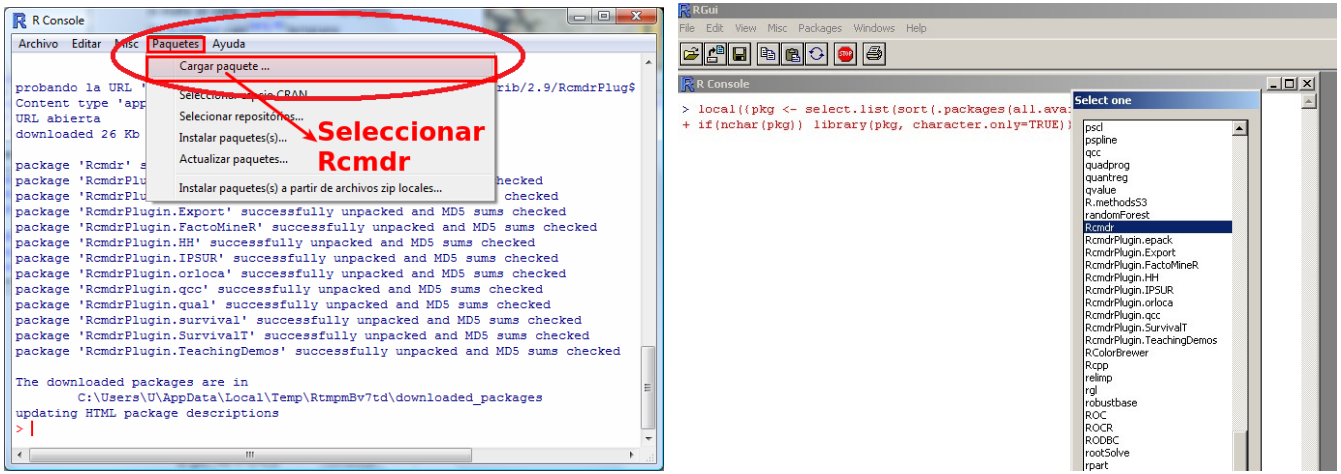
Práctica 6. Regresión lineal

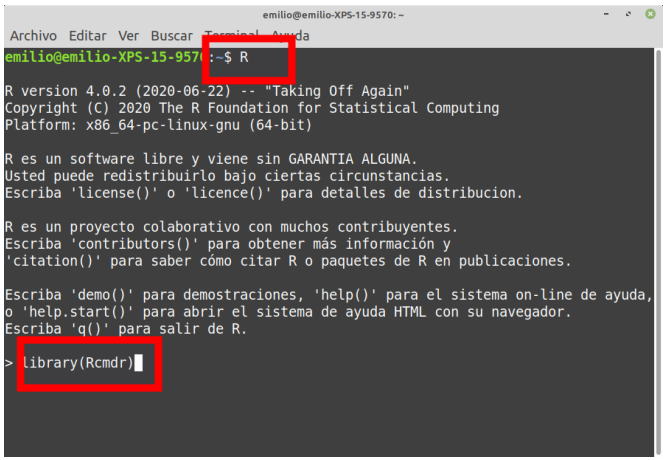
Departamento de Estadística
Universidad de Oviedo

