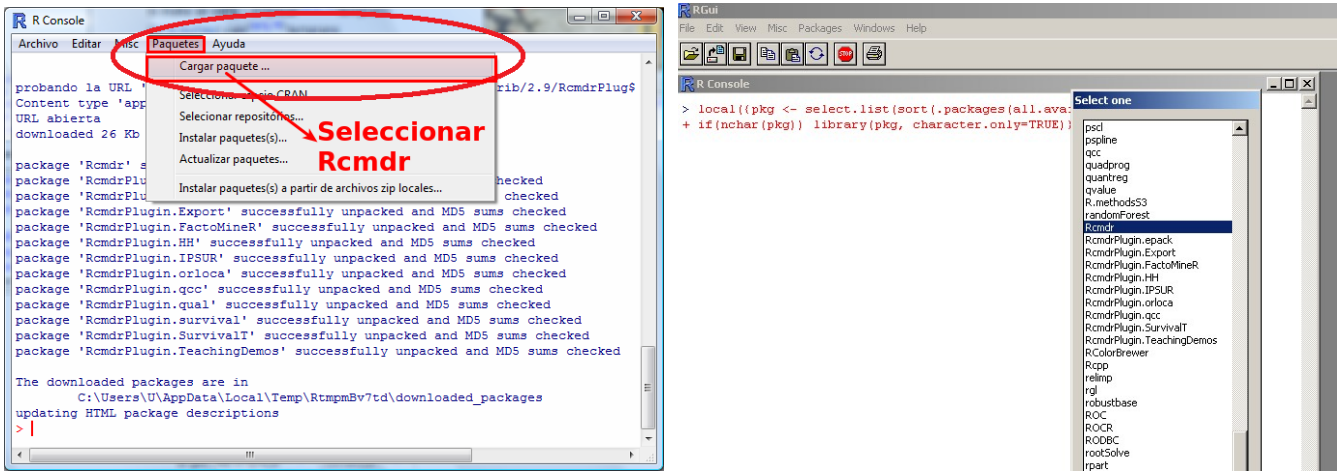
Práctica 5. Contrastes de independencia y correlación lineal

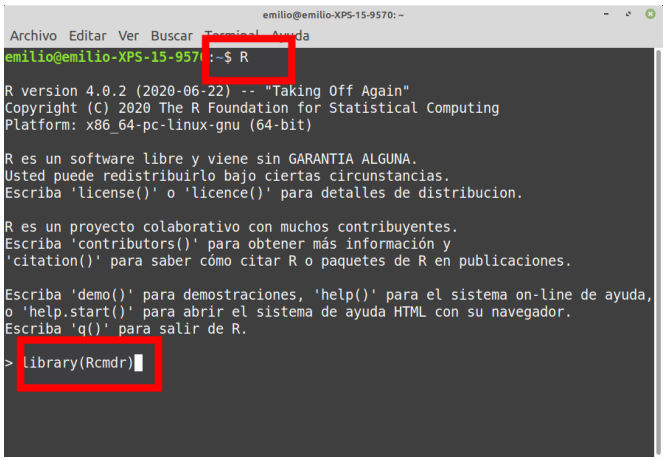
Departamento de Estadística
Universidad de Oviedo

