
```

donde `version` refiere a la versión concreta que hayáis descargado. Conviene recordar que *RStudio* no es R, ni tan sólo lo contiene: hay que instalar ambos programas. De hecho, las instalaciones de R y *RStudio* son independientes una de la otra, de manera que cuando se pone al día uno de estos programas, no se modifica el otro.

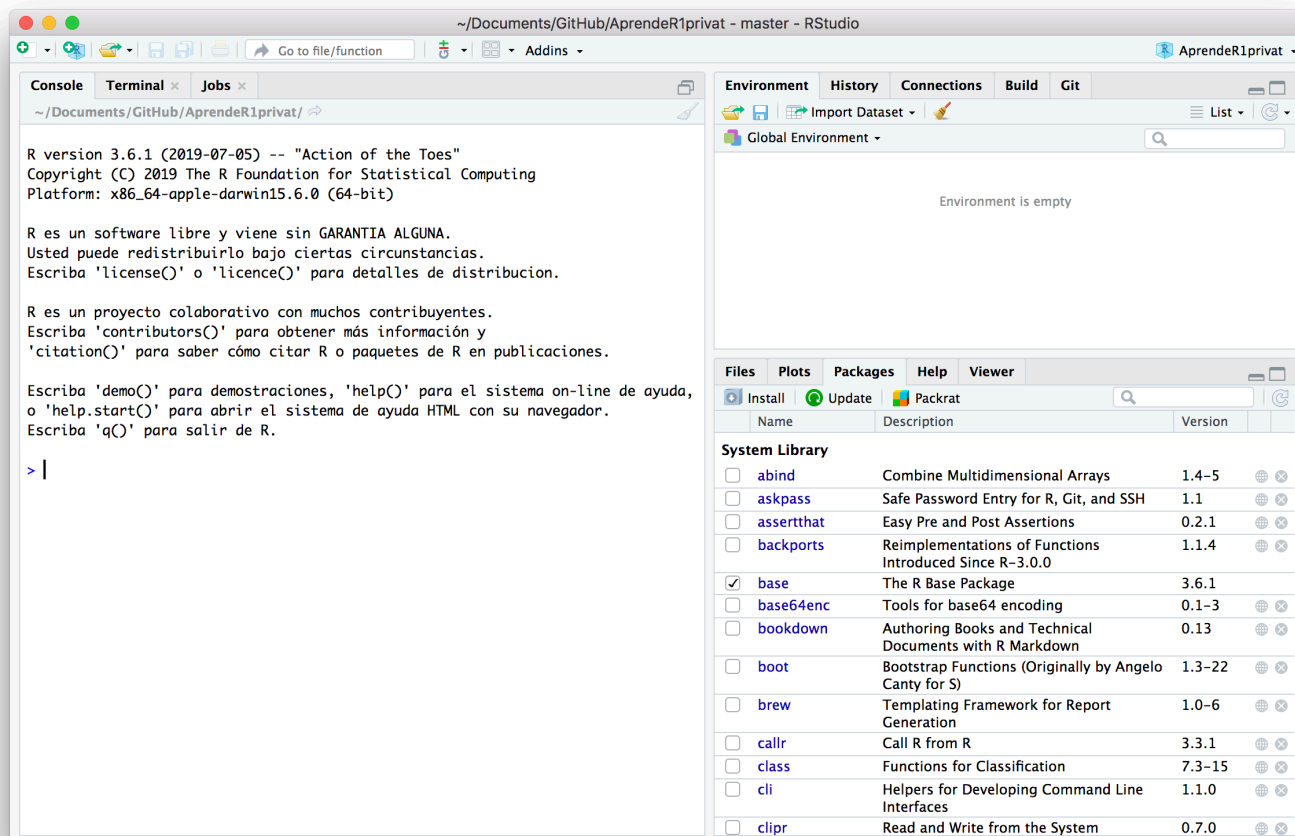


Figura 1.1: Ventana de RStudio para Mac OS X.

Cuando se abre *RStudio*, aparece una ventana similar a la que muestra la Figura 1.1: su apariencia exacta dependerá del sistema operativo, de la versión de *RStudio* e incluso de los paquetes que estemos usando. De momento, nos concentraremos en la ventana de la izquierda, llamada la **consola** de R (la pestaña *Console*). Observaréis que en el momento de abrir la aplicación, dicha ventana contiene una serie de información (versión, créditos etc.) y al final una línea en blanco encabezada por el símbolo `>`. Este símbolo es la **marca de inicio** e indica que R espera que escribáis alguna instrucción y la ejecutéis.

Durante la mayor parte de este curso, usaremos *RStudio* de manera interactiva:

1. Escribiremos una instrucción en la consola, a la derecha de la marca de inicio de su última línea.
2. La ejecutaremos pulsando la tecla *Entrar* (\leftarrow).
3. R la evaluará y, si corresponde, escribirá el resultado en la línea siguiente de la consola (como veremos, no todas las instrucciones hacen que R escriba algo).
4. R abrirá una nueva línea en blanco encabezada por una marca de inicio, donde esperará una nueva instrucción.

Haced una prueba: escribid `1+1` junto a la marca de inicio y pulsad *Entrar*; R escribirá en la línea siguiente el resultado de la suma, `2`, y a continuación una nueva línea en blanco encabezada por la marca de inicio. Ya hablaremos en la Lección ?? del `[1]` que os habrá aparecido delante del 2 en el resultado. Hasta entonces, no os preocupéis por él. En los

bloques de código de este libro no incluimos la marca de inicio, para que podáis copiar tranquilamente el código y luego pegarlo y ejecutarlo en vuestra consola, y el resultado aparece precedido de `##`, para que si por descuido copiáis un resultado, no se ejecute: el símbolo `#` sirve para indicar a R que no ejecute lo que venga a continuación en la misma línea. Así, en este libro el cálculo anterior corresponde a:

```
1+1
```

```
## [1] 2
```

Para facilitarnos el trabajo, la consola dispone de un mecanismo para acceder a las instrucciones ya ejecutadas y modificarlas si queremos. Si situamos el cursor a la derecha de la marca de inicio de la línea inferior y pulsamos la tecla de la flecha vertical ascendente \uparrow , iremos obteniendo de manera consecutiva, en esa línea, las instrucciones escritas hasta el momento en la misma sesión; si nos pasamos, podemos usar la tecla \downarrow para retroceder dentro de esta lista; una vez alcanzada la instrucción deseada, podemos volver a ejecutarla o, con las teclas de flechas horizontales, ir al lugar de la instrucción que queramos y reescribir un trozo antes de ejecutarla. Otra posibilidad es usar la pestaña *History* de la ventana superior derecha de *RStudio*, que contiene la lista de todas las instrucciones que se han ejecutado en la sesión actual. Si seleccionamos una instrucción de esta lista y pulsamos el botón *To console* del menú superior de la pestaña, la instrucción se copiará en la consola y la podremos modificar o ejecutar directamente.

También podemos copiar instrucciones de otros ficheros y pegarlas a la derecha de la marca de inicio de la manera habitual en el sistema operativo de nuestro ordenador. Pero hay que ir con cuidado: las instrucciones copiadas de ficheros en formato que no sea texto simple pueden contener caracteres invisibles a simple vista que generen errores al intentar ejecutar la instrucción copiada. En particular, esto afecta a las instrucciones que podáis copiar de ficheros en formato PDF, procurad no hacerlo. En cambio, no hay ningún problema en copiar y pegar instrucciones de ficheros html como los de estas lecciones.

Volvamos a la ventana de *RStudio* de la Figura 1.1. Observaréis que está dividida a su vez en tres ventanas. La de la izquierda es la consola, donde trabajamos en modo interactivo. La ventana inferior derecha tiene algunas pestañas, entre las que destacamos:

- *Files*, que muestra el contenido de la carpeta de trabajo actual (véase la Sección 1.2).

Al hacer clic sobre un fichero en esta lista, se abrirá en la ventana de ficheros (véase la Sección 1.3).

- *Plots*, que muestra los gráficos que hayamos producido durante la sesión. Se puede navegar entre ellos con las flechas de la barra superior de la pestaña.
- *Packages*, que muestra todos los paquetes instalados y, marcados, los que están cargados en la sesión actual (véase la Sección 1.5).
- *Help*, donde aparecerá la ayuda que pidamos (véase la Sección 1.4).

Por lo que se refiere a la ventana superior izquierda, destacamos las dos pestañas siguientes:

- *Environment*, con la lista de los objetos actualmente definidos (véase la Lección 2).
- *History*, de la que ya hemos hablado, que contiene la lista de todas las instrucciones que hayamos ejecutado durante la sesión.

Aparte de estas tres ventanas, *RStudio* dispone de una cuarta ventana para ficheros, que se abre en el sector superior izquierdo, sobre la consola (véase la Sección 1.3).

Para cerrar *RStudio*, basta elegir *Quit RStudio* del menú *RStudio* o pulsar la combinación de teclas usual para cerrar un programa en vuestro sistema operativo.

1.2 Cómo guardar el trabajo realizado

Antes de empezar a utilizar R en serio, lo primero que tenéis que hacer es crear en vuestro ordenador una carpeta específica que será vuestra **carpeta de trabajo** con R. A continuación, en las *Preferencias* de *RStudio*, que podréis abrir desde el menú *RStudio*, tenéis que declarar esta carpeta como *Default working directory*. A partir de este momento, por defecto, todo el trabajo que realicéis quedará guardado dentro de esta carpeta, y *RStudio* buscará dentro de esta carpeta todo lo que queráis que lea. Si en un momento determinado queréis cambiar temporalmente de carpeta de trabajo, tenéis dos opciones:

- Podéis usar el menú *Session* → *Set Working Directory* → *Choose Directory...* para escoger una carpeta.
- Podéis abrir la pestaña *Files* de la ventana inferior derecha y navegar por el árbol de directorios que aparece en su barra superior hasta llegar a la carpeta deseada.

Tanto de una manera como de la otra, la carpeta que especifiquéis será la carpeta de trabajo durante lo que queda de sesión o hasta que la volváis a cambiar.

En cualquier momento podéis guardar la sesión en la que estéis trabajando usando el menú *Session* → *Save Workspace as....* Además, si no habéis modificado esta opción en las *Preferencias*, cuando cerréis *RStudio* se os pedirá si queréis guardar la sesión; si contestáis que sí, *RStudio* guardará en la carpeta de trabajo dos ficheros, `.RData` y `.RHistory`, que se cargarán automáticamente al volver a abrir *RStudio* y estaréis exactamente donde lo habíais dejado. Nuestro consejo es que digáis que no: normalmente, no os interesará arrastrar todo lo que hayáis hecho en sesiones anteriores. Y si queréis guardar algunas definiciones e instrucciones de una sesión, lo más práctico es guardarlas en un *guión* (véase la Sección [1.3](#)).

Los gráficos que generéis con *RStudio* aparecerán en la ventana inferior derecha, en la pestaña *Plots* que se activa automáticamente cuando se crea alguno. La apariencia del gráfico dependerá de las dimensiones de esta ventana, por lo que es conveniente que sea cuadrada si queréis que el gráfico no aparezca achatado o estirado. Si modificáis la forma de la ventana, las dimensiones del gráfico que aparezca en ella se modificarán de manera automática.

Para guardar un gráfico, hay que ir al menú *Export* de esta pestaña y seleccionar cómo queréis guardarlo: como una imagen en uno de los formatos estándares de imágenes (.png, .jpeg, .tiff, etc.) o en formato PDF. Entonces, se abrirá una ventana donde podéis darle nombre, modificar sus dimensiones y especificar el directorio donde queráis que se guarde, entre otras opciones.

1.3 Cómo trabajar con guiones y otros ficheros

R admite la posibilidad de crear y usar ficheros de instrucciones que se pueden ejecutar y guardar llamados **guiones** (*scripts*). Estos guiones son una alternativa muy cómoda a las sesiones interactivas, porque permiten guardar las versiones finales de las instrucciones usadas, y no toda la sesión con pruebas, errores y resultados provisionales, y facilitan la ejecución de secuencias de instrucciones en un solo paso. Además, un guión se puede guardar, volver a abrir más adelante, editar, etc. Como ya hemos comentado, el símbolo `#` sirve para indicar a R que omita todo lo que hay a su derecha en la misma línea, lo que permite añadir comentarios a un guión.

Para crear un guión con *RStudio*, tenéis que ir al menú *File* → *New File* → *R Script*. Veréis que os aparece una ventana nueva en el sector superior izquierdo de la ventana de *RStudio*, sobre la consola: la llamaremos **ventana de ficheros**. En ella podéis escribir, línea a línea, las instrucciones que queráis. Para ejecutar instrucciones de esta ventana, basta que las seleccionéis y pulséis el botón *Run* que aparece en la barra superior de esta ventana.

Para guardar un guión, basta pulsar el botón con el icono de un disquete de ordenador que aparece en la barra superior de su ventana. Otra posibilidad es usar el menú *File* → *Save*, o pulsar la combinación de teclas usual para guardar un fichero en vuestro sistema operativo, siempre y cuando la **ventana activa** de *RStudio* (donde esté activo el cursor en ese momento) sea la del guión. Al guardar un guión por primera vez, se abre una ventana de diálogo donde *RStudio* espera que le demos un nombre; la costumbre es usar para los guiones la extensión `.R`.

Podéis abrir un guión ya preexistente con *RStudio* usando el menú *File* → *Open File* de *RStudio* o pulsando sobre él en la pestaña *Files*. También podéis arrastrar el icono del guión sobre el de *RStudio* o (si habéis declarado que la aplicación por defecto para abrir ficheros con extensión `.R` sea *RStudio*) simplemente abrir el fichero de la manera usual en vuestro sistema operativo.

Además de guiones, con *RStudio* también podemos crear otros tipos de ficheros que combinen instrucciones de R con instrucciones de otro lenguaje. En este curso lo usaremos para crear ficheros *R Markdown*, que permiten generar de manera muy cómoda informes y presentaciones que incorporen instrucciones de R (o sólo sus resultados). Para crear un fichero *R Markdown*, tenéis que ir al menú *File* → *New File* → *R Markdown...*, donde os aparecerá una ventana que os pedirá el tipo de documento (*Document*, *Presentation...*), su título y el formato de salida. Una vez completada esta información, se abrirá el fichero en la ventana superior izquierda.

Por poner un ejemplo, supongamos que habéis elegido realizar un informe (*Document*) con formato de salida html (los formatos posibles son: PDF, HTML o Word); entonces, para generar un informe básico basta sustituir las palabras clave que ha generado *RStudio* en esta ventana. Probadlo: cambiad el título y el texto; a continuación, guardad el fichero con el nombre que queráis y extensión `.Rmd`, y pulsad el botón *Knit* situado en la barra

superior de la ventana; se generará un fichero HTML en la carpeta de trabajo, con el texto del fichero *R Markdown* y el mismo título cambiando la extensión `.Rmd` por `.html`, y se abrirá en una ventana aparte.

Aprender los primeros pasos de *R Markdown* es sencillo. Para ello, podéis consultar el manual de referencia rápida de *R Markdown* que encontraréis en el url <https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf>, que se puede leer en 15 minutos y que para la mayoría de ejercicios de este curso es más que suficiente. También os puede ser útiles las “chuletas” de *R Markdown* siguientes:

- <https://github.com/rstudio/cheatsheets/raw/master/rmarkdown-2.0.pdf>
- <https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf>

En la Lección **??**, al final de esta primera parte, explicamos algunas técnicas para mejorar los ficheros resultantes.

1.4 Cómo obtener ayuda

Para conocer toda la información (qué hace, cuál es la sintaxis correcta, qué parámetros tiene, algunos ejemplos de uso...) sobre una función o un objeto, se puede usar el campo de búsqueda, marcado con una lupa, en la esquina superior derecha de la pestaña de **Ayuda** (*Help*), situada en la ventana inferior derecha de *RStudio*. Como alternativa, se pueden usar las instrucciones

```
help(nombre del objeto)
```

o, equivalentemente,

```
?nombre del objeto
```

Por ejemplo, si entramos en el campo de búsqueda de la pestaña de Ayuda la palabra `sum`, o si **entramos** en la consola (es decir, si escribimos a la derecha de la marca de inicio y a continuación pulsamos la tecla *Entrar*) la instrucción

```
help(sum)
```


obtenemos en la pestaña de Ayuda toda la información sobre la función `sum`.

Cuando hayamos avanzado un poco en este curso, la Ayuda os será muy útil. Aquí sólo veremos alguna aplicación simple de la mayoría de las funciones que estudiemos, con los parámetros más importantes y suficientes para nuestros propósitos, y necesitaréis consultar su Ayuda para conocer todos sus usos, todos sus parámetros u otra información relevante.

Si queremos pedir ayuda sobre un tema concreto, pero no sabemos el nombre exacto de la función, podemos entrar una palabra clave en el campo de búsqueda de la pestaña de Ayuda, o usar la función

