Solución a la práctica 1 con R

En esta práctica vamos a utilizar los operadores abline, cov, cor, data.frame, edit, head, lm, mean, names, plot, summary, tail y View. También vamos a utilizar el operador read_excel. Para ello necesitamos instalar, si no lo hemos hecho previamente, y cargar el paquete *readxl*. Para instalar el paquete ejecutamos *install.packages("readxl")*, y para cargarlo *library(readxl).*

Ejercicio 1. Supongamos los siguientes 10 datos para la variable *x* y la variable *y*.

y	**1**	**4**	**3**	**7**	**8**	**7**	**10**	**12**	**15**	**20**
x	**9**	**2**	**10**	**7**	**5**	**6**	**4**	**8**	**1**	**3**

Se pide:

1) Cree un marco de trabajo adecuado e introduzca los datos.

Hay dos maneras de introducir los datos en R. La primera consiste en crear las variables X e Y como vectores con los códigos:

```
x <- c(9,2,10,7,5,6,4,8,1,3)
y <- c(1,4,3,7,8,7,10,12,15,20)
```

Para agrupar los datos de estas variables en un *data.frame* de nombre *df*, ejecutamos el código: *df <- data.frame(x,y)*. Si escribimos *df* en la consola y ejecutamos (*Intro*), veremos el contenido de este data.frame, Figura 1.1.

Figura 1.1: df

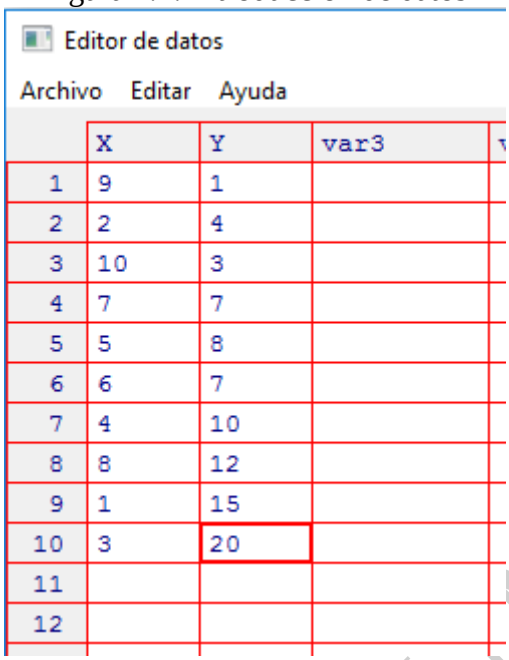
	x	y
1	9	1
2	2	4
3	10	3
4	7	7
5	5	8
6	6	7
7	4	10
8	8	12
9	1	15
10	3	20

La segunda manera de introducir datos en R consiste en crear primero un *data.frame* que contenga las variables X e Y con el código:

```
df2 <- data.frame(X = numeric(0),Y = numeric(0))
```

A este *data.frame* le hemos llamado *df2*. En este momento *df2* está vacío y para introducir los datos ejecutamos el código: *df2 <- edit(df2)*. Se nos abrirá la ventana que aparece en la Figura 1.2, en la cual debemos introducir los datos del mismo modo que en una hoja de cálculo.

Figura 1.2: Introducción de datos



	X	Y	var3
1	9	1	
2	2	4	
3	10	3	
4	7	7	
5	5	8	
6	6	7	
7	4	10	
8	8	12	
9	1	15	
10	3	20	
11			
12			

2) Obtenga los principales estadísticos descriptivos de la variable *y*.

Para responder a esta pregunta ejecutamos el código: *summary(df$y)*. La información con la respuesta nos aparece en la ventana inferior izquierda, Figura 1.3. El signo *$* nos permite seleccionar una variable o columna contenida en el data.frame *df*.