Cargar el programa R y el paquete Rcommander

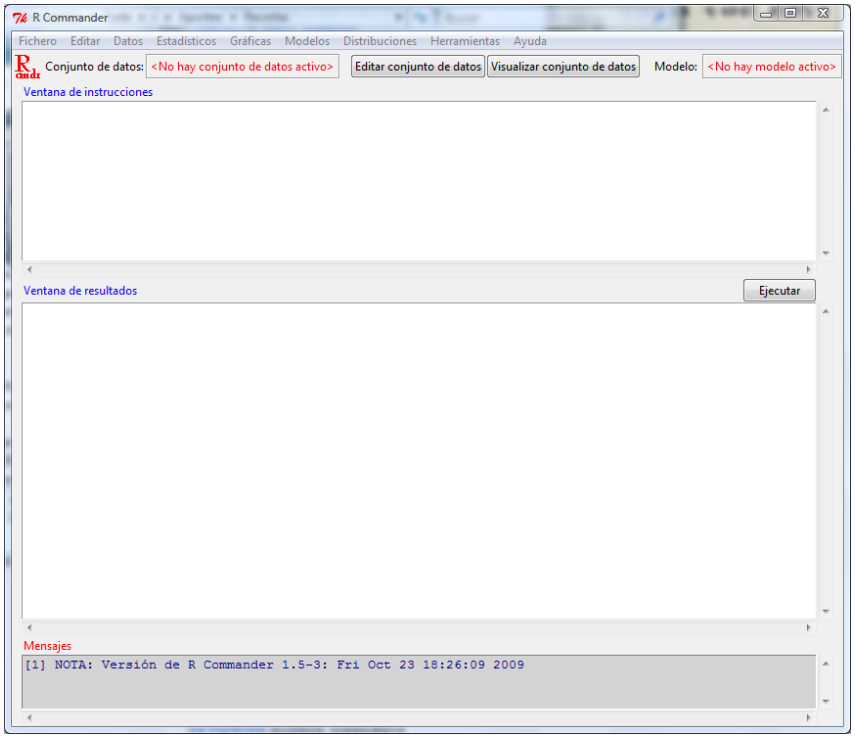
- Iniciamos el programa R.
- Cargamos el paquete RCommader. Dos opciones:
  1. Menú *Paquetes* → *Cargar paquete* → Seleccionamos **Rcmdr**.

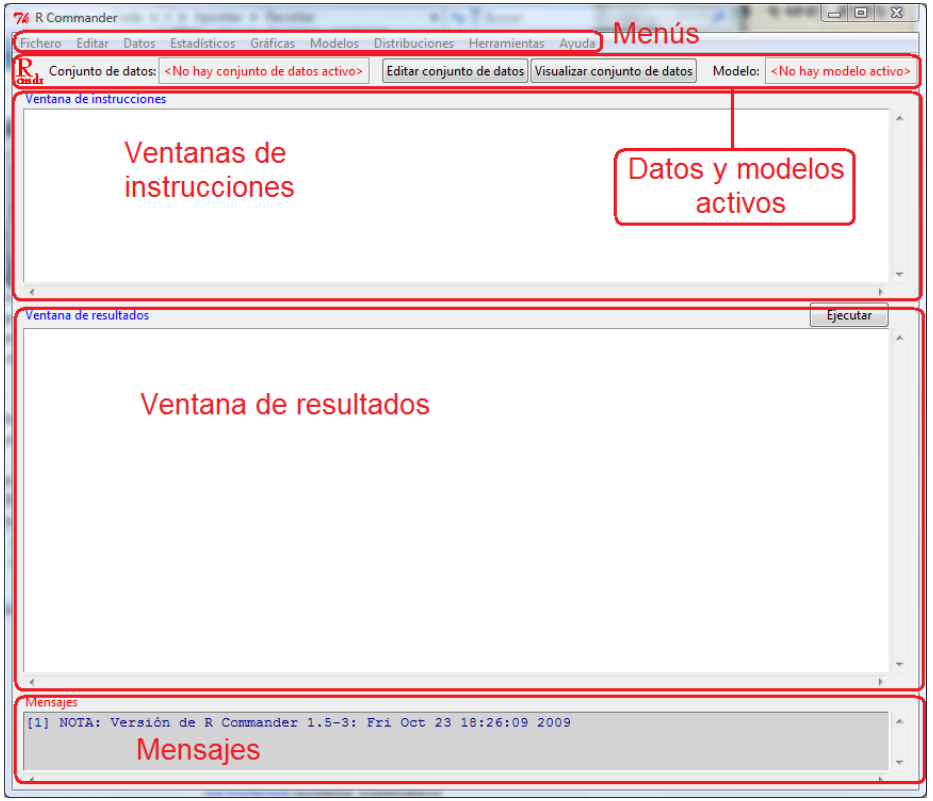


  2. Escribimos **library(Rcmdr)** en la consola y pulsamos retorno de carro.



Cargar Rcommander





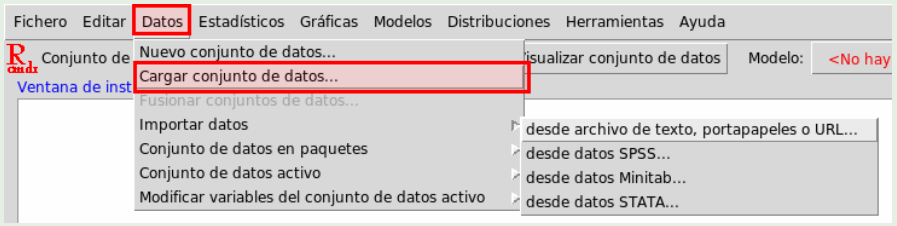
El fichero `acero.rda` se encuentra en el Campus Virtual. Hay que haberlo descargado previamente.