Cargar el programa R y el paquete Rcommander

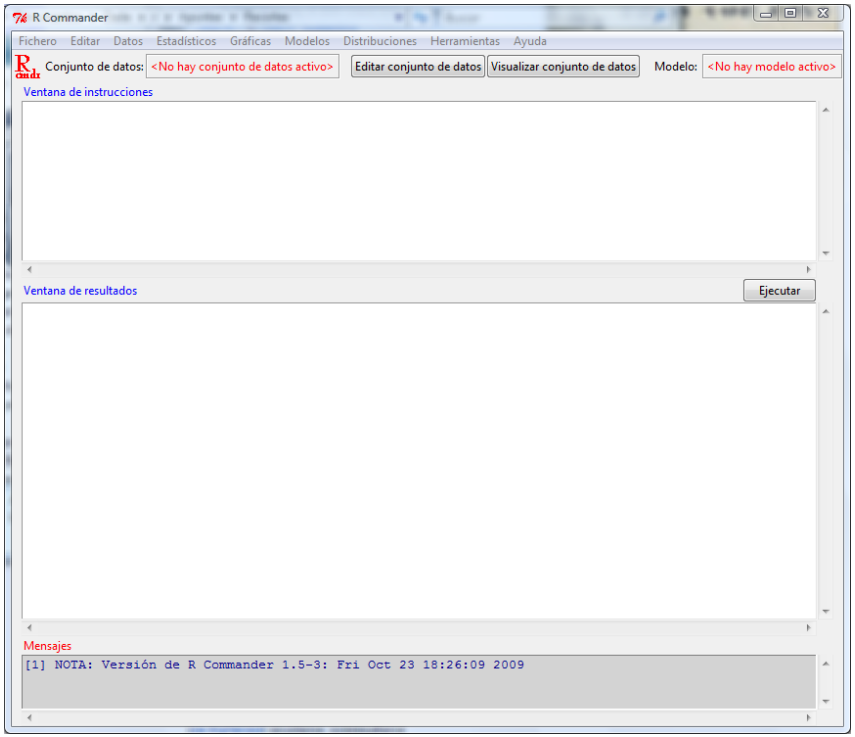
- Iniciamos el programa R.
- Cargamos el paquete RCommader. Dos opciones:
  1. Menú *Paquetes* → *Cargar paquete* → Seleccionamos **Rcmdr**.

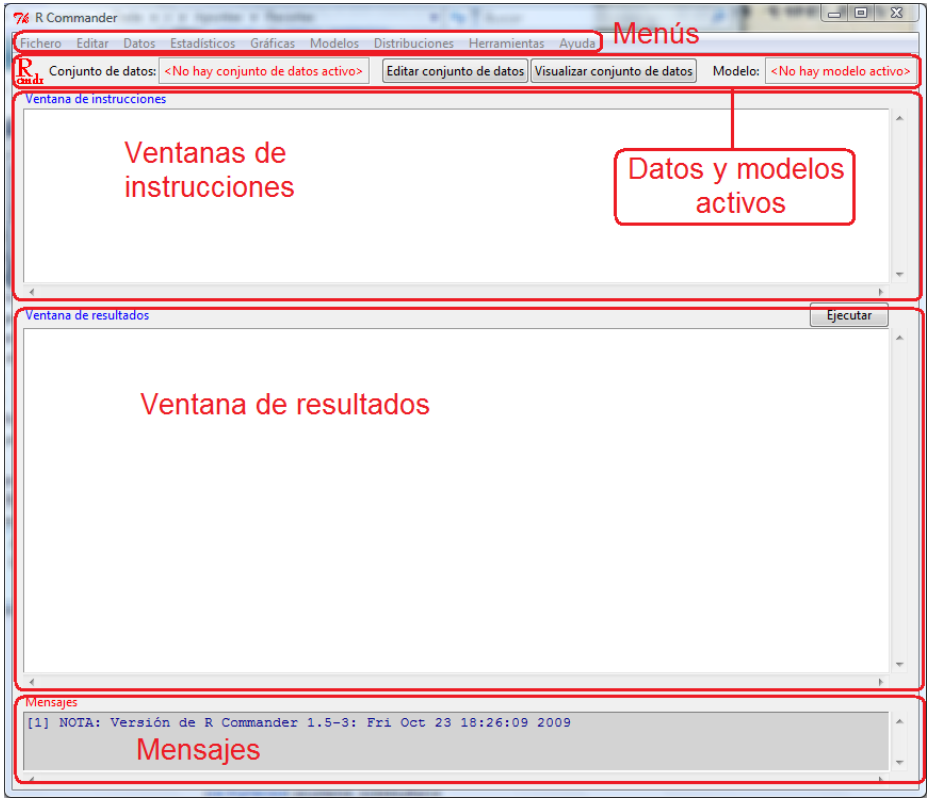


  2. Escribimos **library(Rcmdr)** en la consola y pulsamos retorno de carro.



Cargar Rcommander





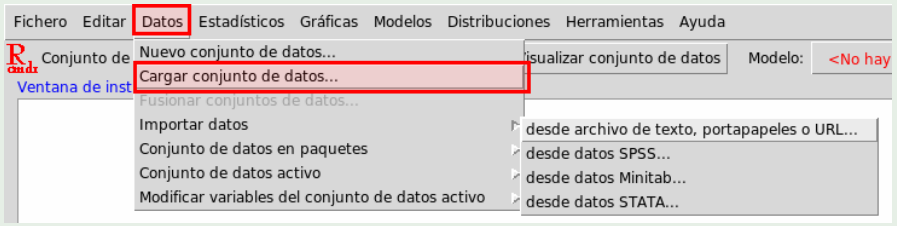
El fichero `acero.rda` se encuentra en el Campus Virtual. Hay que haberlo descargado previamente.