Figura 1.3: summary(df$y)

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.00    4.75    7.50    8.70   11.50   20.00
```

El operador summary nos proporciona el valor mínimo, el primer cuartil, la mediana, la media, el tercer cuartil y valor máximo de la variable *y*.

3) Calcule e interprete la covarianza y la correlación muestrales entre las dos variables.

Sea S_{XY} la covarianza entre *x* e *y*, y sea S_{XY}^C la cuasi-covarianza entre *x* e *y*.

$$S_{XY} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{N} \qquad S_{XY}^C = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{N-1}$$

Con el operador *cov* se obtiene la matriz de cuasivarianzas-cuasicovarianzas. Aplicando este operador al data.frame *df*, *cov(df)*, obtenemos:

Figura 1.4: cov(df)

```
          x         y
x  9.166667 -9.944444
y -9.944444 33.344444
```

Dado que $N \cdot S_{XY} = (N-1) \cdot S^{C}_{XY}$, tenemos que $S_{XY} = (N-1) \cdot S^{C}_{XY} / N$. Por tanto, ejecutando el código *cov(df)*9/10* obtenemos la matriz de varianzas-covarianzas:

Figura 1.5: cov(df)*9/10

```
     x     y
x  8.25 -8.95
y -8.95 30.01
```

Con el operador *cor* obtenemos la correlación entre ambas variables. De modo que ejecutando el código *cor(df)* obtenemos:

Figura 1.6: cor(df)

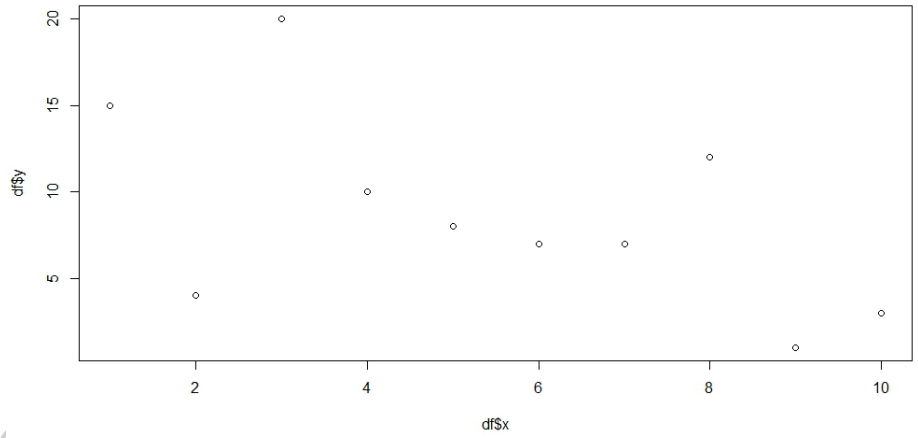
```
           x          y
x  1.0000000 -0.5688046
y -0.5688046  1.0000000
```

Notar que estos estadísticos muestran una relación negativa entre las variables *x* e *y*.

4) Represente gráficamente la nube de puntos y la recta de regresión de *y* sobre *x*.

Para obtener la nube de puntos o gráfico de dispersión entre *x* e *y* utilizamos el operador *plot*. Con el código *plot(dfx,dfy)* obtenemos el gráfico de dispersión entre *x* e *y*:

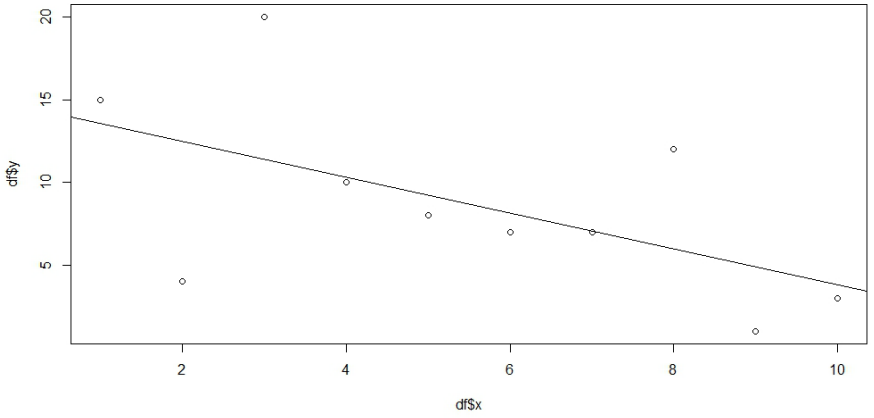
Figura 1.7: plot(dfx,dfy)



Recordar que el signo *$* permite seleccionar una variable o columna contenida en un data.frame. Por último, para incluir la recta de regresión en la nube de puntos de la Figura 1.7, ejecutamos el código: *abline(lm(y~x,data = df))*.[1] De este modo obtenemos el gráfico que aparece en la Figura 1.8.

[1] El símbolo ~ se escribe en R y RStudio pulsando simultáneamente Ctrl, Alt y 4. Es decir, Ctrl + Alt + 4.

Figura 1.8: abline(lm(y~x,data = df))