Cargar la base de datos `acero.rda`

Datos

➥ Cargar conjunto de datos

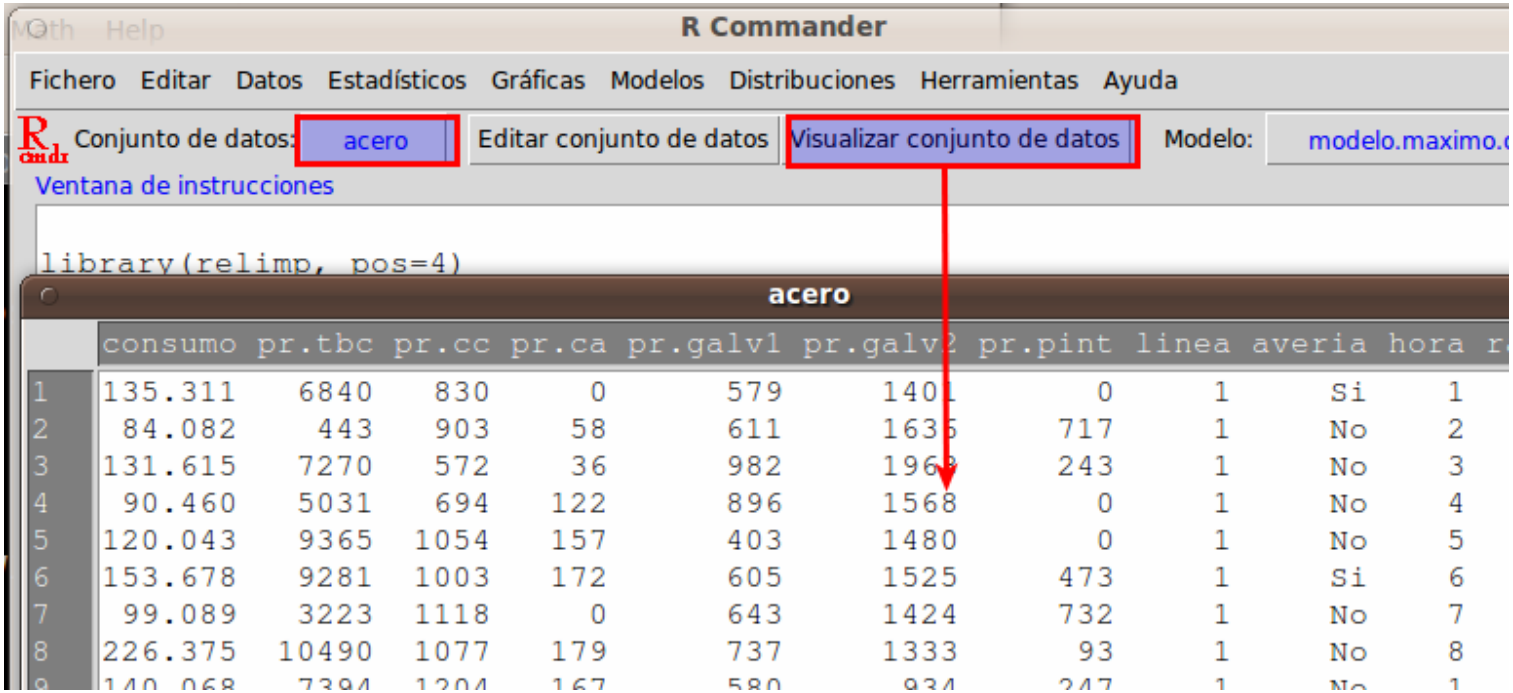
➥ Seleccionar **acero.rda**



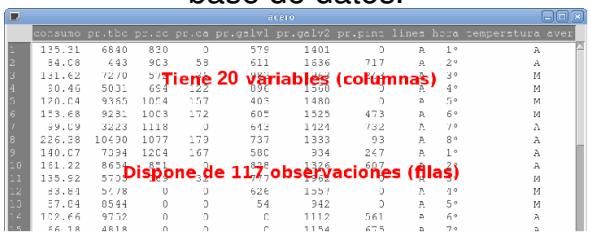

```
> load("/home/emilio/clases/acero.rda")
```

```
NOTA: El conjunto de datos acero tiene 117 filas y 20 columnas.
```

- Para visualizar la base de datos:



Aparece una ventana con los datos disponibles. Moviendo el cursor hacia la izquierda o hacia abajo podemos recorrer toda la base de datos.



Variables de la base de datos `acero`

1. `consumo` Consumo energético de la empresa (Megavatios/hora).
2. `pr.tbc` Producción del tren de bandas calientes (Toneladas de acero).
3. `pr.cc` Producción de colada continua (Toneladas de acero).
4. `pr.ca` Producción del convertidor de acero (Toneladas de acero).
5. `pr.galv1` Producción de galvanizado de tipo I (Tns. de acero).
6. `pr.galv2` Producción de galvanizado de tipo II (Tns. de acero).
7. `pr.pint` Producción de chapa pintada (Tns. de acero).
8. `linea` Línea de producción empleada (A o B).
9. `turno` Turno de mañana (M), tarde (T), noche (N).
10. `temperatura` Temperatura del sistema: Alta, Media y Baja.
11. `pres.aver` Presencia de averías: hubo Averías (A), no hubo averías (NoA).
12. `nun.aver` Número de averías detectadas.
13. `sistema` Activación de un sistema de detección de sobrecalientamiento: encendido (ON), apagado (OFF).
14. ...