```
help.search("palabra clave")
```

o, equivalentemente,

```
??"palabra clave"
```

(las comillas ahora son obligatorias). De esta manera, conseguiremos en la ventana de Ayuda una lista de las funciones que R entiende que están relacionadas con la palabra clave entrada. Entonces, pulsando en la función que nos interese de esta lista, aparecerá la información específica sobre ella. Como podéis imaginar, conviene que la palabra clave esté en inglés.

Además de la ayuda que incorpora el mismo R, siempre podéis acudir a foros y listas de discusión para encontrar ayuda sobre cualquier duda que podáis tener. Algunos recursos que nosotros encontramos especialmente útiles son los siguientes:

- La sección dedicada a R del foro [stackoverflow](#)
- El archivo de la lista de discusión [R-help](#)
- El grupo de Facebook [R project en español](#)

Si tenéis alguna dificultad, es muy probable que alguien ya la haya tenido y se la hayan resuelto en alguno de estos foros.

Existe también una comunidad muy activa de usuarios hispanos de R, en cuyo [portal web](#) encontraréis muchos recursos útiles para mejorar vuestro conocimiento de este lenguaje.

1.5 Cómo instalar y cargar paquetes

Muchas funciones y tablas de datos útiles no vienen con la instalación básica de R, sino que forman parte de **paquetes** (*packages*), que se tienen que instalar y cargar para poderlos usar. Por citar un par de ejemplos, el paquete **magic** lleva una función `magic` que crea **cuadrados mágicos** (tablas cuadradas de números naturales diferentes dos a dos tales que las sumas de todas sus columnas, de todas sus filas y de sus dos diagonales principales valgan todas lo mismo), y para usarla tenemos que instalar y cargar este paquete. De manera similar, el paquete **ggplot2** incorpora una serie de funciones para dibujar gráficos avanzados que no podemos usar si primero no instalamos y cargamos este paquete.

Podemos consultar en la pestaña *Packages* la lista de paquetes que tenemos instalados. Los paquetes que aparecen marcados en esta lista son los que tenemos cargados en la sesión actual. Si queremos cargar un paquete ya instalado, basta marcarlo en esta lista; podemos hacerlo también desde la consola, con la instrucción

```
library(nombre del paquete)
```

En caso de necesitar un paquete que no tengamos instalado, hay que instalarlo antes de poderlo cargar. La mayoría de los paquetes se pueden instalar desde el repositorio del CRAN; esto se puede hacer de dos maneras:

- Desde la consola, entrando la instrucción

```
install.packages("nombre del paquete", dep=TRUE)
```

(las comillas son obligatorias, y fijaos en el plural de `packages` aunque sólo queráis instalar uno). El parámetro `dep=TRUE` hace que R instale no sólo el paquete requerido, sino todos aquellos de los que dependa para funcionar correctamente.

- Pulsando el botón *Install* de la barra superior de la pestaña de paquetes; al hacerlo, *RStudio* abre una ventana dónde se nos pide el nombre del paquete a instalar. Conviene dejar marcada la opción *Install dependencies*, para que se instalen también los paquetes necesarios para su funcionamiento.

Así, supongamos que queremos construir cuadrados mágicos, pero aún no hemos cargado el paquete `magic`.

```
magic(10)
```

```
## Error in magic(10): no se pudo encontrar la función "magic"
```

Así que instalamos y cargamos dicho paquete (también lo podríamos hacer desde la ventana *Packages*):

```
install.packages("magic", dep=TRUE)  
library(magic)
```

Ahora ya podemos usar la función `magic`:

```
magic(10)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]  
## [1,]   34   35    6    7   98   99   70   71   42   43  
## [2,]   36   33    8    5  100   97   72   69   44   41  
## [3,]   11   10   83   82   75   74   47   46   39   38  
## [4,]   12    9   84   81   73   76   48   45   40   37  
## [5,]   87   86   79   78   51   50   23   22   15   14  
## [6,]   85   88   77   80   52   49   21   24   13   16  
## [7,]   63   62   55   54   27   26   19   18   91   90  
## [8,]   61   64   53   56   25   28   17   20   89   92  
## [9,]   59   58   31   30    3    2   95   94   67   66  
## [10,]   57   60   29   32    1    4   93   96   65   68
```

Cuando cerramos *RStudio*, los paquetes instalados en la sesión siguen instalados, pero los cargados se pierden; por lo tanto, si queremos volver a usarlos en otra sesión, tendremos que volver a cargarlos.

Hay paquetes que no se encuentran en el CRAN y que, por lo tanto, no se pueden instalar de la forma que hemos visto. Cuando sea necesario, ya explicaremos la manera de instalarlos y cargarlos en cada caso.

Para terminar, observad que a la derecha del nombre de cada paquete en la pestaña *Packages* aparecen dos símbolos. Al pulsar el primero, seis puntitos, se abrirá en el navegador la página de información del paquete, y al pulsar el segundo, una crucecita, desinstalamos el paquete. Asimismo, en la barra superior de la pestaña *Packages* encontraréis un botón *Update* que sirve para poner al día los paquetes instalados, obteniendo sus últimas versiones publicadas.

1.6 Guía rápida

- `help` o `?` permiten pedir información sobre una función. También se puede usar el campo de búsqueda de la pestaña *Help* en la ventana inferior derecha de *RStudio*.
- `help.search` o `??` permiten pedir información sobre una palabra clave (entrada entre comillas). De nuevo, también se puede usar el campo de búsqueda de la pestaña *Help* en la ventana inferior derecha de *RStudio*.
- `install.packages("paquete", dep=TRUE)` instala el `paquete` y todos los otros paquetes de los que dependa. También se puede usar el botón *Install* de la pestaña *Packages* en la ventana inferior derecha de *RStudio*.
- `library(paquete)` carga el `paquete`. También se puede cargar marcándolo en la ventana *Packages* de *RStudio*.

Lección 2 La calculadora

Cuando se trabaja en modo interactivo en la consola de R, hay que escribir las instrucciones a la derecha de la marca de inicio `>` de la línea inferior (que omitimos en los bloques de código de este libro). Para evaluar una instrucción al terminar de escribirla, se tiene que pulsar la tecla *Entrar* (\leftarrow); así, por ejemplo, si junto a la marca de inicio escribimos `2+3` y pulsamos *Entrar*, R escribirá en la línea siguiente el resultado, 5, y a continuación una nueva línea en blanco encabezada por la marca de inicio, donde podremos continuar entrando instrucciones.

```
2+3 #Y ahora aquí pulsamos Entrar
```

```
## [1] 5
```

Bueno, hemos hecho trampa. Como ya habíamos comentado en la lección anterior, se pueden escribir comentarios: R ignora todo lo que se escribe en la línea después de un signo `#`. También podéis observar que R ha dado el resultado en una línea que empieza con `[1]`; ya discutiremos en la Lección ?? qué significa este `[1]`.

Si la expresión que entramos no está completa, R no la evaluará y en la línea siguiente esperará a que la acabemos, indicándolo con la **marca de continuación**, por defecto un signo `+`. (En estas notas, y excepto en el ejemplo que damos a continuación, no mostraremos este signo `+` para no confundirlo con una suma.) Además, si cometemos algún error de sintaxis, R nos avisará con un mensaje de error.

```
2*(3+5 #Pulsamos Entrar, pero no hemos acabado  
+ ) #ahora sí
```

```
## [1] 16
```

```
2*3+5)
```

```
## Error: <text>:1:6: inesperado ' )'  
## 1: 2*3+5)  
##           ^
```

Como podemos ver, al ejecutar la segunda instrucción, R nos avisa de que el paréntesis no está en su sitio.

Se puede agrupar más de una instrucción en una sola línea separándolas con signos de punto y coma. Al pulsar la tecla *Entrar*, R las ejecutará todas, una tras otra y en el orden en el que las hayamos escrito.

```
2+3; 2+4; 2+5
```

```
## [1] 5
```

```
## [1] 6
```

```
## [1] 7
```

2.1 Números reales: operaciones y funciones básicas

La separación entre la parte entera y la parte decimal en los números reales se indica con un punto, no con una coma. Por consistencia, en el texto también seguiremos el convenio angloamericano de usar un punto en lugar de una coma como separador decimal.

```
2+2,5
```



```
## Error: <text>:1:4: inesperado ',',
```

```
## 1: 2+2,
```

```
##      ^
```

```
2+2.5
```

```
## [1] 4.5
```

Las operaciones usuales se indican en R con los signos que damos en la lista siguiente. Por lo que se refiere a los dos últimos operadores en esta lista, recordad que si a y b son dos números reales, con $b > 0$, la **división entera** de a por b da como **cociente entero** el mayor número entero q tal que $q \cdot b \leq a$, y como **resto** la diferencia $a - q \cdot b$. Por ejemplo, la división entera de 29.5 entre 6.3 es $29.5 = 4 \cdot 6.3 + 4.3$, con cociente entero 4 y resto 4.3. (Cuando $b < 0$, R da como cociente entero el menor número entero q tal que $q \cdot b \geq a$, y como resto la diferencia $a - q \cdot b$, que en este caso es negativa.)

- **Suma:** `+`
- **Resta:** `-`
- **Multiplicación:** `*`
- **División:** `/`
- **Potencia:** `^`
- **Cociente entero:** `%%`
- **Resto de la división entera:** `%%`

A continuación, damos algunos ejemplos de manejo de estas operaciones. Observad el uso natural de los paréntesis para indicar la precedencia de las operaciones.

```
2*3+5/2
```

```
## [1] 8.5
```

```
2*(3+5/2) #Aquí lo único que dividimos entre 2 es 5
```

```
## [1] 11
```

```
2*((3+5)/2)
```

```
## [1] 8
```

```
2/3+4 #Aquí el denominador de la fracción es 3
```

```
## [1] 4.666667
```

```
2/(3+4)
```

```
## [1] 0.2857143
```

```
2^3*5 #Aquí el exponente es 3
```

```
## [1] 40
```

```
2^(3*5)
```

```
## [1] 32768
```

```
2^-5 #En este caso no hacen falta paréntesis...
```

```
## [1] 0.03125
```

```
2^(-5) #Pero queda más claro si se usan
```

```
## [1] 0.03125
```

```
534%/7 #¿Cuántas semanas completas caben en 534 días?
```

```
## [1] 76
```

```
534%%7 #¿Y cuántos días sobran?
```

```
## [1] 2
```

```
534-76*7
```

```
## [1] 2
```

El objeto `pi` representa el número real π .

```
pi
```

```
## [1] 3.141593
```

¡Cuidado! No podemos omitir el signo `*` en las multiplicaciones.

```
2(3+5)
```

```
## Error in eval(expr, envir, enclos): tentativa de aplicar una no-función
```

```
2*(3+5)
```

```
## [1] 16
```

```
2pi
```

```
## Error: <text>:1:2: unexpected symbol
```

```
## 1: 2pi
```

```
##      ^
```

```
2*pi
```

```
## [1] 6.283185
```

Cuando un número es muy grande o muy pequeño, R emplea la llamada **notación científica** para dar una aproximación.

```
2^40
```

```
## [1] 1.099512e+12
```

```
2^(-20)
```

```
## [1] 9.536743e-07
```

En este ejemplo, `1.099512e+12` representa el número $1.099512 \cdot 10^{12}$, es decir, 1099512000000, y `9.536743e-07` representa el número $9.536743 \cdot 10^{-7}$, es decir, 0.0000009536743. Como muestra el ejemplo siguiente, no es necesario que un número

sea especialmente grande o pequeño para que R lo escriba en notación científica: basta que esté rodeado de otros números en esa notación.

```
c(2^40,2^(-20),17/3) #La función c sirve para definir vectores
```

```
## [1] 1.099512e+12 9.536743e-07 5.666667e+00
```

Este `5.666667e+00` representa el número $5.666667 \cdot 10^0$, es decir, 5.666667.

R dispone, entre muchas otras, de las funciones numéricas de la lista siguiente:

- **Valor absoluto**, $|x|$: `abs(x)`
- **Raíz cuadrada**, \sqrt{x} : `sqrt(x)`
- **Exponencial**, e^x : `exp(x)`
- **Logaritmo neperiano**, $\ln(x)$: `log(x)`
- **Logaritmo decimal**, $\log_{10}(x)$: `log10(x)`
- **Logaritmo binario**, $\log_2(x)$: `log2(x)`
- **Logaritmo en base a** , $\log_a(x)$: `log(x,a)`
- **Factorial**, $n!$: `factorial(n)`
- **Número combinatorio**, $\binom{n}{m}$: `choose(n,m)`
- **Seno**, $\sin(x)$: `sin(x)`
- **Coseno**, $\cos(x)$: `cos(x)`
- **Tangente**, $\tan(x)$: `tan(x)`
- **Arcoseno**, $\arcsin(x)$: `asin(x)`
- **Arcocoseno**, $\arccos(x)$: `acos(x)`
- **Arcotangente**, $\arctan(x)$: `atan(x)`

Recordad que el **valor absoluto** $|x|$ de un número x se obtiene tomando x sin signo: $|-8| = |8| = 8$. Recordad también que el **factorial** $n!$ de n , es el producto

$$n! = n \cdot (n - 1) \cdot (n - 2) \cdots 3 \cdot 2 \cdot 1$$

(con el convenio de que $0! = 1$), y es igual al número de maneras posibles de ordenar una lista de n objetos diferentes (su número de **permutaciones**), y que el **número combinatorio** $\binom{n}{m}$, con $m \leq n$, es

$$\binom{n}{m} = \frac{n!}{m! \cdot (n - m)!} = \frac{n(n - 1)(n - 2) \cdots (n - m + 1)}{m(m - 1)(m - 2) \cdots 2 \cdot 1},$$

y es igual al número de maneras posibles de escoger un subconjunto de m elementos de un conjunto de n objetos diferentes.

Las funciones de R se aplican a sus argumentos introduciéndolos siempre entre paréntesis. Si la función se tiene que aplicar a más de un argumento, éstos se tienen que especificar en el orden que toque y separándolos mediante comas; R no tiene en cuenta los espacios en blanco alrededor de las comas. Veamos algunos ejemplos:

```
sqrt(4)
```

```
## [1] 2
```

```
sqrt(8)-8^(1/2)
```

```
## [1] 0
```

```
log10(8)
```

```
## [1] 0.90309
```

```
log(8)/log(10)
```



```
## [1] 0.90309
```

```
7^log(2,7) #7 elevado al logaritmo en base 7 de 2 es 2
```

```
## [1] 2
```

```
10! #R no entiende esta expresión
```

```
## Error: <text>:1:3: inesperado '!'
```

```
## 1: 10!
```

```
##      ^
```

```
factorial(10)
```

```
## [1] 3628800
```

```
exp(sqrt(8))
```

```
## [1] 16.91883
```

```
choose(5,3) #Núm. de subconjuntos de 3 elementos de un conjunto de 5
```

```
## [1] 10
```

```
choose(3,5) #Núm. de subconjuntos de 5 elementos de un conjunto de 3
```

```
## [1] 0
```

R entiende que los argumentos de las funciones `sin`, `cos` y `tan` están en radianes. Si queremos aplicar una de estas funciones a un número de grados, podemos traducir los grados a radianes multiplicándolos por $\pi/180$. De manera similar, los resultados de `asin`, `acos` y `atan` también están en radianes, y se pueden traducir a grados multiplicándolos por $180/\pi$.

```
cos(60) #Coseno de 60 radianes
```

```
## [1] -0.952413
```

```
cos(60*pi/180) #Coseno de 60 grados
```

```
## [1] 0.5
```

```
acos(0.5) #Arcocoseno de 0.5 en radianes
```

```
## [1] 1.047198
```

```
acos(0.5)*180/pi #Arcocoseno de 0.5 en grados
```

```
## [1] 60
```

```
acos(2)
```

```
## [1] NaN
```

Este último `NaN` (acrónimo de `Not a Number`) significa que el resultado no existe; en efecto, $\arccos(2)$ no existe como número real, ya que $\cos(x)$ siempre pertenece al intervalo $[-1, 1]$.

Ya hemos visto que R dispone del signo `pi` para representar el número real π . En cambio, no tiene ningún signo para indicar la constante de Euler e , y hay que emplear `exp(1)` .

```
2*exp(1) #2·e
```

```
## [1] 5.436564
```

```
exp(pi)-pi^exp(1) #e^pi-pi^e
```

```
## [1] 0.6815349
```

Para terminar esta sección, observad el resultado siguiente:

```
sqrt(2)^2-2
```

```
## [1] 4.440892e-16
```

R opera numéricamente con $\sqrt{2}$, no formalmente, y por eso no da como resultado de $(\sqrt{2})^2 - 2$ el valor 0 exacto, sino el número pequeñísimo $4.440892 \cdot 10^{-16}$; de hecho, R trabaja internamente con una precisión de aproximadamente 16 cifras decimales, por lo que no siempre podemos esperar resultados exactos. Si necesitáis trabajar de manera exacta con más cifras significativas, os recomendamos usar las funciones del paquete `Rmpfr` .

2.2 Cifras significativas y redondeos

En cada momento, R decide cuántas cifras muestra de un número según el contexto. También podemos especificar este número de cifras para toda una sesión, entrándolo en lugar de los puntos suspensivos en `options(digits=...)` . Hay que tener presente que ejecutar esta instrucción no cambiará la precisión de los cálculos, sólo cómo se muestran los resultados.

Si queremos conocer una cantidad específica n de cifras significativas de un número x , podemos emplear la función

```
print(x, n)
```

Observad su efecto:

```
sqrt(2)
```

```
## [1] 1.414214
```

```
print(sqrt(2), 20)
```

```
## [1] 1.4142135623730951455
```

```
print(sqrt(2), 2)
```

```
## [1] 1.4
```

```
2^100
```

```
## [1] 1.267651e+30
```

```
print(2^100, 15)
```

```
## [1] 1.26765060022823e+30
```

```
print(2^100, 5)
```

```
## [1] 1.2677e+30
```

El número máximo de cifras que podemos pedir con `print` es 22; si pedimos más, R nos dará un mensaje de error.