Cargar la base de datos `acero.rda`

Datos

➥ Cargar conjunto de datos

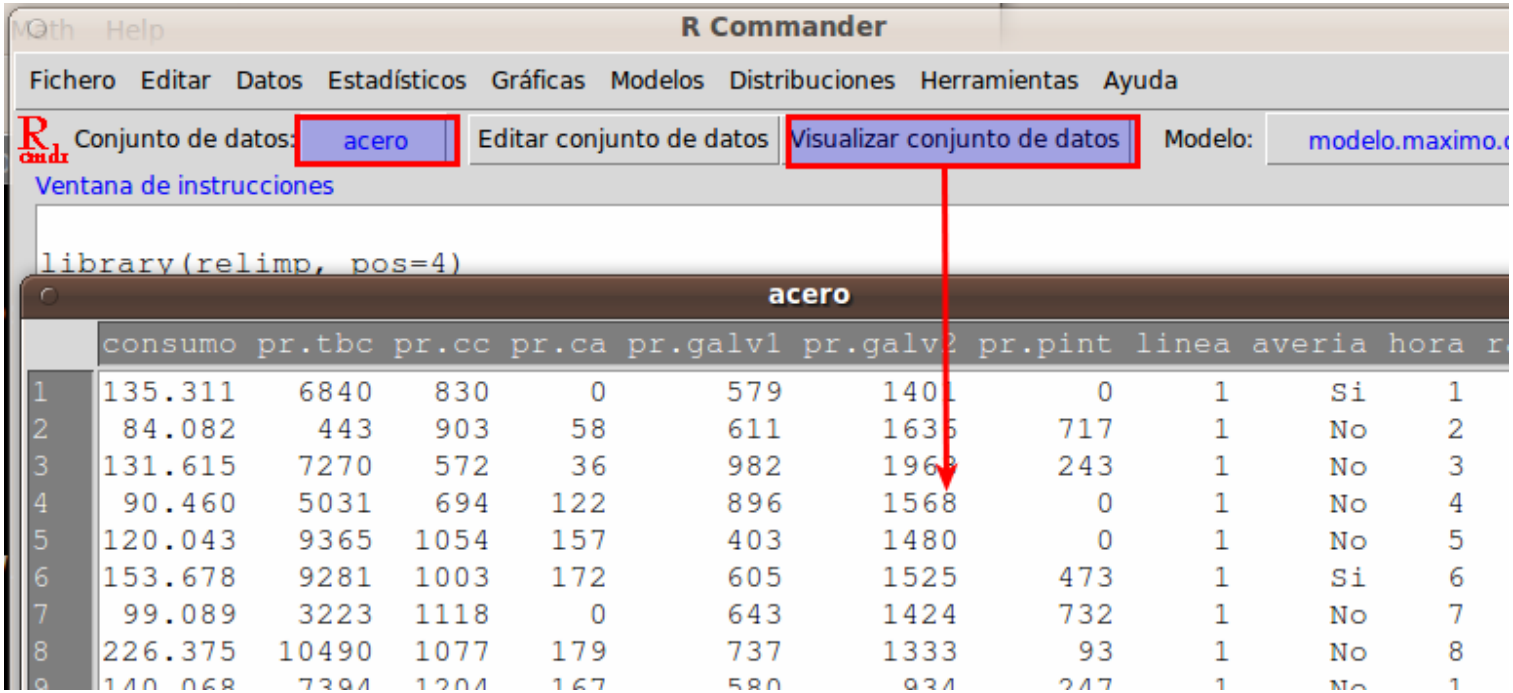
➥ Seleccionar **acero.rda**



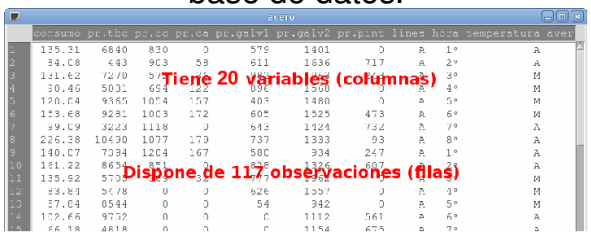

```
> load("/home/emilio/clases/acero.rda")
```

```
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`

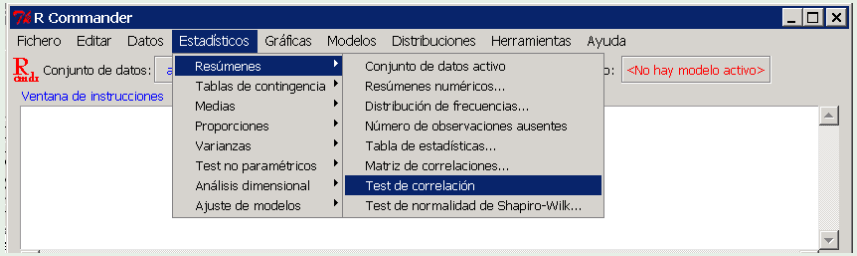
1. `consumo` Consumo energético de la empresa (Megavatios/hora).
2. `pr.tbc` Producción del tren de bandas calientes (Toneladas de acero).
3. `pr.cc` Producción de colada continua (Toneladas de acero).
4. `pr.ca` Producción del convertidor de acero (Toneladas de acero).
5. `pr.galv1` Producción de galvanizado 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. ...