Analizar dos variables simultáneamente

Dos tipos de variables: nominal (sexo, ciudad) y continua (gasto, tiempo)

Sexo	Ciudad	Gasto (euros)	Tiempo (minutos)
Hombre	Gijón	10	240
Mujer	Oviedo	18	231
Hombre	Oviedo	19	125
...	...	...	...

Si seleccionamos dos variables, hay tres posibilidades según sea el tipo de cada variable:

1. una continua (gasto) y otra nominal (sexo) $\rightarrow$ Contraste de medias
2. las dos son nominales (sexo y ciudad) $\rightarrow$ Test de independencia
3. las dos son continuas (gasto y tiempo) $\rightarrow$ correlación

Tests de independencia

Dos variables nominales

¿Son independientes dos variables de tipo nominal?

¿Existe relación entre vivir en Vetusta y ser aficionado del club de fútbol Real Oviedo, S.A.D.?

H_0:	hay independencia estadística entre las dos variables
H_1:	hay dependencia estadística entre las dos variables

Test χ^2 (chi-cuadrado) de independencia:

- Si p < 0.05, rechazamos H_0; por lo tanto, hay relación entre ambas variables.
- Si p $\geq$ 0.05, no rechazamos la independencia. No hay evidencia suficiente para rechazar la independencia entre ambas variables.

¿Existe relación entre que haya o no averías y la temperatura? o, dicho de otro modo, ¿la proporción de averías depende de la temperatura?

Variables cualitativas: utilizar el *Test de independencia Chi-cuadrado*

Estadísticos

➥Tablas de contingencias

➥Tabla de doble entrada...