En este caso hemos utilizado dos operadores *abline* y *lm*. El operador *abline* permite añadir una curva a un gráfico obtenido previamente, y, el operador *lm* permite estimar una recta de regresión con las variables contenidas en el data.frame *df*. En este caso, la variable *y* es la variable dependiente o explicada, y la *x* la independiente o explicativa. Notar que la recta de regresión tiene pendiente negativa.

Ejercicio 2. Queremos examinar la distribución del PIB en una muestra de países. Para ello, contamos con el fichero de Excel PIB.xlsx que contiene información de 183 países sobre el PIB de 2013 en millones de dólares PPP y sobre la población en millones de habitantes de ese mismo año. Los datos se han extraído de la base de datos del programa de desarrollo de la Naciones Unidas.[2] Se pide

1) Importe a R el fichero de Excel.

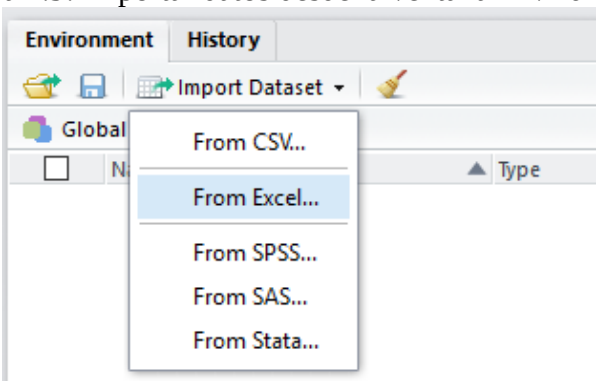
Para importar datos a R, debemos tener en cuenta el formato del fichero que vamos a importar. Es decir, si es un fichero de texto simple, es un fichero en formato CSV, Excel, SPSS, Stata, SAS… Dado que el formato del fichero en el que están los datos de este ejercicio es Excel, a continuación vamos a ver en detalle como importar datos a R desde el fichero PIB.xlsx. El proceso para importar datos a R desde ficheros con formato distinto es análogo.

Hay varios modos de importar datos desde un fichero Excel: i) a través de la ventana *Environment*, ii) a través de la ventana situada debajo de la ventana *Environment*, y iii) directamente desde la consola de R.

Si vamos a importar datos desde la ventana *Environment*, pulsamos la pestaña *Import Dataset*, como aparece en la Figura 1.9, y seleccionamos el formato del fichero del que vamos a importar los datos. En nuestro caso Excel.

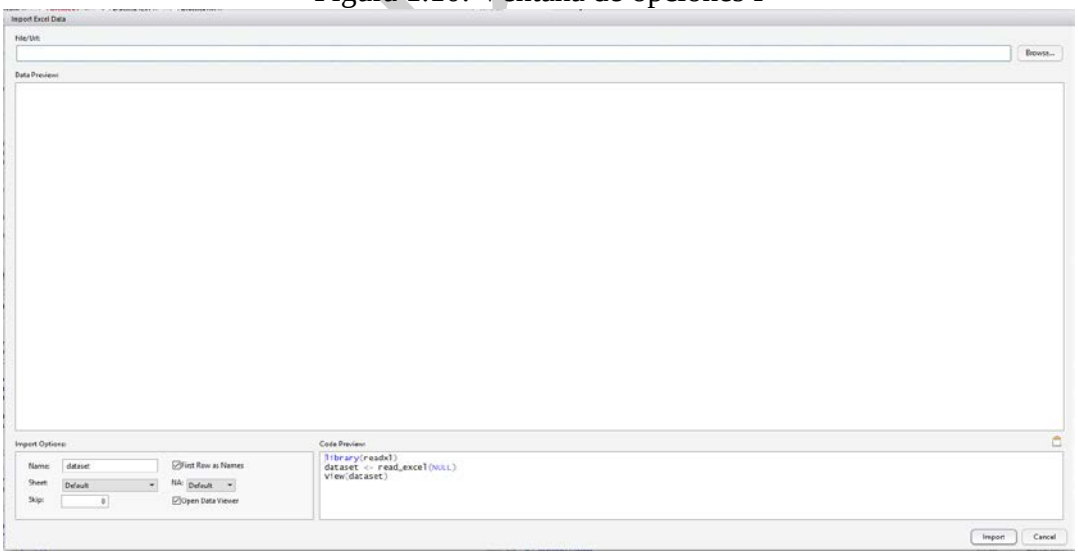
[2] http://hdr.undp.org/es/composite/HDI

Figura 1.9: Importar datos desde la ventana *Environment*



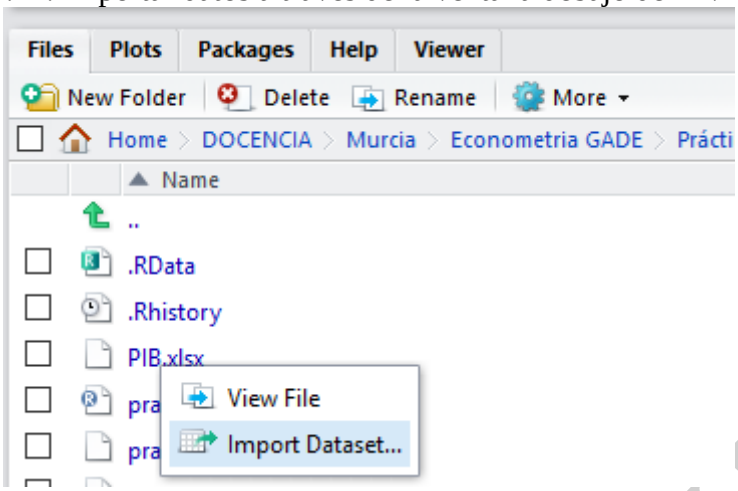