```
print(sqrt(2), 22)
```

```
## [1] 1.414213562373095145475
```

```
print(sqrt(2), 23)
```

```
## Error in print.default(sqrt(2), 23): argumento 'digits' inválido
```

Por otro lado, hay que tener en cuenta que, como ya hemos comentado, R trabaja con una precisión de unas 16 cifras decimales y por lo tanto los dígitos más allá de esta precisión pueden ser incorrectos. Por ejemplo, si le pedimos las 22 primeras cifras de π , obtenemos el resultado siguiente:

```
print(pi, 22)
```

```
## [1] 3.141592653589793115998
```

En cambio, π vale en realidad 3.141592653589793**238462...**, lo que significa que el valor que da R es erróneo a partir de la decimosexta cifra decimal.

La función `print` permite indicar las cifras que queremos *leer*, pero no sirve para especificar las cifras decimales con las que queremos *trabajar*. Para *redondear* un número x a una cantidad específica n de cifras decimales, y trabajar solamente con esas cifras, hay que usar la función

```
round(x, n)
```

La diferencia entre los efectos de `print` y `round` consiste en que `print(sqrt(2), 4)` es igual a $\sqrt{2}$, pero R sólo muestra sus primeras 4 cifras, 1.414, mientras que `round(sqrt(2), 3)` es *igual a* 1.414. Veamos algunos ejemplos

```
print(sqrt(2), 4)
```

```
## [1] 1.414
```

```
print(sqrt(2), 4)^2
```

```
## [1] 1.414
```

```
## [1] 2
```

```
1.414^2
```

```
## [1] 1.999396
```

```
round(sqrt(2), 3)
```



```
## [1] 1.414
```

```
round(sqrt(2), 3)^2
```

```
## [1] 1.999396
```

En caso de empate, R redondea al valor que termina en cifra par, siguiendo la regla de redondeo en caso de empate recomendada por el estándar IEEE 754 para aritmética en coma flotante.

```
round(2.25, 1)
```

```
## [1] 2.2
```

```
round(2.35, 1)
```

```
## [1] 2.4
```

¿Qué pasa si no se indica el número de cifras en el argumento de `round` ?

```
round(sqrt(2))
```

```
## [1] 1
```

```
round(sqrt(2), 0)
```

```
## [1] 1
```

Al entrar `round(sqrt(2))`, R ha entendido que el número de cifras decimales al que queríamos redondear era 0. Esto significa que 0 es el **valor por defecto** de este parámetro. No es necesario especificar los valores por defecto de los parámetros de una función, y para saber cuáles son, hay que consultar su Ayuda. Así, por ejemplo, la Ayuda de `round` indica que su sintaxis es

```
round(x, digits=0)
```

donde el valor de `digits` ha de ser un número entero que indique el número de cifras decimales. Esta sintaxis significa que el valor por defecto del parámetro `digits` es 0.

Escribir `digits=` en el argumento para especificar el número de cifras decimales es optativo, siempre que mantengamos el orden de los argumentos indicado en la Ayuda: en este caso, primero el número y luego las cifras. Este es el motivo por el que podemos escribir `round(sqrt(2), 1)` en lugar de `round(sqrt(2), digits=1)`. Si cambiamos el orden de los argumentos, entonces sí que hay que especificar el nombre del parámetro, como muestra el siguiente ejemplo:

```
round(digits=3, sqrt(2))
```

```
## [1] 1.414
```

```
round(3, sqrt(2))
```

```
## [1] 3
```

En la lista de funciones ya vimos una función de dos argumentos que toma uno por defecto: `log`. Su sintaxis completa es `log(x, base=...)`, y si no especificamos la `base`, toma su valor por defecto, e , y calcula el logaritmo neperiano.

La función `round(x)` redondea x al valor entero más cercano (y en caso de empate, al que termina en cifra par). R también dispone de otras funciones que permiten redondear a números enteros en otros sentidos específicos:

- `floor(x)` redondea x a un número entero **por defecto**, dando el mayor número entero menor o igual que x , que denotamos por $\lfloor x \rfloor$.
- `ceiling(x)` redondea x a un número entero **por exceso**, dando el menor número entero mayor o igual que x , que denotamos por $\lceil x \rceil$.
- `trunc(x)` da la **parte entera** de x , eliminando la parte decimal: es lo que se llama **truncar** x a un entero.

```
floor(8.3) #El mayor entero menor o igual que 8.3
```

```
## [1] 8
```

```
ceiling(8.3) #El menor entero mayor o igual que 8.3
```

```
## [1] 9
```

```
trunc(8.3) #La parte entera de 8.3
```

```
## [1] 8
```

```
round(8.3) #El entero más cercano a 8.3
```

```
## [1] 8
```

```
floor(-3.7) #El mayor entero menor o igual que -3.7
```

```
## [1] -4
```

```
ceiling(-3.7) #El menor entero mayor o igual que -3.7
```

```
## [1] -3
```

```
trunc(-3.7) #La parte entera de -3.7
```

```
## [1] -3
```

```
round(-3.7) #El entero más cercano a -3.7
```

```
## [1] -4
```

2.3 Definición de variables

R funciona mediante **objetos**, estructuras de diferentes tipos que sirven para realizar diferentes tareas. Una **variable** es un tipo de objeto que sirve para guardar datos. Por ejemplo, si queremos crear una variable `x` que contenga el valor π^2 , podemos escribir:

```
x=pi^2
```

Al entrar esta instrucción, R creará el objeto `x` y le asignará el valor que hemos especificado. En general, se puede crear una variable y asignarle un valor, o asignar un nuevo valor a una variable definida anteriormente, mediante la construcción

```
nombre_de_la_variable=valor
```

También se puede conectar el nombre de la variable con el valor por medio de una flecha `->` o `<-`, compuesta de un guión y un signo de desigualdad, de manera que el sentido de la flecha vaya del valor a la variable; por ejemplo, las tres primeras instrucciones

siguientes son equivalentes, y asignan el valor 2 a la variable x , mientras que las dos últimas son incorrectas:

```
x=2
x<-2
2->x
```

```
2=x
```

```
## Error in 2 = x: lado izquierdo de la asignación inválida (do_set)
```

```
2<-x
```

```
## Error in 2 <- x: lado izquierdo de la asignación inválida (do_set)
```

Nosotros usaremos sistemáticamente el signo `=` para hacer asignaciones.

Se puede usar como nombre de una variable cualquier palabra que combine letras mayúsculas y minúsculas (R las distingue), con acentos o sin (aunque os recomendamos que no uséis letras acentuadas, ya que se pueden importar mal de un ordenador a otro), dígitos (0,..., 9), puntos `.` y guiones bajos `_`, siempre que empiece con una letra o un punto. Aunque no esté prohibido, es muy mala idea redefinir nombres que ya sepáis que tienen significado para R, como por ejemplo `pi` o `sqrt`.

Como podéis ver en las instrucciones anteriores y en las que siguen, cuando asignamos un valor a una variable, R no da ningún resultado; después podemos usar el nombre de la variable para referirnos al valor que representa. Es posible asignar varios valores a una misma variable en una misma sesión: naturalmente, en cada momento R empleará el último valor asignado. Incluso se puede redefinir el valor de una variable usando en la nueva definición su valor actual.

```
x=5
x^2
```

```
## [1] 25
```

```
x=x-2 #Redefinimos x como su valor actual menos 2  
x
```

```
## [1] 3
```

```
x^2
```

```
## [1] 9
```

```
x=sqrt(x) #Redefinimos x como la raíz cuadrada de su valor actual  
x
```

```
## [1] 1.732051
```

2.4 Definición de funciones

A menudo queremos definir alguna función. Para ello tenemos que usar, en vez de simplemente `=`, una construcción especial:

```
nombre_de_la_función=function(variables){definición}
```

Una vez definida una función, la podemos aplicar a valores de la variable o variables.

Veamos un ejemplo. Vamos a llamar f a la función $x^2 - 2^x$, usando x como variable, y a continuación la aplicamos a $x = 30$:


```
f=function(x){x^2-2^x}  
f(30)
```

```
## [1] -1073740924
```

Conviene que os acostumbréis a escribir la fórmula que define la función entre llaves `{...}` . A veces es necesario y a veces no, pero no vale la pena discutir cuándo.

El nombre de la variable se indica dentro de los paréntesis que siguen al `function` . En el ejemplo anterior, la variable era x , y por eso hemos escrito `=function(x)` . Si hubiéramos querido definir la función con variable t , habríamos usado `=function(t)` (y, naturalmente, habríamos escrito la fórmula que define la función con la variable t):

```
f=function(t){t^2-2^t}
```

Se pueden definir funciones de dos o más variables con `function` , declarándolas todas. Por ejemplo, para definir la función $f(x, y) = e^{(2x-y)^2}$, tenemos que entrar

```
f=function(x, y){exp((2*x-y)^2)}
```

y ahora ya podemos aplicar esta función a pares de valores:

```
f(0, 1)
```

```
## [1] 2.718282
```

```
f(1, 0)
```

```
## [1] 54.59815
```

Las funciones no tienen por qué tener como argumentos o resultados sólo números reales: pueden involucrar vectores, matrices, tablas de datos, etc. Y se pueden definir por medio de secuencias de instrucciones, no sólo mediante fórmulas numéricas directas; en este caso, hay que separar las diferentes instrucciones con signos de punto y coma o escribir cada instrucción en una nueva línea. Ya iremos viendo ejemplos a medida que avance el curso.

En cada momento se pueden saber los objetos (por ejemplo, variables y funciones) que se han definido en la sesión hasta ese momento entrando la instrucción `ls()` o consultando la pestaña *Environment*. Para borrar la definición de un objeto, hay que aplicarle la función `rm`. Si se quiere hacer limpieza y borrar de golpe las definiciones de todos los objetos que se han definido hasta el momento, se puede emplear la instrucción `rm(list=ls())` o usar el botón con el icono de la escoba de la barra superior de la pestaña *Environment*.

```
rm(list=ls())    #Borramos todas las definiciones
f=function(t){t^2-2^t}
a=1
a
```

```
## [1] 1
```

```
ls()
```

```
## [1] "a" "f"
```

```
rm(a)
ls()
```

```
## [1] "f"
```

```
a
```

```
## Error in eval(expr, envir, enclos): objeto 'a' no encontrado
```

2.5 Números complejos (opcional)

Hasta aquí, hemos operado con números reales. Con R también podemos operar con números complejos. Los signos para las operaciones son los mismos que en el caso real.

```
(2+5i)*3
```

```
## [1] 6+15i
```

```
(2+5i)*(3+7i)
```

```
## [1] -29+29i
```

```
(2+5i)/(3+7i)
```

```
## [1] 0.7068966+0.0172414i
```

Fijaos en que cuando entramos en R un número complejo escrito en forma binomial $a + bi$, *no* escribimos un `*` entre la `i` y su coeficiente; de hecho, *no hay que escribirlo* :

```
2+5*i
```

```
## Error in eval(expr, envir, enclos): objeto 'i' no encontrado
```

Por otro lado, si el coeficiente de i es 1 o -1, hay que escribir el 1: por ejemplo, $3 - i$ se tiene que escribir `3-1i` . Si no lo hacemos, R da un mensaje de error.

```
(3+i)*(2-i)
```

```
## Error in eval(expr, envir, enclos): objeto 'i' no encontrado
```

```
(3+1i)*(2-1i)
```

```
## [1] 7-1i
```

Los complejos que tienen como parte imaginaria un número entero o un racional escrito en forma decimal se pueden entrar directamente en forma binomial, como lo hemos hecho hasta ahora. Para definir números complejos más... complejos, se puede usar la función

```
complex(real=..., imaginary=...)
```

Veamos un ejemplo:

```
1+2/3i #Esto en realidad es 1 más 2 partido por 3i
```

```
## [1] 1-0.666667i
```

```
1+(2/3)i
```

```
## Error: <text>:1:8: unexpected symbol
```

```
## 1: 1+(2/3)i
```

```
##      ^
```

```
complex(real=1,imaginary=2/3)
```

```
## [1] 1+0.666667i
```

```
z=1+sqrt(2)i
```

```
## Error: <text>:1:12: unexpected symbol
```

```
## 1: z=1+sqrt(2)i
```

```
##          ^
```

```
z=complex(real=1, imaginary=sqrt(2))
```

```
z
```

```
## [1] 1+1.414214i
```

Como sabéis, los números complejos se inventaron para poder trabajar con raíces cuadradas de números negativos. Ahora bien, por defecto, cuando calculamos la raíz cuadrada de un número negativo R no devuelve un número complejo, sino que se limita a avisarnos de que no existe.

```
sqrt(-3)
```

```
## Warning in sqrt(-3): Se han producido NaNs
```

```
## [1] NaN
```

Si queremos que R produzca un número complejo al calcular la raíz cuadrada de un número negativo, tenemos que especificar que este número negativo es un número complejo. La mejor manera de hacerlo es declarándolo como complejo aplicándole la función `as.complex`

```
sqrt(as.complex(-3))
```

```
## [1] 0+1.732051i
```

La mayoría de las funciones que hemos dado para los números reales admiten extensiones para números complejos, y con R se calculan con la misma función. Ahora no entraremos a explicar cómo se definen estas extensiones, sólo lo comentamos por si sabéis qué hacen y os interesa calcularlas.

```
sqrt(2+3i)
```

```
## [1] 1.674149+0.895977i
```

```
exp(2+3i)
```

```
## [1] -7.31511+1.042744i
```

```
sin(2+3i)
```

```
## [1] 9.154499-4.168907i
```

```
acos(as.complex(2)) #EL arcocoseno de 2 es un número complejo
```

```
## [1] 0+1.316958i
```

La raíz cuadrada merece un comentario. Naturalmente, `sqrt(2+3i)` calcula un número complejo z tal que $z^2 = 2 + 3i$. Como ocurre con los números reales, todo número complejo diferente de 0 tiene dos raíces cuadradas, y una se obtiene multiplicando la otra por -1. R da como raíz cuadrada de un número real la positiva, y como raíz cuadrada de un complejo la que tiene parte real positiva, y si su parte real es 0, la que tiene parte imaginaria positiva.

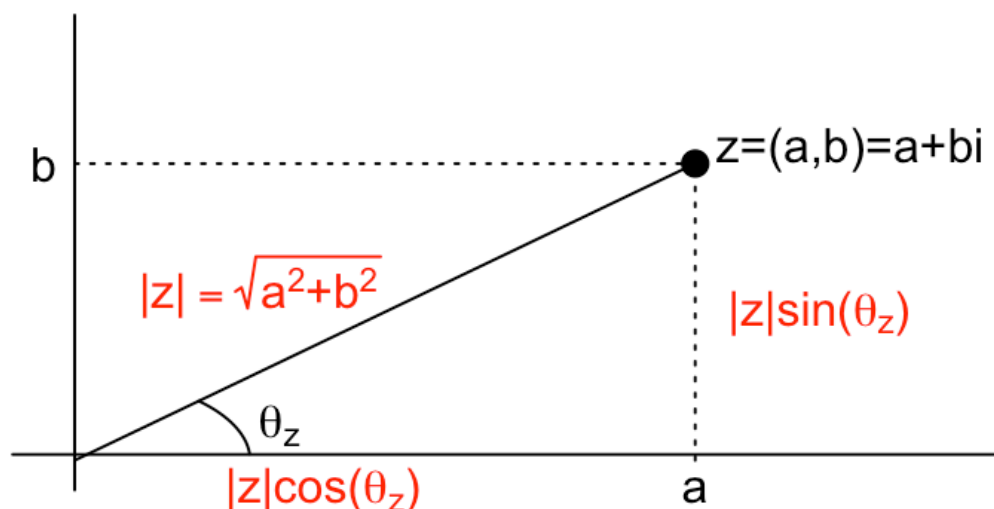


Figura 2.1: Interpretación geométrica de los números complejos.

Un número complejo $z = a + bi$ se puede representar como el punto (a, b) del plano cartesiano \mathbb{R}^2 . Esto permite asociarle dos magnitudes geométricas: véase la Figura 2.1

- El **módulo** de z , que denotaremos por $|z|$, es la distancia euclídea de $(0, 0)$ a (a, b) :

$$|z| = \sqrt{a^2 + b^2}.$$

Si $z \in \mathbb{R}$, su módulo coincide con su valor absoluto; en particular, si $z = 0$, su módulo es 0, y es el único número complejo de módulo 0.

- El **argumento** de z (para $z \neq 0$), que denotaremos por θ_z , es el ángulo que forman el semieje positivo de abscisas y el vector que va de $(0, 0)$ a (a, b) . Este ángulo está determinado por las ecuaciones

$$\cos(\theta_z) = \frac{a}{\sqrt{a^2 + b^2}}, \quad \sin(\theta_z) = \frac{b}{\sqrt{a^2 + b^2}}.$$

R sabe calcular módulos y argumentos de números complejos. Los argumentos los da en radianes y dentro del intervalo $(-\pi, \pi]$. En general, R dispone de las funciones básicas específicas para números complejos de la lista siguiente:

- **Parte real:** `Re`

- **Parte imaginaria:** Im
- **Módulo:** Mod
- **Argumento:** Arg
- **Conjugado:** Conj

Recordad que el **conjugado** de un número complejo $z = a + bi$ es $\bar{z} = a - bi$. Veamos algunos ejemplos de uso de estas funciones:

```
Re(4-7i)
```

```
## [1] 4
```

```
Im(4-7i)
```

```
## [1] -7
```

```
Mod(4-7i)
```

```
## [1] 8.062258
```

```
Arg(4-7i)
```

```
## [1] -1.05165
```

```
Conj(4-7i)
```

```
## [1] 4+7i
```


El módulo y el argumento de un número complejo $z \neq 0$ lo determinan de manera única, porque

$$z = |z| \left(\cos(\theta_z) + \sin(\theta_z)i \right).$$

Si queremos definir un número complejo mediante su módulo y argumento, no hace falta utilizar esta igualdad: podemos usar la instrucción

```
complex(modulus=..., argument=...)
```

Por ejemplo:

```
z=complex(modulus=3, argument=pi/5)
```

```
z
```

```
## [1] 2.427051+1.763356i
```

```
Mod(z)
```

```
## [1] 3
```

```
Arg(z)
```

```
## [1] 0.6283185
```

```
pi/5
```

```
## [1] 0.6283185
```

2.6 Guía rápida

- Signos de operaciones aritméticas:

- Suma: `+`
- Resta: `-`
- Multiplicación: `*`
- División: `/`
- Potencia: `^`
- Cociente entero: `/%`
- Resto de la división entera: `%%`

- Funciones numéricas:

- Valor absoluto: `abs`
- Raíz cuadrada: `sqrt`
- Exponencial de base e : `exp`
- Logaritmo neperiano: `log`
- Logaritmo decimal: `log10`
- Logaritmo binario: `log2`
- Logaritmo en base a : `log(...,base=a)`
- Factorial: `factorial`
- Número combinatorio: `choose`
- Seno: `sin`
- Coseno: `cos`
- Tangente: `tan`
- Arcoseno: `asin`
- Arcocoseno: `acos`
- Arcotangente: `atan`

- `pi` es el número π .
- `print(x, n)` muestra el valor de x con n cifras significativas.
- `round(x, n)` redondea el valor de x a n cifras decimales.
- `floor(x)` redondea x a un número entero por defecto.
- `ceiling(x)` redondea x a un número entero por exceso.
- `trunc(x)` da la parte entera de x .
- `variable=valor` asigna el `valor` a la `variable`. Otras construcciones equivalentes son `variable<-valor` y `valor->variable`.
- `función=function(variables){instrucciones}` define la `función` de variables las especificadas entre los paréntesis mediante las instrucciones especificadas entre las llaves.
- `ls()` nos da la lista de objetos actualmente definidos.
- `rm` borra la definición del objeto u objetos a los que se aplica.
- `rm(list=ls())` borra las definiciones de todos los objetos que hayamos definido.
- `complex` se usa para definir números complejos que no se puedan entrar directamente en forma binomial. Algunos parámetros importantes:
 - `real` e `imaginary` : sirven para especificar su parte real y su parte imaginaria.
 - `modulus` y `argument` : sirven para especificar su módulo y su argumento.
- `as.complex` convierte un número real en complejo.
- Funciones específicas para números complejos:
 - Parte real: `Re`
 - Parte imaginaria: `Im`
 - Módulo: `Mod`
 - Argumento: `Arg`
 - Conjugado: `Conj`

2.7 Ejercicios

Modelo de test

En los tests, tenéis que entrar las respuestas sin dejar ningún espacio en blanco excepto los que se pidan explícitamente. Cuando os pidan que deis una instrucción de R, *no* tenéis que incluir la marca de inicio `>`. Del mismo modo, cuando os pidan que copiéis un resultado dado por R, *no* tenéis que incluir el `[1]`.

(1) Dad una expresión para calcular $(2 + 7)8 + \frac{5}{2} - 3^6 + 8!$, con las operaciones escritas exactamente en el orden dado y sin paréntesis innecesarios, y a continuación, separado por un único espacio en blanco, copiad exactamente el resultado que ha dado R al evaluarla.

(2) Dad una expresión para calcular $|\sin(\sqrt{2}) - e^{\sqrt[5]{2}}|$, con las operaciones y funciones escritas exactamente en el orden dado, y a continuación, separado por un único espacio en blanco, copiad exactamente el resultado que ha dado R al evaluarla.

(3) Dad una expresión para calcular $\sin(37^\circ)$, empleando la construcción explicada en esta lección para calcular funciones trigonométricas de ángulos dados en grados, y a continuación, separado por un único espacio en blanco, copiad exactamente el resultado que ha dado R al evaluarla.

(4) Dad una expresión para calcular $3e - \pi$, con las operaciones escritas exactamente en la orden dado, y a continuación, separado por un único espacio en blanco, copiad exactamente el resultado que ha dado R al evaluarla.

(5) Dad una expresión para calcular $e^{2/3}$ redondeado a 3 cifras decimales y a continuación, separado por un único espacio en blanco, copiad exactamente el resultado que ha dado R al evaluarla.

(6) En una sola línea, definid x como $\sqrt{2}$ e y como $\cos(3\pi)$ y calculad $\ln(x^y)$; separad las tres instrucciones con puntos y comas seguidos de un único espacio en blanco. A continuación, separado por un espacio en blanco (sin punto y coma), copiad exactamente el resultado que ha dado R al evaluar esta secuencia de instrucciones.

(7) Corresponde el número en notación científica `3.3333e10` al número 33333000000? Tenéis que contestar SI (sin acento) o NO.

Ejercicio

Si hubiéramos empezado a contar segundos a partir de las 12 campanadas que marcaron el inicio de 2015, ¿qué día de qué año llegaríamos a los 250 millones de segundos?
¡Cuidado con los años bisiestos!

Respuestas al test

(1) `(2+7)*8+5/2-3^6+factorial(8)` 39665.5

(2) `abs(sin(sqrt(2))-exp(2^(1/5)))` 2.166319

También sería correcto `abs(sin(2^(1/2))-exp(2^(1/5)))` 2.166319

(3) `sin(37*pi/180)` 0.601815

(4) `3*exp(1)-pi` 5.013253

(5) `round(exp(2/3),3)` 1.948

(6) `x=sqrt(2); y=cos(3*pi); log(x^y)` -0.3465736

(7) SI

Lección 3 Un aperitivo: Introducción a la regresión lineal

En muchos libros de texto y artículos científicos encontraréis gráficos donde una línea recta o algún otro tipo de curva se ajusta a una serie de observaciones representadas por medio de puntos en el plano. La situación en general es la siguiente. Supongamos que tenemos una serie de puntos del plano cartesiano \mathbb{R}^2 ,

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n),$$

que representan pares de observaciones de dos variables numéricas: por ejemplo, $x =$ año e $y =$ población, o $x =$ longitud de una rama e $y =$ número de hojas en la rama. Queremos describir cómo depende la variable **dependiente** y de la variable **independiente** x a partir de estas observaciones. Para ello, buscaremos una función $y = f(x)$ cuya gráfica se aproxime lo máximo posible a los puntos $(x_i, y_i)_{i=1, \dots, n}$. Esta función nos dará un modelo matemático del comportamiento de las observaciones realizadas que nos permitirá entender mejor los mecanismos que relacionan las variables estudiadas o hacer predicciones sobre futuras observaciones.

Una primera opción, y la más sencilla, es estudiar si los puntos $(x_i, y_i)_{i=1, \dots, n}$ satisfacen una relación lineal. En este caso, se busca la recta de ecuación $y = b_1x + b_0$, con $b_0, b_1 \in \mathbb{R}$, que aproxime mejor los puntos dados, en el sentido de que la suma de los cuadrados de las diferencias entre los valores y_i y sus aproximaciones $b_1x_i + b_0$,

$$\sum_{i=1}^n (y_i - (b_1x_i + b_0))^2,$$

sea mínima. A esta recta $y = b_1x + b_0$ se la llama **recta de regresión por mínimos cuadrados**; para abreviar, aquí la llamaremos simplemente **recta de regresión**, porque es la única que estudiaremos por ahora.

El objetivo de esta lección es ilustrar el uso de R mediante el cálculo de esta recta de regresión. Para ello, introduciremos algunas funciones de R que ya explicaremos con más detalle en otras lecciones. Utilizaremos también transformaciones logarítmicas para tratar

casos en los que los puntos dados se aproximen mejor mediante una función potencial o exponencial.

3.1 Cálculo de rectas de regresión

Consideremos la Tabla 3.1, que da la altura media de los niños a determinadas edades. Los datos se han extraído de http://www.cdc.gov/growthcharts/clinical_charts.htm. Queremos determinar a partir de estos datos si hay una relación lineal entre la edad y la altura media de los niños.

Tabla 3.1: Alturas medias de niños por edad.

edad (años)	altura (cm)
1	76.11
2	86.45
3	95.27
5	109.18
7	122.03
9	133.73
11	143.73
13	156.41

Cuando tenemos una serie de observaciones emparejadas como las de esta tabla, la manera natural de almacenarlas en R es mediante una **tabla de datos**, un **data frame** en el argot de R. Aunque en este ejemplo concreto no sería necesario, lo haremos así para que empecéis a acostumbraros. La ventaja de tener los datos organizados en forma de *data frame* es que con ellos luego se pueden hacer muchas más cosas. Estudiaremos en detalle los *data frames* en la Lección ??.

Para crear este *data frame*, en primer lugar guardaremos cada fila de la Tabla 3.1 como un **vector**, es decir, como una lista ordenada de números, y le pondremos un nombre adecuado. Para definir un vector, podemos aplicar la función `c` a la secuencia ordenada de números, separados por comas:

```
edad=c(1,2,3,5,7,9,11,13)
altura=c(76.11,86.45,95.27,109.18,122.03,133.73,143.73,156.41)
edad
```

```
## [1]  1  2  3  5  7  9 11 13
```

```
altura
```

```
## [1]  76.11  86.45  95.27 109.18 122.03 133.73 143.73 156.41
```

Ahora vamos a construir un *data frame* de dos columnas, una para la edad y otra para la altura, y lo llamaremos `datos1`. Estas columnas serán las **variables** de nuestra tabla de datos. Para organizar diversos vectores de la misma longitud en un *data frame*, podemos aplicar la función `data.frame` a los vectores:

```
datos1=data.frame(edad,altura)
datos1
```

```
##   edad altura
## 1     1  76.11
## 2     2  86.45
## 3     3  95.27
## 4     5 109.18
## 5     7 122.03
## 6     9 133.73
## 7    11 143.73
## 8    13 156.41
```

Observad que las filas del *data frame* resultante corresponden a los pares (edad, altura) de la Tabla 3.1.

Al analizar unos datos, siempre es conveniente empezar con una representación gráfica que nos permita hacernos una idea de sus características. En este caso, lo primero que haremos será dibujar los pares (edad, altura) usando la función `plot`. Esta función tiene muchos parámetros que permiten mejorar el resultado, pero ya los veremos al estudiarla en detalle en la Lección [??](#). Por ahora nos conformamos con un gráfico básico de estos puntos que nos muestre su distribución.

Dada una familia de puntos $(x_n, y_n)_{n=1, \dots, k}$, si llamamos `x` al vector $(x_n)_{n=1, \dots, k}$ de sus abscisas e `y` al vector $(y_n)_{n=1, \dots, k}$ de sus ordenadas, podemos obtener el gráfico de los puntos $(x_n, y_n)_{n=1, \dots, k}$ mediante la instrucción

```
plot(x,y)
```

Si los vectores `x` e `y` son, en este orden, la primera y la segunda columna de un *data frame* de dos variables, como es nuestro caso, es suficiente aplicar la función `plot` al *data frame*. Así, por ejemplo, para dibujar el gráfico de la Figura [3.1](#) de los puntos $(\text{edad}_n, \text{altura}_n)_{n=1, \dots, 8}$, basta entrar la siguiente instrucción:

```
plot(datos1)
```

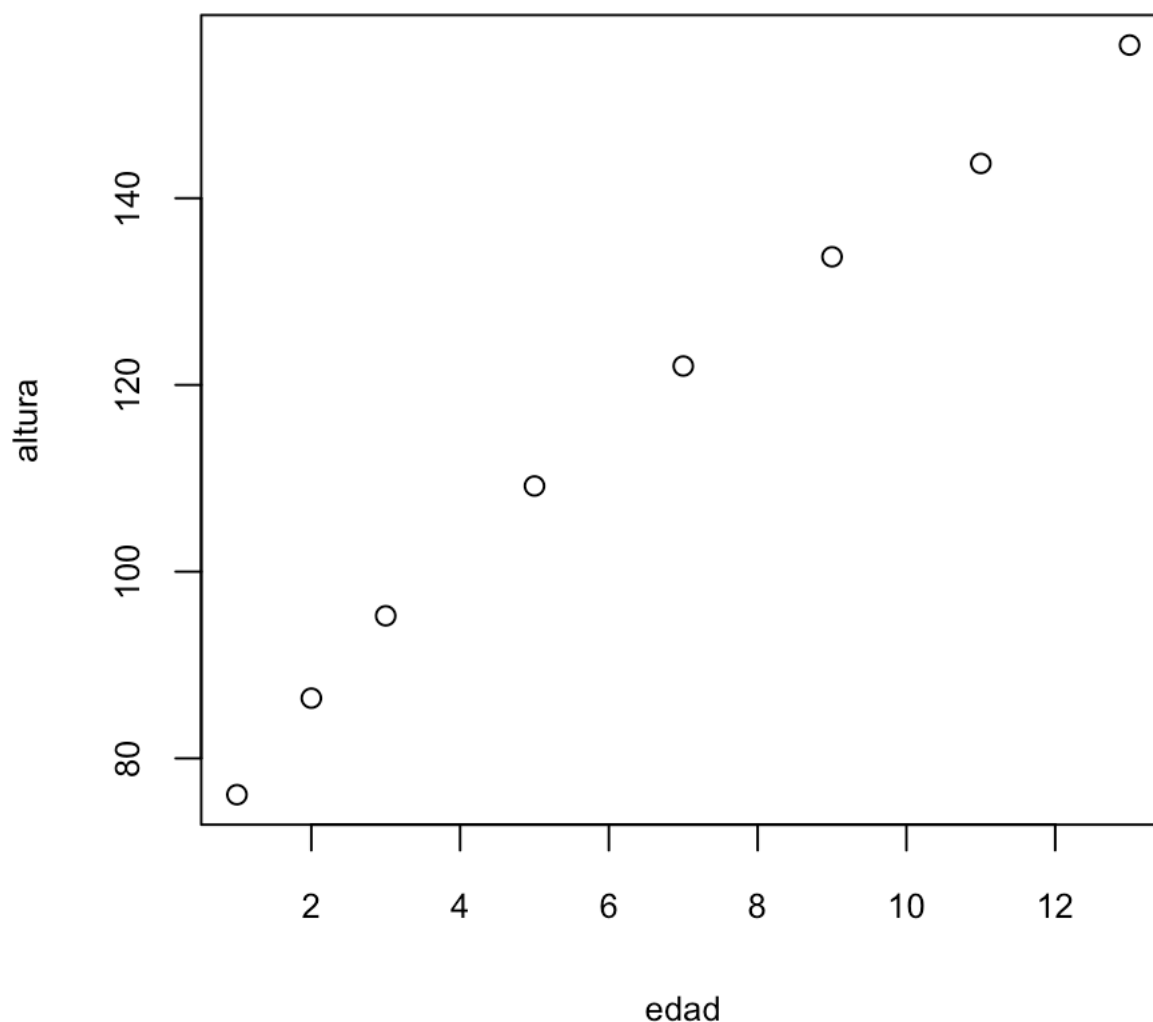


Figura 3.1: Representación gráfica de la altura media de los niños a determinadas edades.

Al ejecutar esta instrucción en la consola de `Rstudio`, el gráfico resultante se abrirá en la pestaña *Plots*, y en él se puede observar a simple vista que nuestros puntos siguen aproximadamente una recta. Vamos a calcular ahora su recta de regresión.

Dada una familia de puntos $(x_n, y_n)_{n=1, \dots, k}$, si llamamos `x` al vector $(x_n)_{n=1, \dots, k}$ de sus abscisas e `y` al vector $(y_n)_{n=1, \dots, k}$ de sus ordenadas, su recta de regresión se calcula con R por medio de la instrucción