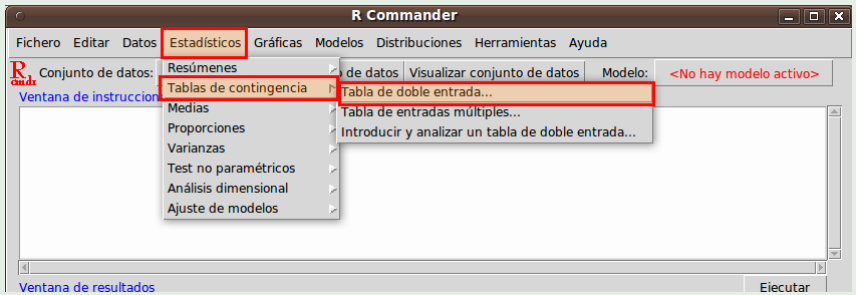
Seleccionar

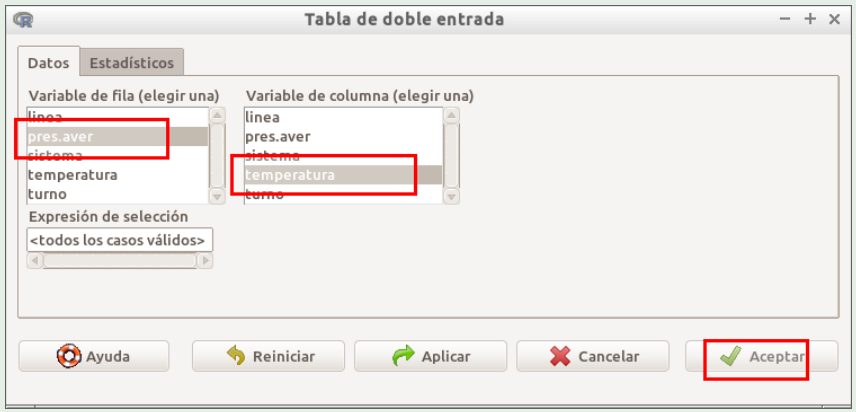
variable fila: `pres.aver`

variable columna: `temperatura`

➥Marcar **Test de independencia Chi-cuadrado**

➥Aceptar





(Sigue →)

H_0: hay independencia estadística entre las dos variables
H_1: hay dependencia estadística entre las dos variables

```
Frequency table:
          temperatura
pres.aver Alta Baja Media
      A      8   14     6
      NoA   38   24    27

        Pearson's Chi-squared test

data:  .Table
X-squared = 5.1595, df = 2, p-value = 0.07579
```

Como el p-valor (0.07199) es mayor que cualquier nivel de significación habitual α, no se rechaza la hipótesis nula. Por lo tanto concluimos que no hay evidencias estadísticas de que la temperatura afecte a que haya o no averías.

(Sigue →)

```
Frequency table:
          temperatura
pres.aver Alta Baja Media
       A     8   14     6
       NoA  38   24    27
```

¿Cómo se interpreta el valor 8?:
En 8 ocasiones se detectó que había averías cuando la temperatura era alta.

Calcule e interprete los siguientes porcentajes: a) 6.8%; b) 17.4%; and c) 27.6%

a)

	Alta	Baj.	Med.	
A	**8**	14	6	28
NoA	38	24	27	89
	46	38	33	**117**

$$\frac{8}{117} \cdot 100 = 6.8\%$$

b)

	Alta	Baj.	Med.	
A	**8**	14	6	28
NoA	38	24	27	89
	46	38	33	117

$$\frac{8}{46} \cdot 100 = 17.4\%$$

c)

	Alta	Baj.	Med.	
A	**8**	14	6	**28**
NoA	38	24	27	89
	46	38	33	117

$$\frac{8}{28} \cdot 100 = 28.6\%$$

Repetir este ejercicio marcando la opción de *frecuencias esperadas*

- Estadísticos
  ➥Tablas de contingencias
    ➥Tabla de doble entrada...
- Seleccionar
  variable fila: `pres.aver`
  variable columna: `temperatura`
  ➥Marcar **Imprimir las frecuencias esperadas**
    ➥Aceptar

```
Expected counts:
          temperatura
pres.aver     Alta      Baja     Media
       A   11.00855  9.094017  7.897436
       NoA 34.99145 28.905983 25.102564
```

¿Cómo se interpreta 11.008?
Bajo condiciones de independencia, el número esperado de observaciones cuando había averías y la temperatura alta sería de 11.008 casos.

Repetir este ejercicio marcando la opción de *Porcentajes totales*

- Estadísticos
  ➥Tablas de contingencias
  ➥Tabla de doble entrada...
- Seleccionar
  variable fila: `pres.aver`
  variable columna: `temperatura`
  ➥Marcar **Porcentajes totales**
  ➥Aceptar

```
> totPercents(.Table) # Percentage of Total
Total percentages:
      Alta Baja Media Total
A      6.8 12.0   5.1  23.9
NoA   32.5 20.5  23.1  76.1
Total 39.3 32.5  28.2 100.0
```

¿Cómo se interpreta 6.8?
El número de observaciones cuando había averías y la temperatura era alta fue del 6.8%.