A continuación se nos abre una nueva ventana. Como se puede ver en la Figura 1.10, esta ventana consta de 4 partes. En *File/Url* debemos indicar el lugar en el que está ubicado el fichero que vamos a importar (se puede hacer a través de la pestaña *Browse…*) o la dirección de Internet de la página web que contenga los datos. En este último caso, el ordenador que utilicemos debe estar conectado a Internet y el servidor que aloja el fichero con los datos debe funcionar y no tiene que requerir una contraseña para acceder. En *Data Preview* podemos ver previamente nuestros datos. *Import Options* nos muestra las distintas opciones que debemos tener en cuenta para importar datos. Finalmente, en *Code Preview*, nos aparece el código necesario para importar nuestro fichero Excel. Una vez hemos incorporada toda la información relativa al fichero que vamos a importar pulsamos *Import*, y nos aparecerá dicho fichero en la ventana *Environment*. Para guardar el fichero pulsamos el icono que aparece en la ventana *Environment*,  . El fichero se guardará en formato *.RData* y aparecerá como un *data.frame*.

Figura 1.10: Ventana de opciones I



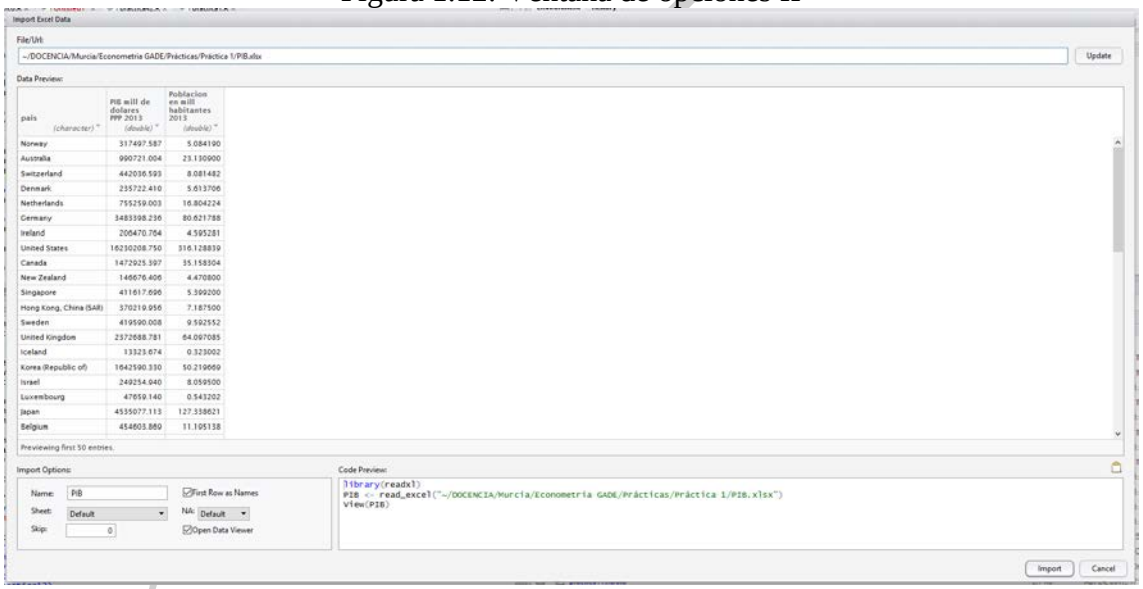
Para importar datos a través de la ventana situada debajo de la ventana *Environment*, lo primero que debemos hacer si no lo hemos hecho previamente, es fijar como directorio de trabajo la carpeta que contenga el fichero. A continuación pulsamos el fichero con los datos y nos aparecerá una ventana con dos opciones. Seleccionamos *Import Dataset* (ver Figura 1.11).

Figura 1.11: Importar datos a través de la ventana debajo de *Environment*



Después de seleccionar *Import Dataset*, nos aparecerá la ventana de la Figura 1.10, pero en esta ocasión con la información necesaria para importar el fichero de datos que hemos seleccionado (ver Figura 1.12). Si nos parece correcta la vista previa de nuestro fichero de datos pulsamos *Import*, y nos aparecerá dicho fichero en la ventana *Environment*. Igual que hemos indicado arriba para guardar el fichero pulsamos el icono , el cual aparece en la ventana *Environment*.

Figura 1.12: Ventana de opciones II



Como podemos ver en la Figura 1.12 los nombres de las variables que contienen el PIB y la población de cada país son largos y engorrosos. Una vez creado el archivo en la ventana *Environment*, podemos indicar a R que nos muestre el nombre de las variables que contiene el data.frame PIB con el código *names(PIB)*. A continuación vamos a modificar los nombres de las variables del fichero, de modo que la variable que contiene el PIB de cada país se llame *pib* y la que contiene su población se llame *poblacion*. Para ello ejecutamos el siguiente código: *names(PIB) = c("pais", "pib", "poblacion")*.