Calcule la matriz de correlaciones de la emisión de óxido nitroso (N2O), óxidos de monóxido de carbono (CO), dióxido de carbono (CO2), mezcla de óxidos de nitrógeno (NOx) y dióxido de azufre (SO2)

Estadísticos

➥Resúmenes

➥Matriz de correlaciones...

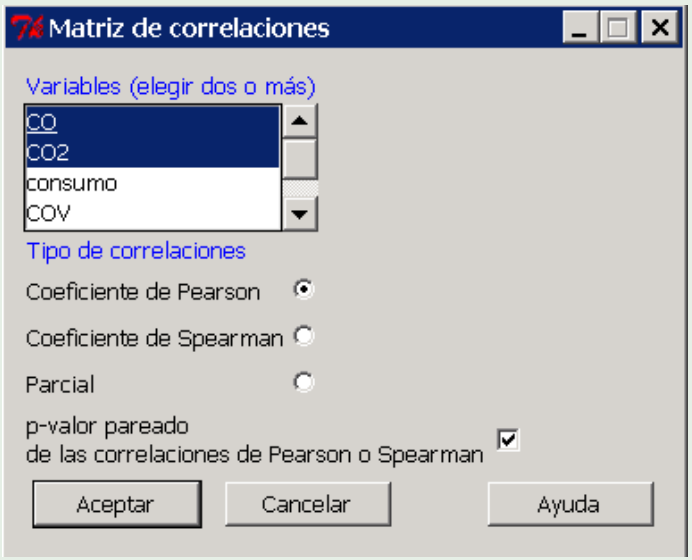


Seleccionamos CO, CO2, N2O, NOx y SO2.

➥Coeficiente de Pearson

➥p-valor pareado de las correlaciones de Pearson o Spearman

➥Aceptar



(Sigue →)

Matriz de correlaciones:

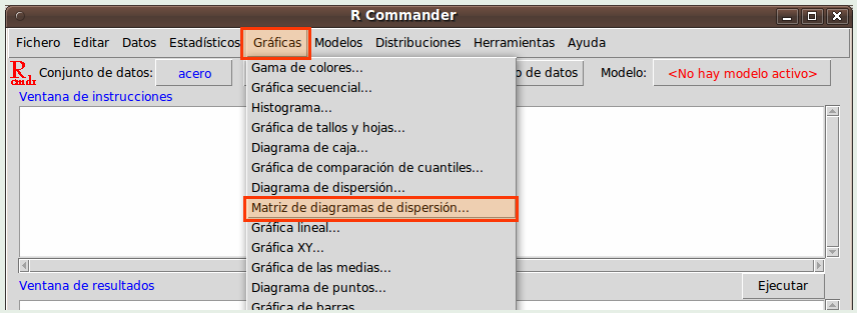
```
       CO    CO2   N2O    NOx    SO2
CO   1.00   0.94 0.82   0.52   0.04
CO2  0.94   1.00 0.85   0.57 -0.03
N2O  0.82   0.85 1.00   0.53   0.01
NOx  0.52   0.57 0.53   1.00 -0.13
SO2  0.04 -0.03 0.01 -0.13   1.00
```

¿Qué variable es la que tiene más relación con N2O?