Repetir este ejercicio marcando la opción de *Porcentajes por filas*

- Estadísticos
  ➥Tablas de contingencias
  ➥Tabla de doble entrada...
- Seleccionar
  variable fila: `pres.aver`
  variable columna: `temperatura`
  ➥Marcar **Porcentajes por filas**
  ➥Aceptar

```
Row percentages:
          temperatura
pres.aver Alta Baja Media Total Count
       A   28.6   50  21.4   100    28
       NoA 42.7   27  30.3   100    89
```

¿Cómo se interpreta 28.6?
El 28.6% de las veces que había averías se produjeron cuando la temperatura era alta.

Repetir este ejercicio marcando la opción de *Porcentajes por columnas*

- Estadísticos
  ➥Tablas de contingencias
  ➥Tabla de doble entrada...
- Seleccionar
  variable fila: `pres.aver`
  variable columna: `temperatura`
  ➥Marcar **Porcentajes por columnas**
  ➥Aceptar

```
Column percentages:
          temperatura
pres.aver  Alta  Baja Media
    A      17.4  36.8  18.2
    NoA    82.6  63.2  81.8
    Total 100.0 100.0 100.0
    Count  46.0  38.0  33.0
```

¿Cómo se interpreta 17.4?
En el 17.4% de las observaciones realizadas cuando la temperatura era alta, había averías.

Analizar dos variables continuas simultáneamente

Producto escalar

Gasto (euros)	Tiempo (minutos)
10	240
18	231
19	125
...	...

Sea θ el ángulo que forman estos dos vectores.
Sea $\rho = cos(\theta)$. Entonces

$$-1 \leq \rho \leq 1$$

1. Si $\rho = cos(\theta) = 1, \implies \theta = 0$. Misma dirección y sentido.
2. Si $\rho = cos(\theta) = 0, \implies \theta = \pi/2, 3\pi/2$. Ortogonales
3. Si $\rho = cos(\theta) = -1, \implies \theta = \pi$. Misma dirección y sentido contrario.

Tests de independencia

Dos variables continuas

¿Hay relación entre dos variables de tipo continuo?

¿Existe relación entre el gasto en alcohol y el gasto en telefonía?

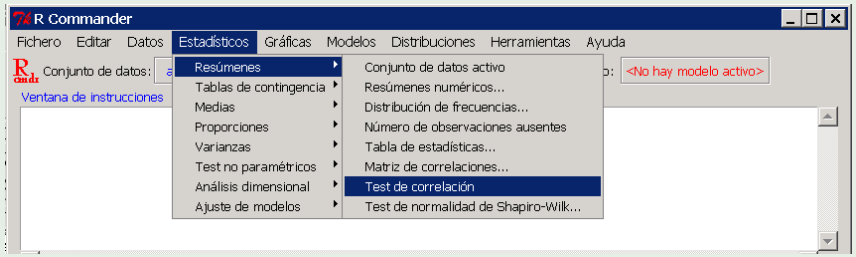
$$\begin{array}{ll} H_0: & \rho = 0 \text{ (la correlación es nula)} \\ H_1: & \rho \neq 0 \text{ (la correlación es no nula)} \end{array}$$

Test de correlación de Pearson:

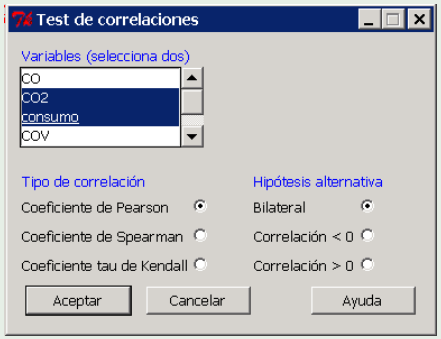
- Si p < 0.05, rechazamos H_0; por lo tanto, hay correlación entre ambas variables.
- Si p $\geq$ 0.05, no rechazamos la presencia de correlación entre ambas variables. No hay evidencia suficiente para rechazar la independencia entre ambas variables.