Para ver cómo queda el fichero después de este cambio podemos ejecutar el código *View(PIB)* o el código *head(PIB)*. Si ejecutamos *View(PIB)* aparece una hoja con los datos del data.frame *PIB* en la ventana superior-izquierda, pero si ejecutamos *head(PIB)* obtendremos el contenido de las primeras seis filas de dicho data.frame como aparece en la Figura 1.13. Por defecto, R muestra las seis primeras filas, pero si deseamos que nos muestre un número distinto debemos indicárselo a R. Por ejemplo, si queremos ver ocho filas ejecutamos el código: *head(PIB,8)*. Por otro lado, si queremos ver las seis últimas filas del data.frame *PIB* ejecutamos: *tail(PIB)*.

Figura 1.13: head(PIB) I

	pais	pib	poblacion
1	Norway	317497.6	5.084190
2	Australia	990721.0	23.130900
3	Switzerland	442036.6	8.081482
4	Denmark	235722.4	5.613706
5	Netherlands	755259.0	16.804224
6	Germany	3483398.2	80.621788

Para importar datos directamente desde la consola de R debemos tener en cuenta el formato del fichero. Pues dicho formato determinará el operador que vamos a utilizar para importar los datos. Siguiendo con nuestro ejemplo de importar datos desde un fichero Excel, debemos usar el operador *read_excel* e indicar donde está ubicado el fichero Excel. Para utilizar dicho operador debemos previamente cargar el paquete *readxl*, con el código *library(readxl)*. Si el fichero de datos está ubicado en nuestro directorio de trabajo, no es necesario indicar la ubicación del fichero. Por tanto, para importar los datos y guardarlos en el objeto PIB ejecutamos el siguiente código:

PIB <- read_excel("PIB.xlsx")

2) Genere una variable que contenga al logaritmo del PIB.

A continuación vamos a generar la variable *lpib*. Esta variable contendrá el logaritmo del PIB de cada país. Para ello ejecutamos: *PIB$lpib = log(PIB$pib)*. Notar que la variable *lpib* ha sido directamente incorporada al data.frame *PIB*. Para comprobar que hemos creado la variable correctamente ejecutamos *head(PIB)* (ver Figura 1.14). También se puede comprobar en la ventana *Environment*.

Figura 1.14: head(PIB) II

	pais	pib	poblacion	lpib