```
       CO    CO2   N2O    NOx    SO2
CO   1.00   0.94 0.82   0.52   0.04
CO2  0.94   1.00 0.85   0.57 -0.03
N2O  0.82   0.85 1.00   0.53   0.01
NOx  0.52   0.57 0.53   1.00 -0.13
SO2  0.04 -0.03 0.01 -0.13   1.00
```

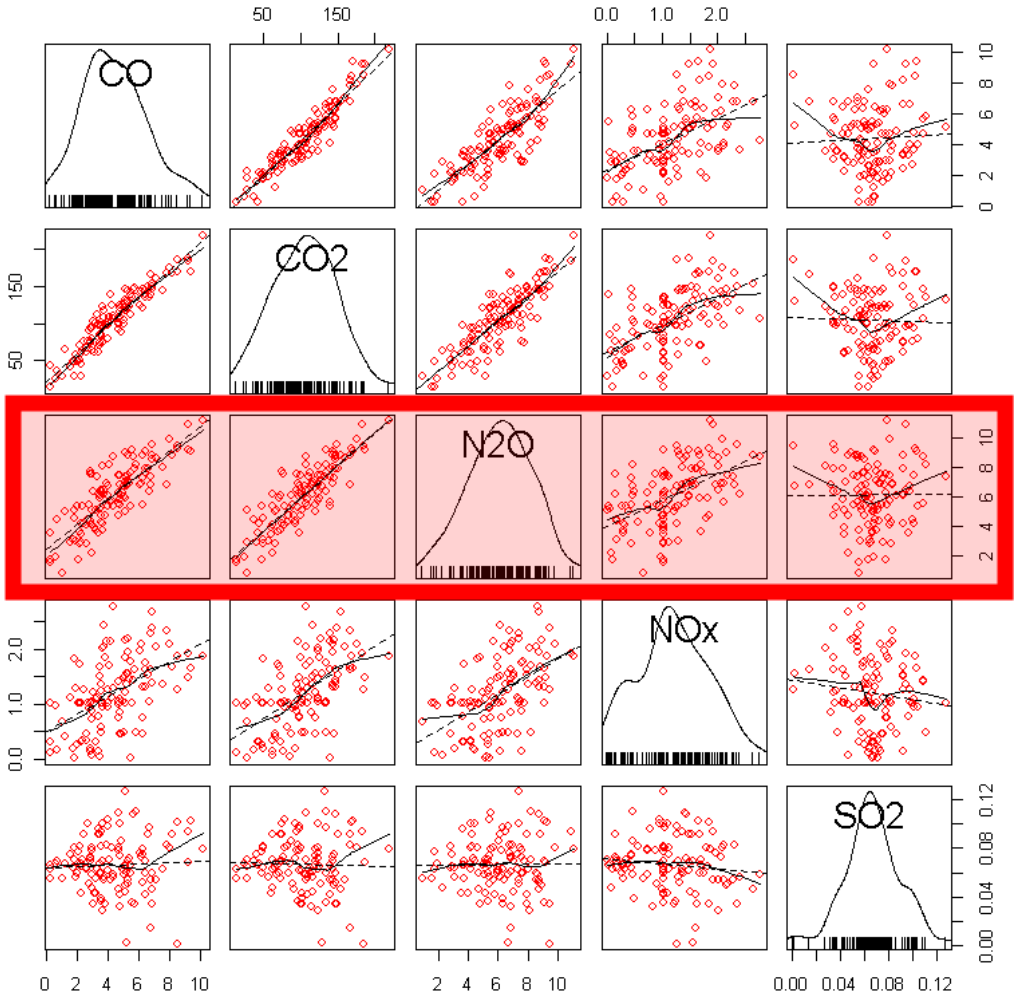
Dibuje la matriz de diagrama de dispersión de la emisión de óxido nitroso (N2O), óxidos de monóxido de carbono (CO), dióxido de carbono (CO2), mezcla de óxidos de nitrógeno (NOx) y dióxido de azufre (SO2)

Gráficas
➥Matriz de diagrama de dispersión...



Seleccionamos CO, CO2, N2O, NOx y SO2.
➥Aceptar

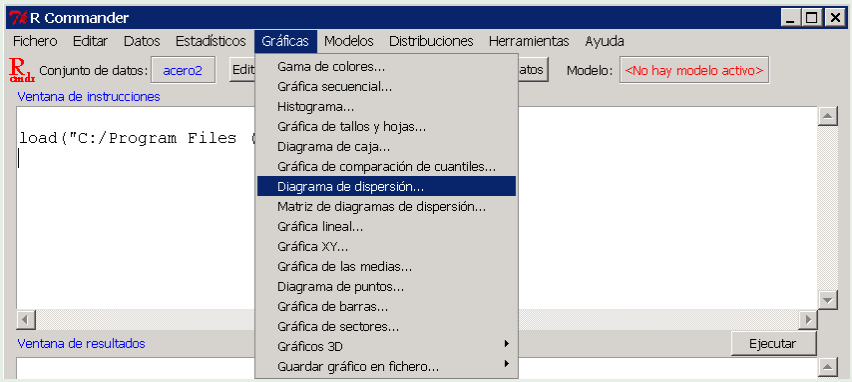
(Sigue →)



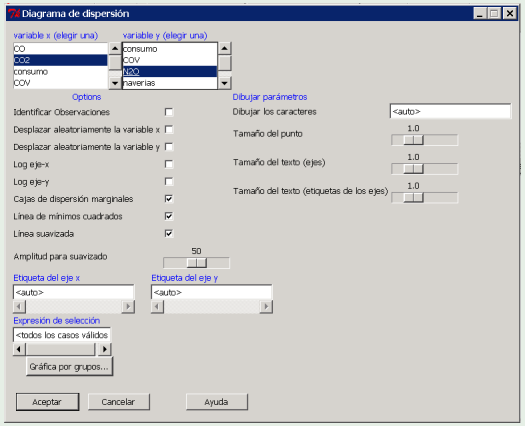
En la tercera fila, el `N2O` aparece en representado en el eje de ordenadas, y el resto de variables en el de abscisas. ¿Qué diagrama de dispersión de la tercera fila muestra un patrón más claro de relación? Si bien usualmente no se puede responder de forma concluyente a esta pregunta a través de estos gráficos, se ve claramente en este caso cómo no parece haber relación con `SO2`, no hay una relación lineal clara con `NOx` y la relación lineal es fuerte con `CO` y `CO2`.

Dibuje el diagrama de dispersión con la emisión de óxido nitroso (N2O) en el eje de ordenadas y la emisión de dióxido de carbono (CO2) en el eje de abcisas.

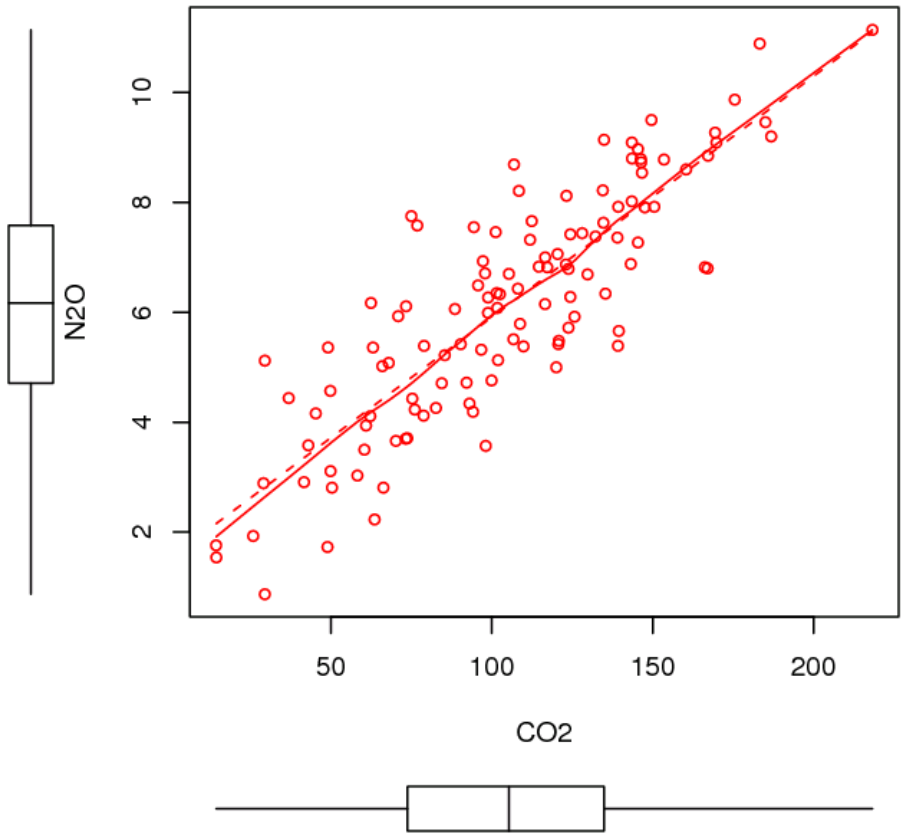
Gráficas