Con la base de datos `acero`, ¿qué se puede decir sobre si existe o no relación lineal entre el consumo energético (*consumo*) y la emisión de dióxido de carbono (*CO2*)?

Estadísticos
➥Resúmenes...
➥Test de correlación



Seleccionar las variables **CO2** y **consumo**
➥Aceptar




```
data:  CO2 and consumo
t = 35.1003, df = 115, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9376074 0.9695667
sample estimates:
      cor
0.9563613
```

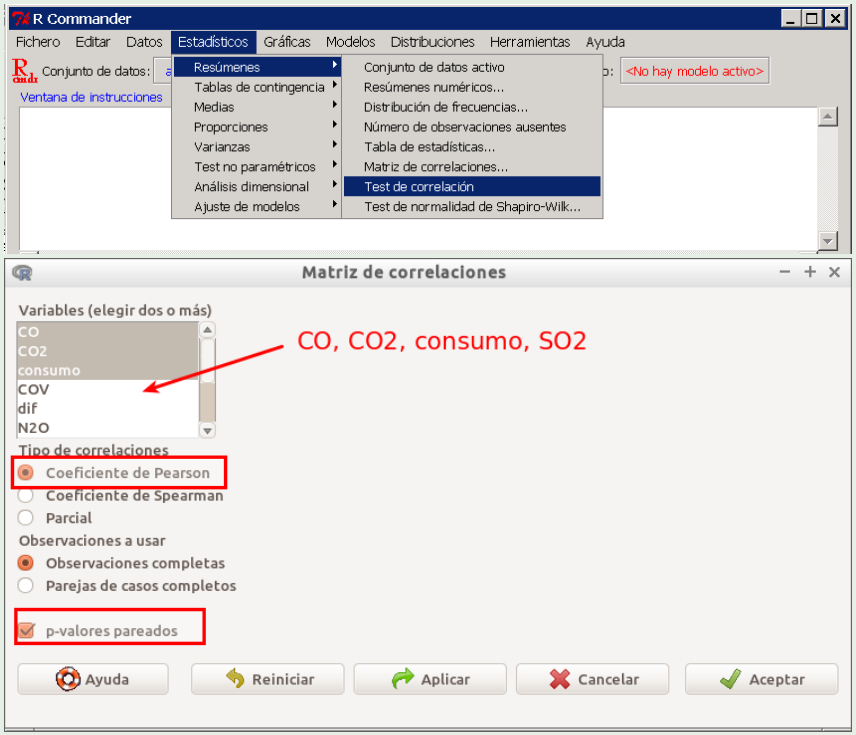
Hay evidencias estadísticas de relación lineal entre el consumo y la emisión de CO2 (es del 0.95).

Test de correlación

Analizar la relación lineal del *consumo* con la emisión de *CO*, *CO2* y *SO2*.

Estadísticos
➥Resúmenes
➥Matriz de correlaciones...

Seleccionamos **CO**, **CO2**, **consumo** y **SO2**.
➥Coeficiente de Pearson
➥p-valor pareado de las correlaciones de Pearson o Spearman
➥Aceptar



(Sigue →)

Correlaciones:

```
> rcorr.adjust(acero[,c("CO","CO2","consumo","SO2")], type="pearson")
Pearson correlations:
            CO     CO2 consumo     SO2
CO      1.0000  0.9442  0.9198  0.0444
CO2     0.9442  1.0000  0.9564 -0.0286
consumo 0.9198  0.9564  1.0000 -0.0076
SO2     0.0444 -0.0286 -0.0076  1.0000

 Number of observations: 117

 Pairwise two-sided p-values:
        CO     CO2    consumo SO2
CO              <.0001 <.0001  0.6347
CO2     <.0001         <.0001  0.7599
consumo <.0001 <.0001          0.9352
SO2     0.6347 0.7599 0.9352
```

De la primera matriz (p-valores), el alta la correlación del consumo energético y la emisión de `CO` o la de `CO2`, pero no así con la de `SO2`. La mayor relación correlación del consumo es con la emisión de `CO2`, puesto que el coeficiente de correlación (primera matriz) 0.9564.