```
lm(y~x)
```

Fijaos en la sintaxis: dentro del argumento de `lm`, primero va el vector `y`, seguido del vector `x` conectado a `y` por una tilde `~`. Para R, esta tilde significa **en función de**: es decir, `lm(y~x)` significa **la recta de regresión de y en función de x** . Para obtener este signo, los usuarios de Windows y Linux tienen que pulsar Ctrl+Alt+4 seguido de un espacio en blanco y los de Mac OS X con teclado español pueden pulsar Alt+Ñ seguido de un espacio en blanco.

Si los vectores `y` y `x` son dos columnas de un *data frame*, para calcular la recta de regresión de y en función de x podemos usar la instrucción

```
lm(y~x, data=nombre del data frame)
```

Así pues, para calcular la recta de regresión de los puntos $(\text{edad}_n, \text{altura}_n)_{n=1,\dots,8}$, entramos la siguiente instrucción:

```
lm(altura~edad, data=datos1)
```

```
##
## Call:
## lm(formula = altura ~ edad, data = datos1)
##
## Coefficients:
## (Intercept)      edad
##      73.968      6.493
```

El resultado que hemos obtenido significa que la recta de regresión tiene término independiente 73.968 (el punto donde la recta *interseca* al eje de las y) y el coeficiente de x es 6.493 (el coeficiente de la variable `edad`). Es decir, es la recta

$$y = 6.493x + 73.968.$$

Ahora la podemos superponer al gráfico anterior, empleando la función `abline`. Esta función permite añadir una recta al gráfico activo en la pestaña *Plots*. Por lo tanto, si no hemos cerrado el gráfico anterior, la instrucción

```
abline(lm(altura~edad, data=datos1))
```

le añade la recta de regresión, produciendo la Figura 3.2. Se ve a simple vista que, efectivamente, esta recta aproxima muy bien los datos.

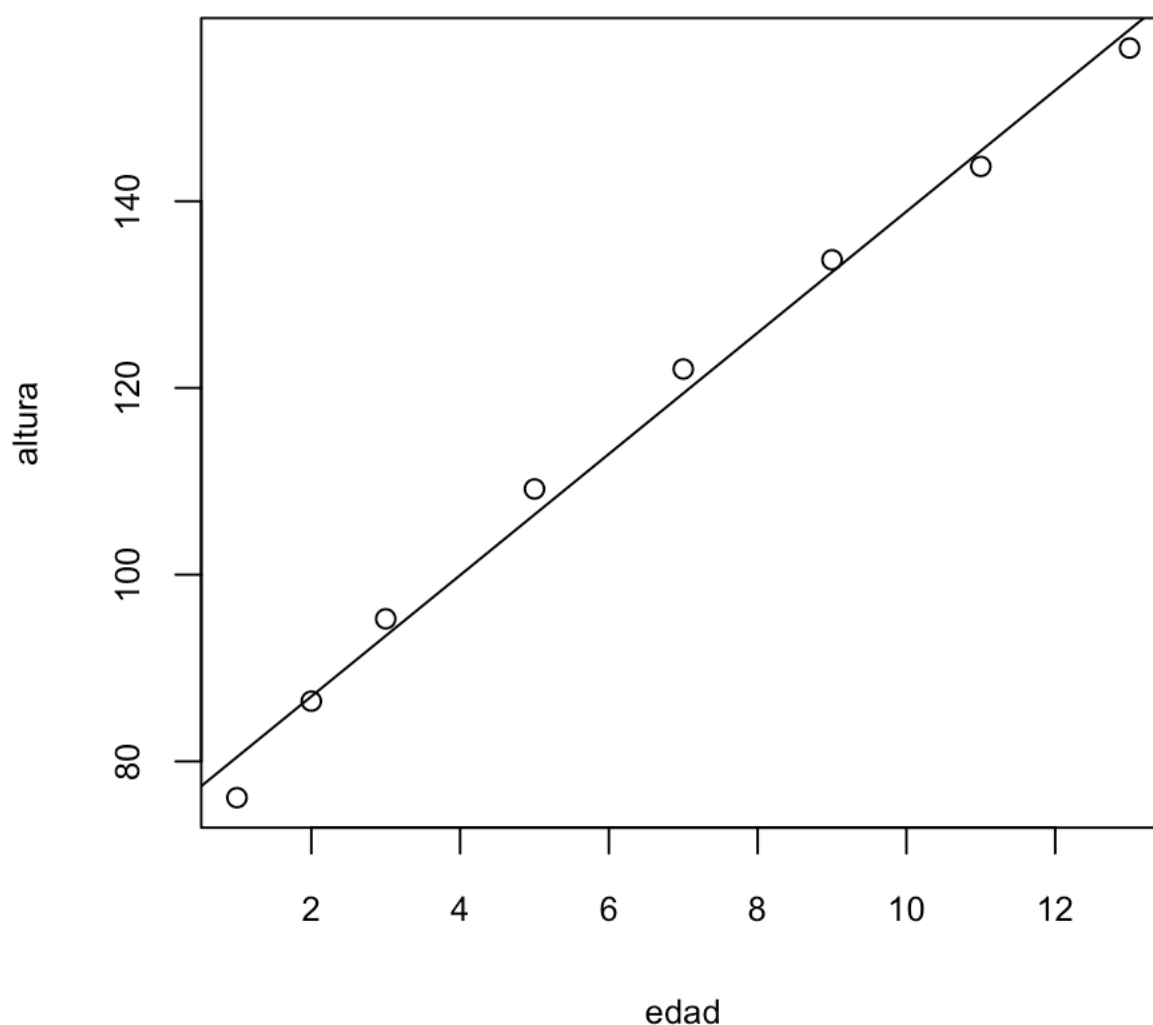


Figura 3.2: Ajuste mediante la recta de regresión de la altura media de los niños respecto de su edad.

Es importante tener presente que el análisis que hemos realizado de los pares de valores $(\text{edad}_n, \text{altura}_n)_{n=1, \dots, 8}$ ha sido puramente descriptivo: hemos mostrado que estos datos son consistentes con una función lineal, pero *no hemos demostrado* que la altura media sea función aproximadamente lineal de la edad. Esto último requeriría una demostración matemática o un argumento biológico, no una simple comprobación numérica para una muestra pequeña de valores, que, al fin y al cabo, es lo único que hemos hecho.

Lo que sí que podemos hacer ahora es usar la relación lineal observada para predecir la altura media de los niños de otras edades. Por ejemplo, ¿qué altura media estimamos que tienen los niños de 10 años? Si aplicamos la regla

$$\text{altura} = 6.493 \cdot \text{edad} + 73.968,$$

podemos predecir que la altura media a los 10 años es $6.493 \cdot 10 + 73.968 = 138.898$, es decir, de unos 139 cm.

Para evaluar numéricamente si la relación lineal que hemos encontrado es significativa o no, podemos usar el *coeficiente de determinación* R^2 . No explicaremos aquí cómo se define, ya lo haremos en la Lección [??](#). Es suficiente saber que es un valor entre 0 y 1 y que cuanto más se aproxime la recta de regresión al conjunto de puntos, más cercano será a 1. Por el momento, y como regla general, si este coeficiente de determinación R^2 es mayor que 0.9, consideraremos que el ajuste de los puntos a la recta es bueno.

Cuando R calcula la recta de regresión también obtiene este valor, pero no lo muestra si no se lo pedimos. Si queremos saber todo lo que ha calculado R con la función `lm`, tenemos que emplear la construcción `summary(lm(...))`. En general, la función `summary` aplicada a un objeto de R nos da un resumen de los contenidos de este objeto, resumen que depende de la clase de objeto que se trate.

Veamos cuál es el resultado de esta instrucción en nuestro ejemplo:

```
summary(lm(altura~edad, data=datos1))
```

```
##
## Call:
## lm(formula = altura ~ edad, data = datos1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.351  -1.743   0.408   2.018   2.745
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  73.9681     1.7979   41.14 1.38e-08 ***
## edad         6.4934     0.2374   27.36 1.58e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.746 on 6 degrees of freedom
## Multiple R-squared:  0.992, Adjusted R-squared:  0.9907
## F-statistic: 748.4 on 1 and 6 DF, p-value: 1.577e-07
```

Por ahora podemos prescindir de casi toda esta información (en todo caso, observad que la columna `Estimate` nos da los coeficientes de la recta de regresión) y fijarnos sólo en el primer valor de la penúltima línea, `Multiple R-squared`. Éste es el coeficiente de determinación R^2 que nos interesa. En este caso ha sido de 0.992, lo que confirma que la recta de regresión aproxima muy bien los datos.

Podemos pedir a R que nos dé el valor `Multiple R-squared` sin tener que obtener todo el `summary`, añadiendo el sufijo `$r.squared` a la construcción `summary(lm(...))`.

```
summary(lm(altura~edad, data=datos1))$r.squared
```

```
## [1] 0.9920466
```

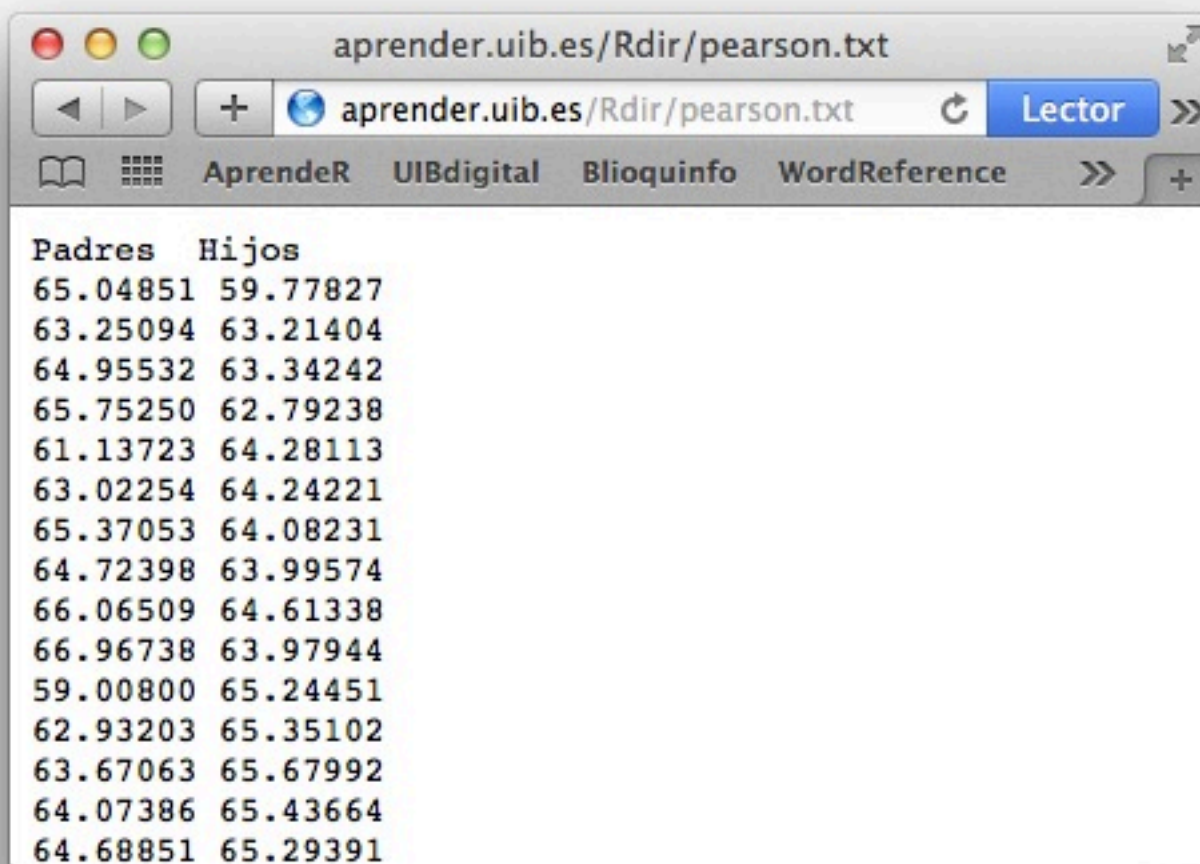
Los sufijos que empiezan con `$` suelen usarse en R para obtener componentes de un objeto. Por ejemplo, si al nombre de un *data frame* le añadimos el sufijo formado por `$` seguido del nombre de una de sus variables, obtenemos el contenido de esta variable.

```
datos1$edad
```

```
## [1] 1 2 3 5 7 9 11 13
```

Veamos otro ejemplo de cálculo de recta de regresión.

Ejemplo 3.1 Karl Pearson recopiló en 1903 las alturas de 1078 parejas formadas por un padre y un hijo. Hemos guardado en el `url` <http://aprender.uib.es/Rdir/pearson.txt> un fichero que contiene estas alturas. Si lo abrís en un navegador, veréis que es una tabla de dos columnas, etiquetadas `Padres` e `Hijos` (Figura 3.3). Cada fila contiene las alturas en pulgadas de un par Padre-Hijo.



Padres	Hijos
65.04851	59.77827
63.25094	63.21404
64.95532	63.34242
65.75250	62.79238
61.13723	64.28113
63.02254	64.24221
65.37053	64.08231
64.72398	63.99574
66.06509	64.61338
66.96738	63.97944
59.00800	65.24451
62.93203	65.35102
63.67063	65.67992
64.07386	65.43664
64.68851	65.29391

Figura 3.3: Vista en un navegador del fichero pearson.txt.

Vamos a usar estos datos para estudiar si hay dependencia lineal entre la altura de un hijo y la de su padre. Para ello, lo primero que haremos será cargarlos en un *data frame*. Esto se puede llevar a cabo de dos maneras:

- Usando el menú *Import Dataset* de la pestaña *Environment* de la ventana superior derecha de *RStudio*, sobre el que volveremos en la Lección ???. Al pulsar sobre este menú, se nos ofrece la posibilidad de importar un fichero de diferentes maneras; en este ejemplo, vamos a usar *From Text (readr)...*, que es la adecuada para importar tablas de Internet. Al seleccionarla, se nos pide el `url` del fichero y se nos dan a escoger una serie de opciones donde podemos especificar el nombre del *data frame* que queremos crear, si el fichero tiene o no una primera fila con los nombres de las columnas, cuál es el signo usado para separar columnas, etc. Pulsando el botón *Update* podremos ver en el campo *Data Preview* de esta ventana de diálogo el aspecto del *data frame* que obtendremos con las opciones seleccionadas; se trata entonces de escoger las opciones adecuadas para que se cree la versión correcta del *data frame*. En el caso concreto de esta tabla `pearson.txt`, se tiene que seleccionar la casilla de *First Row as Names* y escoger el valor *Whitespace* en *Delimiter* (Figura 3.4). Al pulsar el botón *Import*, se importará el fichero en un *data frame* con el nombre

especificado en el campo *Name* y se verá su contenido en la ventana de ficheros si se ha seleccionado la casilla *Open Data Viewer*.

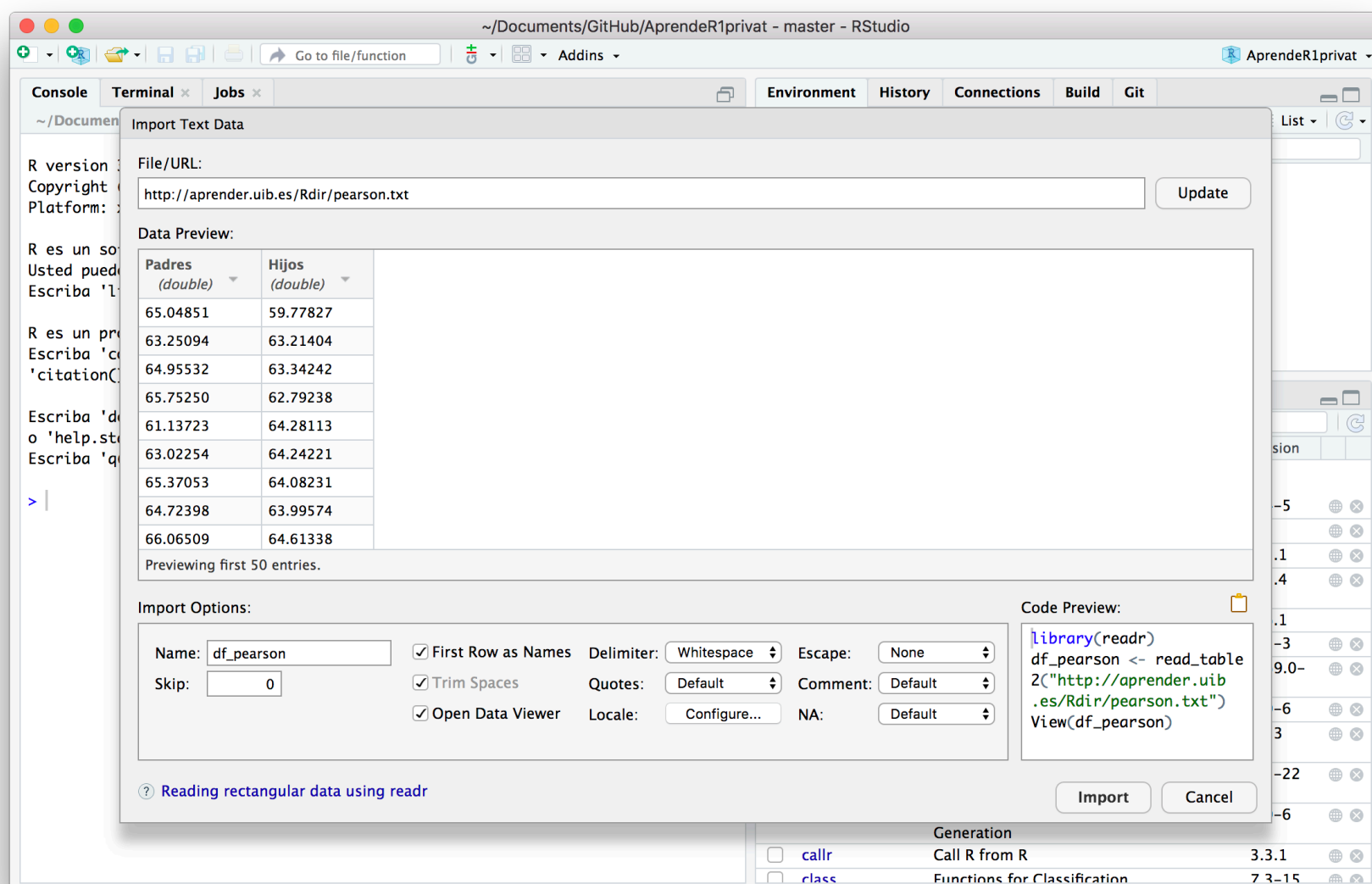


Figura 3.4: Opciones para guardar el fichero *pearson.txt* en un *data frame* llamado *df_pearson* usando el menú *Import Dataset*.

- Usando la instrucción `read.table`, de la que también hablaremos en la Lección ??; por ahora simplemente hay que saber que se ha de aplicar al nombre del fichero entre comillas, si está en el directorio de trabajo, o a su `url`, también escrito entre comillas. Si además el fichero contiene una primera fila con los nombres de las columnas, hay que añadir el parámetro `header=TRUE`.

Así pues, para cargar esta tabla de datos concreta en un *data frame* llamado `df_pearson`, podemos usar el menú *Import Dataset*, o entrar la instrucción siguiente:

```
df_pearson=read.table("http://aprender.uib.es/Rdir/pearson.txt", header=TRUE)
```

En ambos casos, para comprobar que se ha cargado bien, podemos usar las funciones `str`, que muestra la estructura del *data frame*, y `head`, que muestra sus primeras filas.


```
str(df_pearson)
```

```
## 'data.frame':    1078 obs. of  2 variables:
##  $ Padres: num  65 63.3 65 65.8 61.1 ...
##  $ Hijos : num  59.8 63.2 63.3 62.8 64.3 ...
```

```
head(df_pearson)
```

```
##      Padres    Hijos
## 1 65.04851 59.77827
## 2 63.25094 63.21404
## 3 64.95532 63.34242
## 4 65.75250 62.79238
## 5 61.13723 64.28113
## 6 63.02254 64.24221
```

El resultado de `str(df_pearson)` nos dice que este *data frame* está formado por 1078 observaciones (filas) de dos variables (columnas) llamadas `Padres` e `Hijos`. El resultado de `head(df_pearson)` nos muestra sus primeras seis filas, que podemos comprobar que coinciden con las del fichero original mostrado en la Figura 3.3.

Calculemos la recta de regresión de las alturas de los hijos respecto de las de los padres: ahora las siguientes instrucciones:

```
lm(Hijos~Padres, data=df_pearson)
```

```
##  
## Call:  
## lm(formula = Hijos ~ Padres, data = df_pearson)  
##  
## Coefficients:  
## (Intercept)      Padres  
##      33.8866      0.5141
```

```
summary(lm(Hijos~Padres, data=df_pearson))$r.squared
```

```
## [1] 0.2513401
```

Obtenemos la recta de regresión

$$y = 33.8866 + 0.5141x,$$

donde y representa la altura de un hijo y x la de su padre, y un coeficiente de determinación $R^2 = 0.25$, muy bajo. La regresión no es muy buena, como se puede observar en la Figura 3.5 que generamos con el código siguiente:

```
plot(df_pearson)  
abline(lm(Hijos~Padres, data=df_pearson),col="red")
```

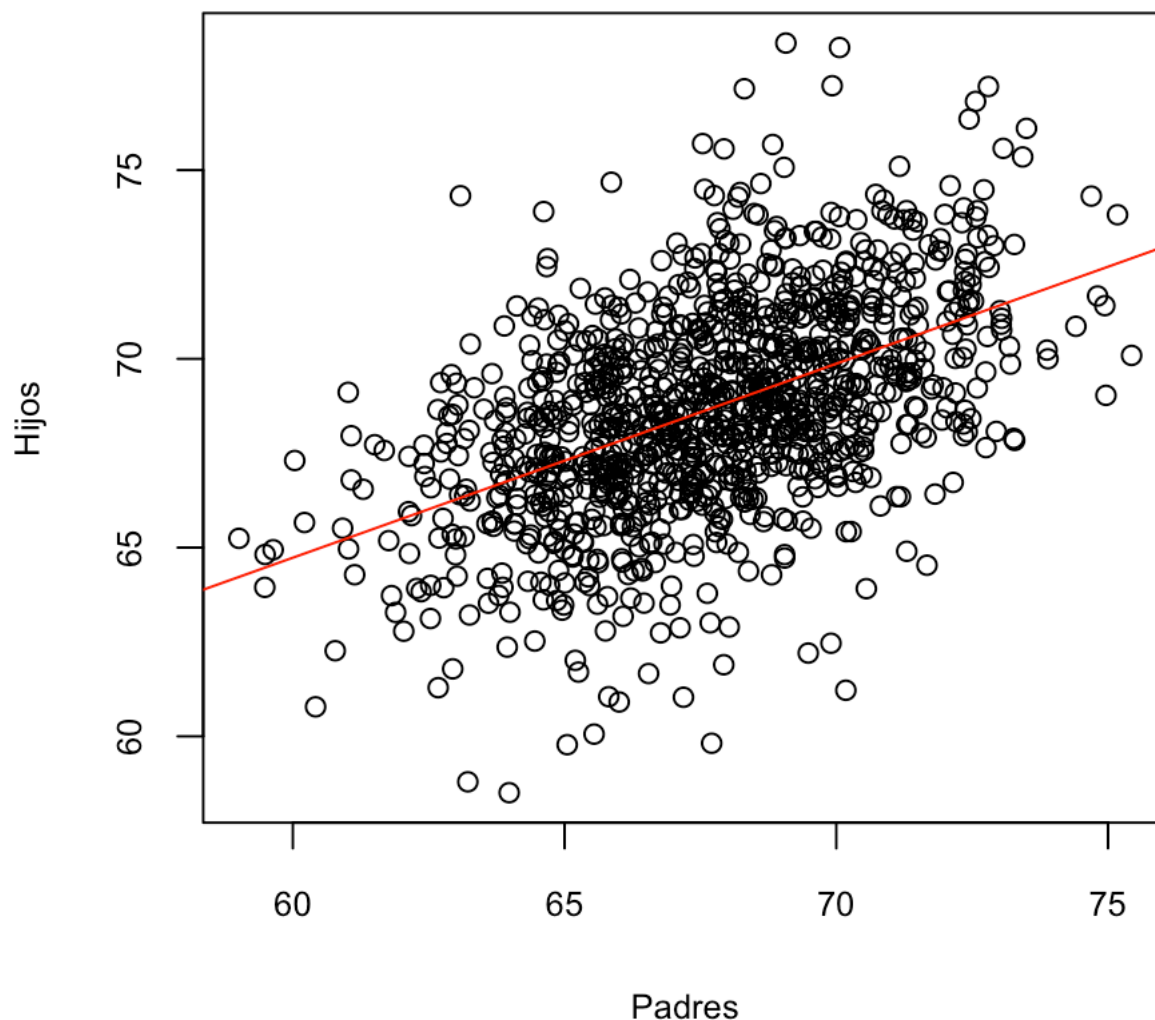


Figura 3.5: Representación gráfica de las alturas de los hijos en función de la de sus padres, junto con su recta de regresión.

Hemos usado el parámetro `col="red"` en el `abline` para que la recta de regresión sea roja y facilitar así su visualización en medio de la nube de puntos.

3.2 Rectas de regresión y transformaciones logarítmicas

La dependencia de un valor en función de otro no siempre es lineal. A veces podremos detectar otras dependencias (en concreto, exponenciales o potenciales) realizando un **cambio de escala** adecuado en el gráfico.

Cuando dibujamos un gráfico, lo normal es marcar cada eje de manera que la misma distancia entre marcas signifique la misma diferencia entre sus valores; por ejemplo, en el gráfico de la Figura 3.1, las marcas sobre cada uno de los ejes están igualmente espaciadas, de manera que entre cada par de marcas consecutivas en el eje de abscisas

hay una diferencia de 2 años y entre cada par de marcas consecutivas en el eje de ordenadas hay una diferencia de 20 cm. Decimos entonces que los ejes están en **escala lineal**. Pero a veces es conveniente dibujar algún eje en **escala logarítmica**, situando las marcas de tal manera que la misma distancia entre marcas signifique el mismo *cociente* entre sus valores. Como el logaritmo transforma cocientes en restas, un eje en escala logarítmica representa el logaritmo de sus valores en escala lineal.

Decimos que un gráfico está en **escala semilogarítmica** cuando su eje de abscisas está en escala lineal y su eje de ordenadas en escala logarítmica. Salvo por los valores en las marcas sobre el eje de las y , esto significa que dibujamos en escala lineal el gráfico de $\log(y)$ en función de x . Así pues, si al representar unos puntos (x, y) en escala semilogarítmica observamos que siguen aproximadamente una recta, esto querrá decir que los valores $\log(y)$ siguen una ley aproximadamente lineal en los valores x , y, por lo tanto, que y sigue una ley aproximadamente exponencial en x . En efecto, si $\log(y) = ax + b$, entonces

$$y = 10^{\log(y)} = 10^{ax+b} = 10^{ax} \cdot 10^b = 10^b \cdot (10^a)^x = \beta \cdot \alpha^x,$$

donde $\beta = 10^b$ y $\alpha = 10^a$.

De manera similar, decimos que un gráfico está en **escala doble logarítmica** cuando ambos ejes están en escala logarítmica. Esto es equivalente, de nuevo salvo por los valores en las marcas sobre los ejes, a dibujar en escala lineal el gráfico de $\log(y)$ en función de $\log(x)$. Por consiguiente, si al dibujar unos puntos (x, y) en escala doble logarítmica observamos que siguen aproximadamente una recta, esto querrá decir que los valores $\log(y)$ siguen una ley aproximadamente lineal en los valores $\log(x)$, y, por lo tanto, que y sigue una ley aproximadamente potencial en x . En efecto, si $\log(y) = a \log(x) + b$, entonces

$$y = 10^{\log(y)} = 10^{a \log(x) + b} = 10^{a \log(x)} \cdot 10^b = 10^b \cdot (10^{\log(x)})^a = 10^b \cdot x^a = \beta \cdot x^a,$$

donde $\beta = 10^b$.

Veamos algunos ejemplos de regresiones lineales con cambios de escala.

Ejemplo 3.2 La serotonina se asocia a la estabilidad emocional en el hombre. En un experimento (véase el artículo de B. Peskar y S. Spector “Serotonin: Radioimmunoassay” en *Science* 179 (1973), pp. 1340-1341) se midió, para algunas cantidades de serotonina (expresadas en *nanogramos*, la milmillonésima parte de un gramo), el porcentaje de

inhibición de un cierto proceso bioquímico en el que se observaba su presencia. El objetivo era estimar la cantidad de serotonina presente en un tejido a partir del porcentaje de inhibición observado. Los datos que se obtuvieron son los de la Tabla 3.2.

Tabla 3.2: Porcentajes de inhibición de un cierto proceso bioquímico en presencia de serotonina.

serotonina (ng)	inhibición (%)
1.2	19
3.6	36
12.0	60
33.0	84

Como queremos predecir la cantidad de serotonina en función de la inhibición observada, consideraremos los pares (inhibición,serotonina). En esta ocasión, en vez de trabajar con un *data frame*, trabajaremos directamente con los vectores.

```
inh=c(19,36,60,84)
ser=c(1.2,3.6,12,33)
```

Con la instrucción siguiente obtenemos la Figura 3.6, donde vemos claramente que la cantidad de serotonina no es función lineal de la inhibición.

```
plot(inh,ser)
```

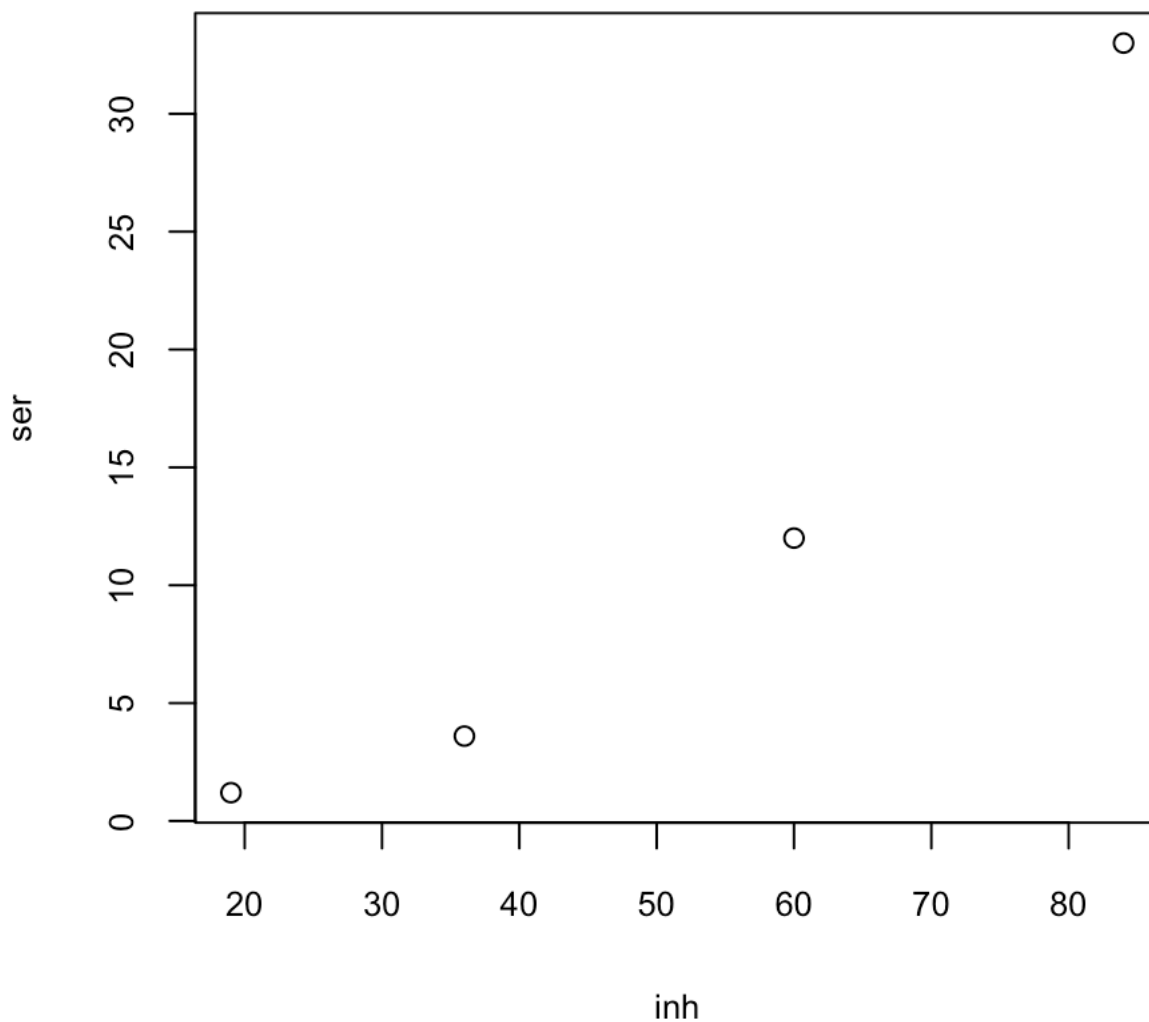


Figura 3.6: Representación gráfica en escala lineal del porcentaje de inhibición en función de la cantidad de serotonina.

Vamos a dibujar ahora el gráfico semilogarítmico de estos puntos, para ver si de esta manera quedan sobre una recta. Para ello, tenemos que añadir al argumento de `plot` el parámetro `log="y"`.