➥Diagrama de dispersión...



Seleccionamos: **CO2** y **N2O**
En Opciones, marcar **Cajas de dispersión marginales**, **Línea de mínimos cuadrados** y **Líneas suavizadas**
➥Aceptar



(Sigue →)



El eje de abscisas representa la emisión de `CO2` (con un gráfico de cajas) y el de ordenadas la de `N2O` (con un gráfico de cajas).
Se observa una relación creciente entre ambas magnitudes.
En el gráfico aparecen dos líneas. Una es la recta de regresión (el modelo más simple) y la otra la línea de regresión no paramétrica (el mejor ajuste posible a los datos, respecto de mínimos cuadrados). Si ambas líneas son muy similares, el ajuste lineal resulta adecuado.
En este caso la línea recta sigue muy bien el comportamiento de la línea no paramétrica, por lo que el modelo lineal parece ajustar bien los datos.

Estime la emisión de `N2O` a partir de la emisión de `CO2`. Llame al modelo como **RegModel.1**

Determinar los coeficientes β_0, β_1 y σ_ϵ tales que

$$\texttt{N2O} = \beta_0 + \beta_1 \cdot \texttt{CO2} + \epsilon, \quad \epsilon \sim \textit{Normal}(0, \sigma_\epsilon)$$

donde ϵ denota el error aleatorio.

Estadísticos

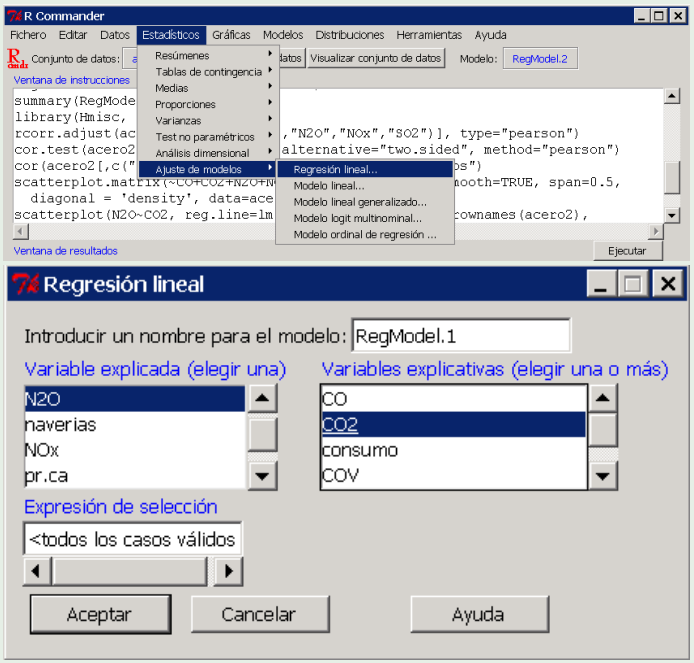
➥Ajuste de modelos

➥Regresión lineal

Variable explicada: **N2O**

➥Variables explicativas: **CO2**

➥Aceptar



(El modelo por defecto se llama `RegModel.1`)

(Sigue →)