1	Norway	317497.6	5.084190	12.66823
2	Australia	990721.0	23.130900	13.80619
3	Switzerland	442036.6	8.081482	12.99915
4	Denmark	235722.4	5.613706	12.37041
5	Netherlands	755259.0	16.804224	13.53482
6	Germany	3483398.2	80.621788	15.06352

3) Genere una variable que contenga al PIB por habitante.

Para generar la variable que contenga el PIB por habitante, la cual llamaremos *pibpc*, ejecutamos: *PIB$pibpc = PIB$pib/PIB$poblacion*. De este modo la variable *pibpc* directamente se incorporará al data.frame *PIB*, como podemos comprobar ejecutando el código *head(PIB)* o viendo la ventana *Environment*.

4) Sabiendo que los primeros 102 países tienen un nivel de desarrollo medio o alto, calcule la media del PIB por habitante para estos países, y para los países con menor nivel de desarrollo. Calcule también la media para los países con población mayor de 80 millones de habitantes. Compare y comente las medias de los países más y menos desarrollados y la de los países más poblados.

Para calcular la media del PIB por habitante de los primeros 102 países que aparecen en el data.frame PIB ejecutamos: *mean(PIB[1:102,]$pibpc)*. Analicemos este código. Con *PIB[1:102,]* indicamos a R que seleccione las 102 primeras filas del data.frame *PIB*. Con *$pibpc* indicamos a R que seleccione la variable PIB por habitante. Finalmente, el operador *mean* nos permite calcular la media que buscamos, la cual es 26918.84 dólares.

Para calcular la media del PIB por habitante del resto de países que aparecen en el data.frame *PIB* ejecutamos: *mean(PIB[103:183,]$pibpc)*. Y obtendremos que el PIB por habitante medio baja a 4520.688 dólares para los países con menor nivel de desarrollo.

Por último, calculamos el PIB por habitante medio para aquellos países con una población mayor de 80 millones de habitantes. Para ello ejecutamos el código: *mean(PIB$pibpc[PIB$poblacion > 80])*. Analicemos este código. Con *PIB$pibpc* indicamos a R que seleccione la variable PIB por habitante del data.frame *PIB*. Y con *[PIB$poblacion > 80]* seleccionamos aquellos países cuya población es superior a 80 millones de habitantes. Notar que la unidad de la variable POBLACION es millones de habitantes. Por cierto, la media para los países con una población mayor de 80 millones es 15427.43 dólares.