```
plot(inh, ser, log="y")
```

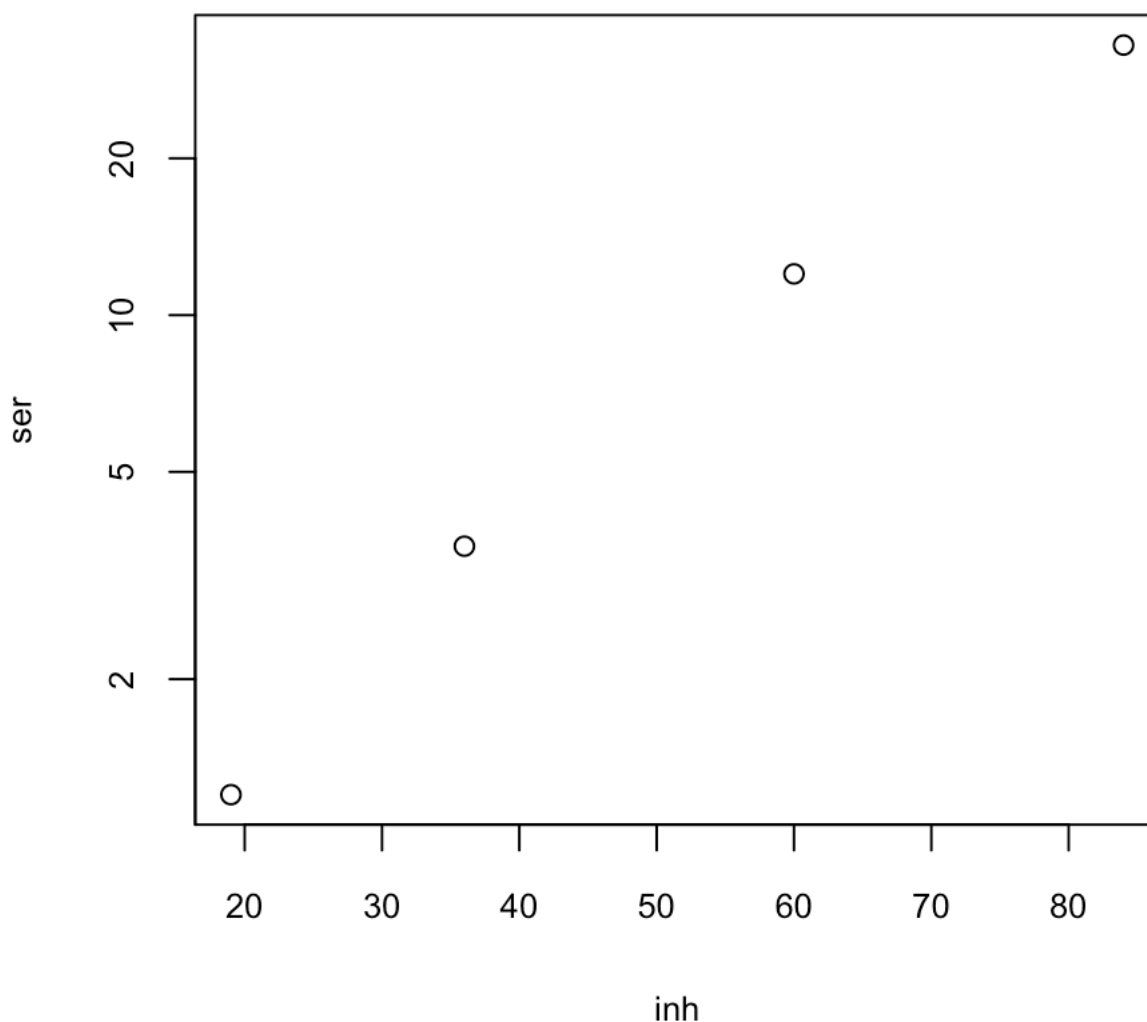


Figura 3.7: Representación gráfica en escala semilogarítmica del porcentaje de inhibición en función de la cantidad de serotonina.

Obtenemos la Figura 3.7. Observad cómo las marcas en el eje de ordenadas no están distribuidas de manera lineal: la distancia de 5 a 10 es la misma que de 10 a 20. Los puntos en este gráfico sí que parecen seguir una recta. Por lo tanto, parece que el logaritmo de la cantidad de serotonina es una función aproximadamente lineal del porcentaje de inhibición. Para confirmarlo, calcularemos la recta de regresión de los puntos

$$(\text{inhibición}_n, \log(\text{serotonina}_n))_{n=1,\dots,4}.$$

Para calcular los logaritmos en base 10 de todas las cantidades de serotonina en un solo paso, podemos aplicar la función `log10` directamente al vector `ser`.

```
log10(ser)
```

```
## [1] 0.07918125 0.55630250 1.07918125 1.51851394
```

```
lm(log10(ser)~inh)
```

```
##  
## Call:  
## lm(formula = log10(ser) ~ inh)  
##  
## Coefficients:  
## (Intercept)          inh  
##      -0.28427      0.02196
```

```
summary(lm(log10(ser)~inh))$r.squared
```

```
## [1] 0.9921146
```

El resultado indica que la recta de regresión de estos puntos es $y = 0.02196x - 0.28427$, con un valor de R^2 de 0.992, muy bueno. Por lo tanto, podemos afirmar que, aproximadamente,

$$\log(\text{serotonina}) = 0.02196 \cdot \text{inhibición} - 0.28427.$$

Elevando 10 a cada uno de los lados de esta identidad, obtenemos

$$\begin{aligned}\text{serotonina} &= 10^{\log(\text{serotonina})} = 10^{-0.28427} \cdot 10^{0.02196 \cdot \text{inhibición}} \\ &= 0.52 \cdot 1.052^{\text{inhibición}}.\end{aligned}$$

Es decir, los puntos de partida siguen aproximadamente la función exponencial

$$y = 0.52 \cdot 1.052^x.$$

Vamos ahora a dibujar en un mismo gráfico los puntos (inhibición_n , serotonina_n) y esta función exponencial. Para añadir la gráfica de una función $y = f(x)$ al gráfico activo en la pestaña *Plots* podemos emplear la función

```
curve(f(x), add=TRUE)
```


Así, el código siguiente produce la Figura 3.8; fijaos en cómo hemos especificado la función $y = 0.52 \cdot 1.052^x$ dentro del `curve`.

```
plot(inh, ser)
curve(0.52*1.052^x, add=TRUE)
```

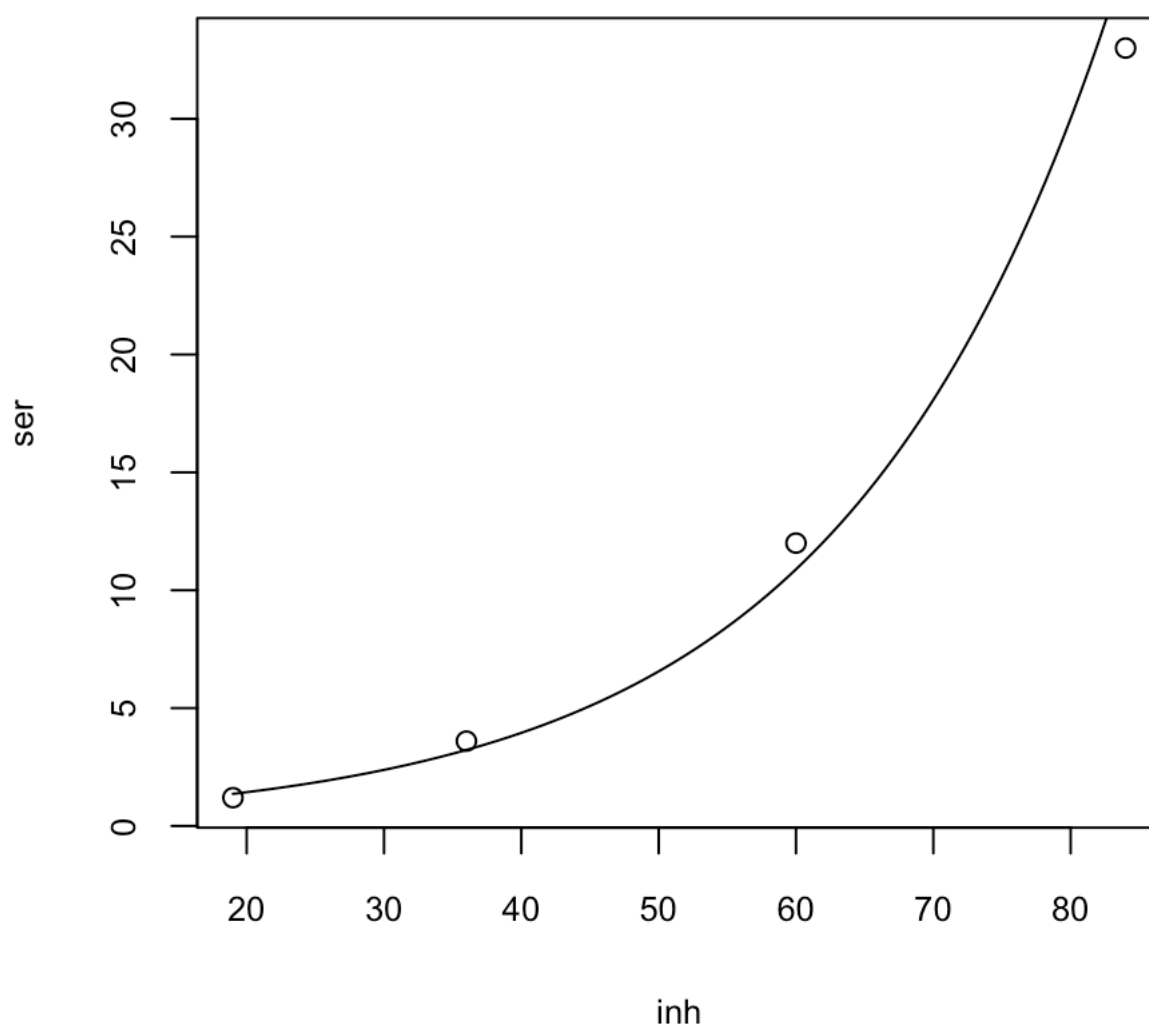


Figura 3.8: Representación gráfica en escala lineal del porcentaje de inhibición en función de la cantidad de serotonina, junto con la función $y = 0.52 \cdot 1.052^x$.

Ahora podemos usar la relación observada,

$$\text{serotonina} = 0.52 \cdot 1.052^{\text{inhibición}},$$

para estimar la cantidad de serotonina presente en el tejido a partir de una inhibición concreta. Por ejemplo, si hemos observado un 25% de inhibición, podemos estimar que la cantidad de serotonina es $0.52 \cdot 1.052^{25} = 1.84$ ng

Ejemplo 3.3 Consideremos ahora los datos de la Tabla 3.3. Se trata de los números acumulados de casos de SIDA en los Estados Unidos desde 1981 hasta 1992, extraídos del *HIV/AIDS Surveillance Report* de 1993 (<http://www.cdc.gov/hiv/topics/surveillance/resources/reports/index.htm>). *Acumulados* significa que, para cada año, se da el número de casos detectados *hasta* entonces.

Tabla 3.3: Números acumulados anuales de casos de SIDA en los Estados Unidos, 1981 a 1992.

año	casos
1981	97
1982	709
1983	2698
1984	6928
1985	15242
1986	29944
1987	52902
1988	83903
1989	120612
1990	161711
1991	206247
1992	257085

Queremos estudiar el comportamiento de estos números acumulados de casos en función del tiempo expresado en años a partir de 1980. Lo primero que hacemos es cargar los datos en un *data frame*. Fijaos en que la lista de años va a ser la secuencia de números consecutivos entre 1 y 12. Para definir la secuencia de números consecutivos entre *a* y *b* podemos usar la construcción `a:b`. Esto nos ahorra trabajo y reduce las oportunidades de cometer errores al escribir los números.

```
tiempo=1:12  
SIDA_acum=c(97,709,2698,6928,15242,29944,52902,83903,120612,161711,206247,257085)  
df_SIDA=data.frame(tiempo, SIDA_acum)
```

Con la instrucción siguiente dibujamos estos datos:

```
plot(df_SIDA)
```

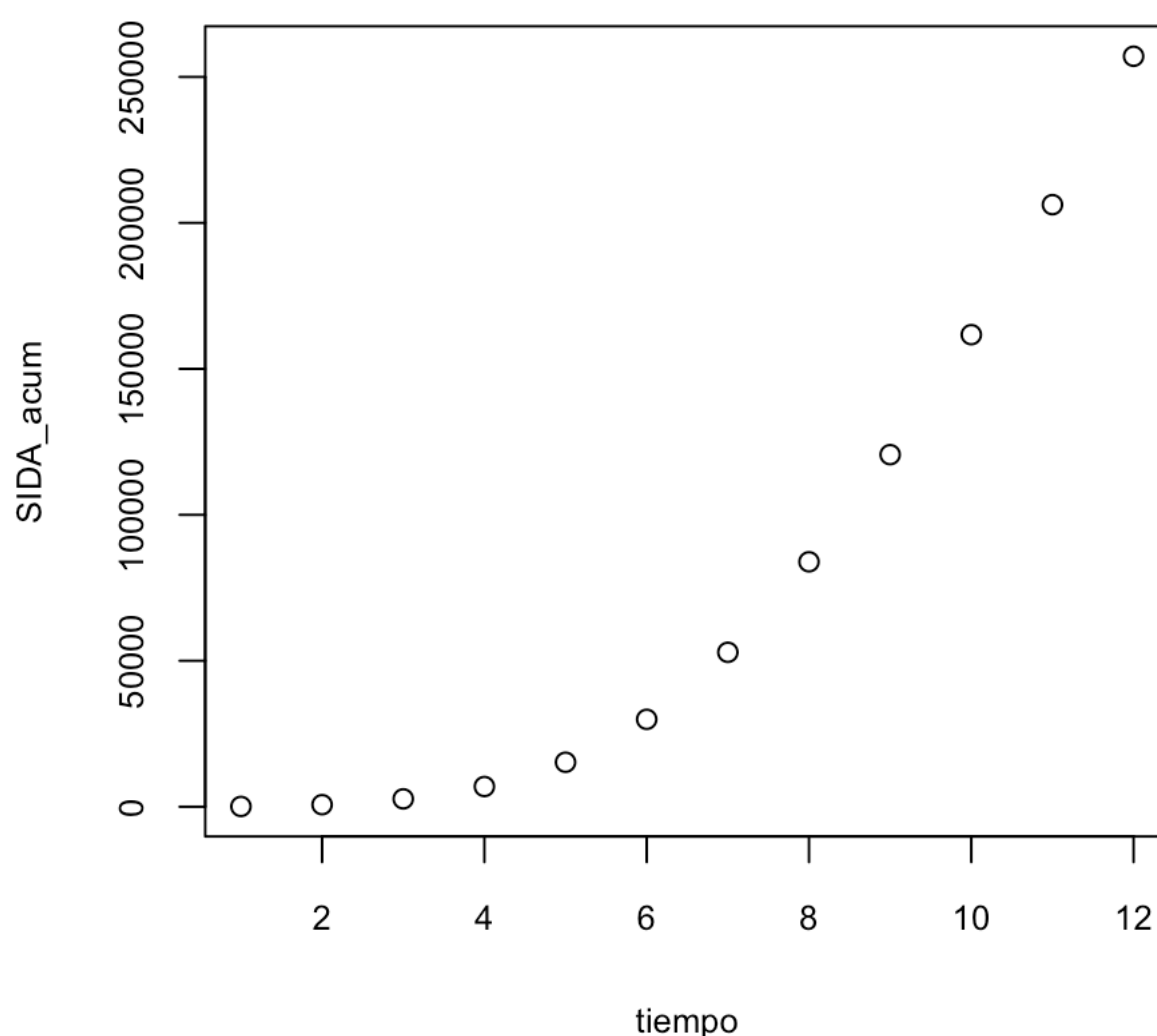


Figura 3.9: Representación gráfica en escala lineal del número acumulado de casos de SIDA en EEUU desde 1980 en función de los años transcurridos desde ese año.

Obtenemos el gráfico de la Figura 3.9, y está claro que los puntos (x_n, y_n) , donde x representa el año e y el número acumulado de casos de SIDA, no se ajustan a una recta. De hecho, a simple vista se diría que el crecimiento de y en función de x es exponencial.

Para confirmar este crecimiento exponencial, dibujamos el gráfico semilogarítmico:

```
plot(df_SIDA, log="y")
```

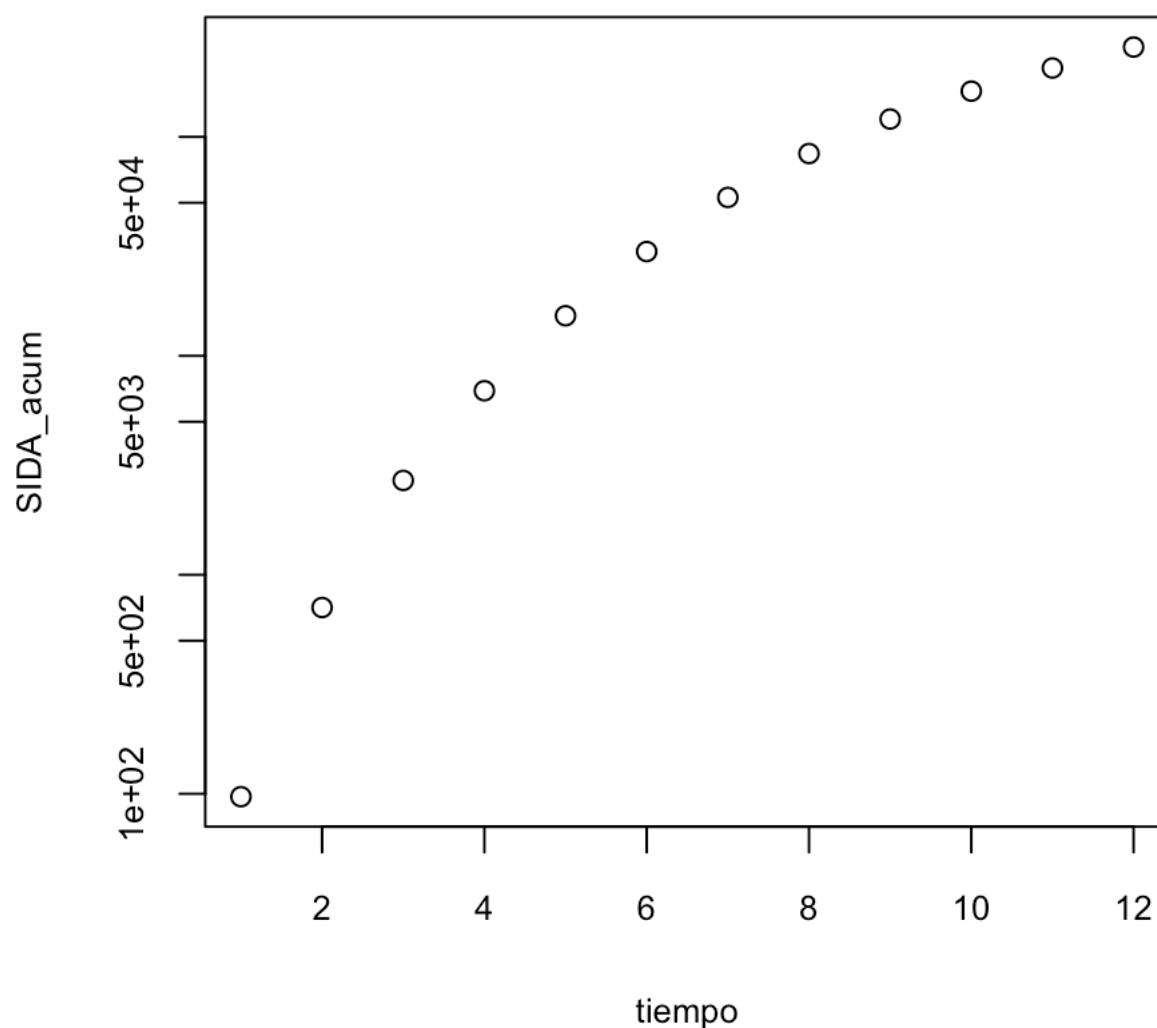


Figura 3.10: Representación gráfica en escala semilogarítmica del número acumulado de casos de SIDA en EEUU desde 1980 en función de los años transcurridos desde ese año.

Obtenemos el gráfico de la Figura 3.10, donde los puntos tampoco siguen una recta. Así pues, resulta que y tampoco parece ser función exponencial de x .

Vamos a ver si el crecimiento de y en función de x es potencial. Para ello, dibujaremos un gráfico doble logarítmico de los puntos (x_n, y_n) , especificando `log="xy"` dentro del argumento de `plot`.

```
plot(df_SIDA, log="xy")
```

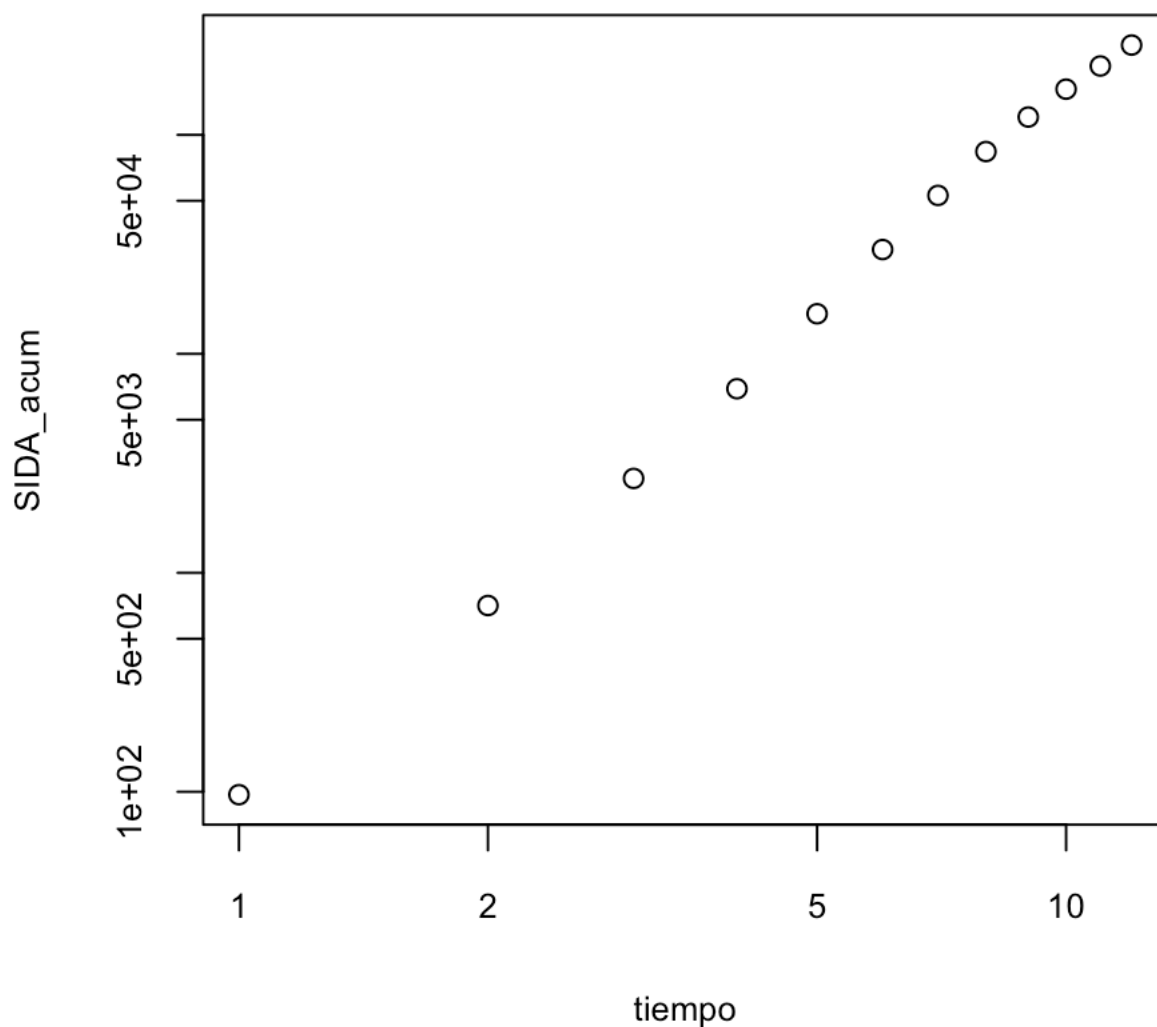


Figura 3.11: Representación gráfica en escala doble logarítmica del número acumulado de casos de SIDA en EEUU desde 1980 en función de los años transcurridos desde ese año.

Obtenemos el gráfico de la Figura 3.11, y ahora sí que parece lineal. Así que parece que los números acumulados de casos de SIDA crecieron potencialmente con el transcurso de los años.

Lo que haremos ahora será calcular la recta de regresión del logaritmo de `SIDA_acum` respecto del logaritmo de `tiempo` y mirar el coeficiente de determinación. Recordad que podemos aplicar una función a todas las entradas de un vector en un solo paso.

```
lm(log10(SIDA_acum)~log10(tiempo), data=df_SIDA)
```

```
##

## Call:
## lm(formula = log10(SIDA_acum) ~ log10(tiempo), data = df_SIDA)
##

## Coefficients:
##      (Intercept)      log10(tiempo)
##           1.918           3.274
```

```
summary(lm(log10(SIDA_acum)~log10(tiempo),data=df_SIDA))$r.squared
```

```
## [1] 0.9983866
```

La regresión que obtenemos es $\log(y) = 1.918 + 3.274 \log(x)$, con un valor de R^2 de 0.998, muy alto. Elevando 10 a ambos lados de esta igualdad, obtenemos

$$\begin{aligned} y = 10^{\log(y)} &= 10^{1.918} \cdot 10^{3.274 \log(x)} = 10^{1.918} \cdot (10^{\log(x)})^{3.274} \\ &= 82.79422 \cdot x^{3.274}. \end{aligned}$$

Para ver si los puntos $(\text{tiempo}_n, \text{SIDA_acum}_n)_{n=1, \dots, 12}$ se ajustan bien a la curva

$$y = 82.79422 \cdot x^{3.274},$$

dibujaremos los puntos y la curva en un único gráfico (en escala lineal):

```
plot(df_SIDA)
curve(82.79422*x^3.274, add=TRUE)
```

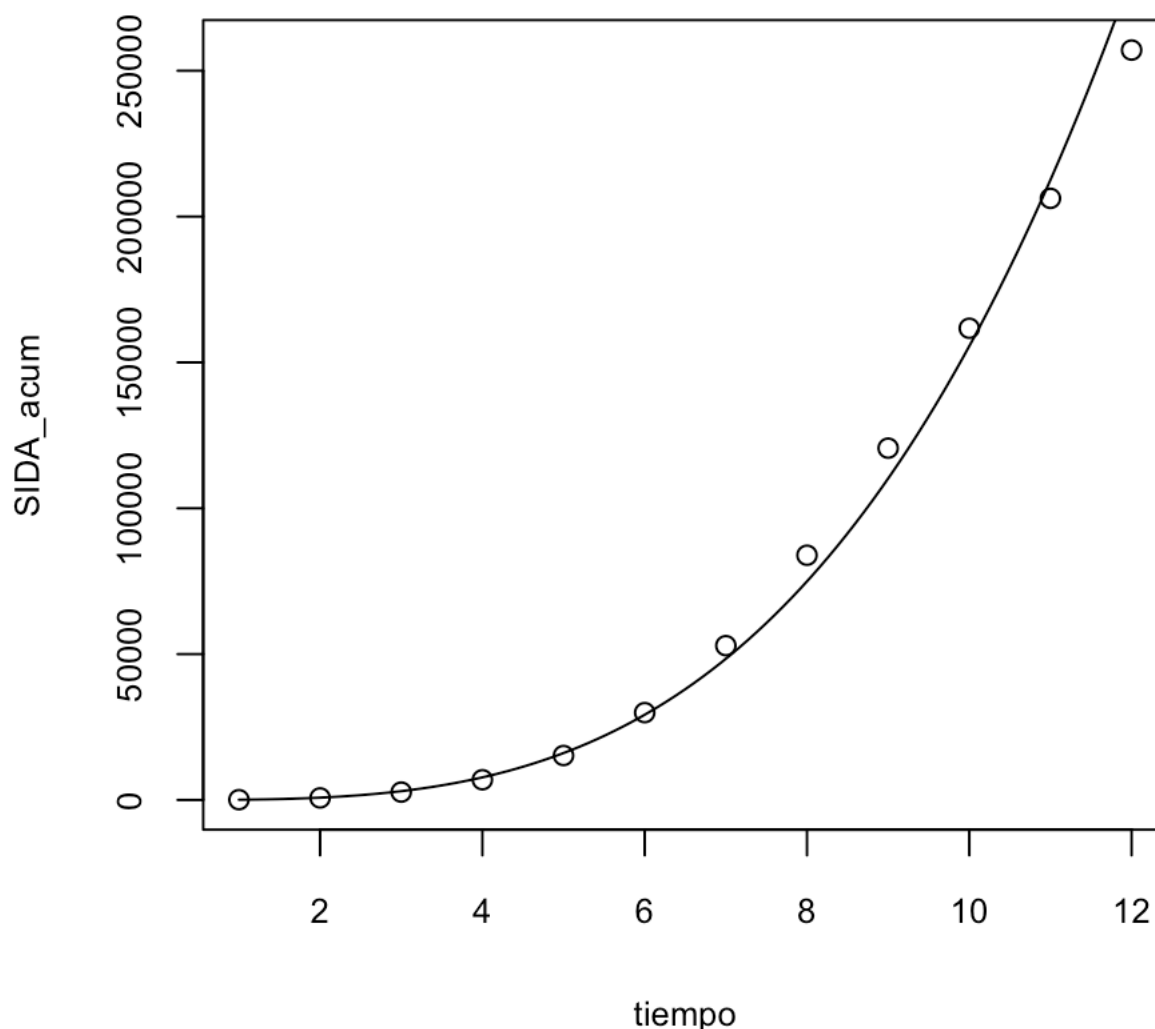


Figura 3.12: Representación gráfica en escala lineal de la cantidad acumulada de enfermos de SIDA en EEUU desde 1980 en función de los años transcurridos desde ese año, junto con su ajuste mediante la función potencial $82.79422 \cdot x^{3.274}$.

Obtenemos la Figura 3.12, donde vemos que la curva se ajusta bastante bien a los puntos.

Hay que mencionar aquí que se han propuesto modelos matemáticos que predicen que, cuando se inicia una epidemia de SIDA en una población, los números acumulados de casos en los primeros años son proporcionales al cubo del tiempo transcurrido desde el inicio; véase, por ejemplo, el artículo de S.A. Colgate, E. A. Stanley, J. M. Hyman, S. P. Layne y C. Qualls “Risk behavior-based model of the cubic growth of acquired immunodeficiency syndrome in the United States”, en *PNAS* 86 (1989), pp. 4793-4797. El resultado del análisis que hemos realizado es consistente con esta predicción teórica.

3.3 Guía rápida

- `c` sirve para definir vectores.
- `a:b`, con $a < b$, define un vector con la secuencia $a, a+1, a+2, \dots, b$.

- `data.frame` , aplicada a unos vectores de la misma longitud, define un *data frame* (el tipo de objetos de R en los que guardamos usualmente las tablas de datos) cuyas columnas serán estos vectores.
- `read.table` define un *data frame* a partir de un fichero externo. También se puede usar el menú *Import Dataset* de la pestaña *Environment* en la ventana superior derecha de *RStudio*.
- `lm(y~x)` calcula la recta de regresión del vector y respecto del vector x . Si x e y son dos columnas de un *data frame*, éste se ha de especificar en el argumento mediante el parámetro `data` igualado al nombre del *data frame*.
- `summary` sirve para obtener un resumen estadístico de un objeto. Este resumen depende del objeto. En el caso de una recta de regresión calculada con `lm` , muestra una serie de información estadística extra obtenida en dicho cálculo.
- `plot(x,y)` produce el gráfico de los puntos (x_n, y_n) . Si x e y son, respectivamente, la primera y la segunda columna de un *data frame* de dos columnas, se le puede entrar directamente el nombre del *data frame* como argumento. El parámetro `log` sirve para indicar los ejes que se desea que estén en escala logarítmica: `"x"` (abscisas), `"y"` (ordenadas) o `"xy"` (ambos).
- `abline` añade una recta al gráfico activo.
- `curve(función, add=TRUE)` añade la gráfica de la `función` al gráfico activo.

3.4 Ejercicios

Ejercicio

Las larvas de *Lymantria dispar*, conocidas como *orugas peludas del alcornoque*, son una plaga en bosques y huertos. En un experimento se quiso determinar la capacidad de atracción de una cierta feromona sobre los machos de esta especie, con el objetivo de emplearla en trampas (véase el artículo de M. Beroza y E. F. Knipling “Gypsy moth control with the sex attractant pheromone” en *Science* 177 (1972), pp. 19-27). En la Tabla 3.4, x representa la cantidad de feromona empleada, en microgramos (la millonésima parte de un gramo) y N el número de machos atrapados en una trampa empleando esta cantidad de feromona para atraerlos.

Tabla 3.4: Cantidades de feromona empleadas en trampas y números de machos atrapados.

x	N
0.1	3
1.0	6
5.0	9
10.0	11
100.0	20

1. Decidid si, en los puntos (x, N) dados en la Tabla 3.4, el valor de N sigue una función aproximadamente lineal, exponencial o potencial en el valor de x .
2. En caso de ser una función de uno de estos tres tipos, calculadla.
3. Representad en un gráfico los puntos (x, N) de la Tabla 3.4 y la función que hayáis calculado en el apartado anterior, para visualizar la bondad del ajuste de la curva a los puntos.
4. Estimad cuánta feromona tenemos que usar en una trampa para atraer a 50 machos.