```
> RegModel.1 <- lm(N2O~CO2, data=acero)
> summary(RegModel.1)
Call:
lm(formula = N2O ~ CO2, data = acero)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2585 -0.7287  0.0404  0.6511  2.9353

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.526865   0.280149    5.45 2.91e-07 ***
CO2         0.043850   0.002491   17.60  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.111 on 115 degrees of freedom
Multiple R-squared: 0.7293,        Adjusted R-squared: 0.7269
F-statistic: 309.8 on 1 and 115 DF,  p-value: < 2.2e-16
```

La columna `Estimate` proporciona los valores de las estimaciones de los coeficientes, con lo que el modelo de regresión lineal simple que mejor se ajusta a estos datos es:

$$\mathrm{N2O} = 1.526865_{\sigma=0.280149} + 0.043850_{\sigma=0.002491} \cdot \mathrm{CO2} + \epsilon, \quad \epsilon \sim \mathit{Normal}(0, 1.111)$$

Así pues, por cada unidad que se incremente la emisión de `CO2`, la emisión promedio de `N2O` se espera que se incremente en 0.043850 unidades [Compruebe los intervalos de confianza asociados: el intervalo de confianza al 95% para este efecto es $0.04385 \pm 1.96 \cdot 0.002491$].

Hemos obtenido que el coeficiente de determinación es: $R^2 = 72.69\%$, que estima el porcentaje de la variación de la variable dependiente que es explicado por la regresión. En este caso, el 72.69% de la variación de la emisión de `N2O` se debe a la emisión de `CO2`.

Pronósticos

Recordemos que los valores que proporciona la recta de regresión para un valor dado de la variable explicativa pueden interpretarse como:

- predicciones del valor de la variable explicada; o
- estimaciones de su media.

Tanto para estas predicciones como para estas estimaciones, podemos proporcionar intervalos de confianza al nivel que se considere apropiado, normalmente al 95%.

Ejemplo:
[interval="prediction"] ¿Cuál es la previsión de gasto medio en alcohol de Juan, que tiene 20 años?
[interval="confidence"] ¿Cuál es la previsión de gasto medio en alcohol de quienes tienen 20 años?

Utilice el modelo `RegModel.1` para estudiar los valores de emisión de `N2O` en las horas en las que se emiten 110t/h de `CO2` (estimación puntual).

Escriba en la `Ventana de instrucciones`:

```
predict(RegModel.1,data.frame(CO2=c(110)))
```

Las salidas de este procedimiento nos indican que la estimación puntual, obtenida a partir de este modelo de regresión lineal simple, de la emisión de `N2O` que se producirá en una hora en la que se hayan emitido 110t de `CO2` es de $6'350341$.

Utilice el modelo `RegModel.1` para estudiar los valores de emisión de `N2O` en las horas en las que se emiten 110t/h (toneladas/hora) de `CO2` (estimación por **intervalo**).

Escriba en la `Ventana de instrucciones`:

```
predict(RegModel.1,data.frame(CO2=c(110)),interval="prediction")
```

```
> predict(RegModel.1,data.frame(CO2=c(110)),interval="prediction")
       fit      lwr     upr
1 6.350341 4.140992 8.55969
```

El valor 6.350341 es la estimación puntual de la emisión de `N2O` que se obtiene a partir de este modelo. Es el valor que se obtiene en la recta de regresión para `CO2`=110.
Se tiene una confianza del 95% de que la emisión de `N2O` estará entre 4.140992t/h y 8.55969t/h para una hora en la que haya una emisión de 110t/h de `CO2`.

Utilice el modelo `RegModel.1` para estudiar los valores de emisión de `N2O` en las horas en las que se emiten 110t/h de `CO2` (estimación por intervalo con una confianza del 99%).

Escriba en la `Ventana de instrucciones`:

```
predict(RegModel.1,data.frame(CO2=c(110)),interval="prediction",level=0.99)
```

```
        fit      lwr      upr
1 6.350341 3.428878 9.271804
```

Utilice el modelo `RegModel.1` para estudiar los valores de emisión *media* de `N2O` en las horas en las que se emiten 110 y 100 t/h de `CO2` (estimación por intervalo para el **promedio**).

Escriba en la `Ventana de instrucciones`:

```
predict(RegModel.1,data.frame(CO2=c(110,100)),interval="confidence")
```

```
        fit       lwr       upr
1 6.350341 6.145248 6.555434
2 5.911843 5.707193 6.116494
```

Con una confianza del 95% la emisión *media* de `N2O` en aquellas horas en la que se emiten 110t/h de `CO2` está entre 6.145248 y 6.555434.
Y, respectivamente, la emisión media cuando la `CO2` es 100 está entre 5.707193 y 6.116494.