Regresión lineal con R Commander

Jose A. Egea, Mathieu Kessler
Departamento de Matemática Aplicada y Estadística
Universidad Politécnica de Cartagena
{josea.egea}{mathieu.kessler}@upct.es

1. Introducción

En este manual se describe paso a paso la manera de realizar ajustes de datos mediante regresión lineal usando R Commander, y se proponen una serie de ejercicios para ser resueltos por los alumnos.

El primer ejemplo se resolverá paso a paso y con él se repasarán los siguientes puntos:

- Detectar visualmente una posible relación entre pares de variables.
- Ajustar la recta de regresión de una variable dependiente dada una variable independiente.
- Realizar predicciones y estimaciones a partir de la recta de regresión.

2. Resolución de un problema paso a paso

Enunciado: En curso de mecanografía se evalúa la progresión de los alumnos registrando el número de pulsaciones por minuto (p.p.m.) empleado por cada uno al redactar un texto. Se ha evaluado a 8 estudiantes que llevan siguiendo el curso un diferente número de semanas. Los resultados se muestran en la Tabla 1.

Nº de semanas	3	5	2	8	6	9	3	4
p.p.m.	87	119	47	195	162	234	72	110

Cuadro 1: Velocidad de tecleado (p.p.m) vs. Nº de semanas de curso

1. Representa el diagrama de dispersión y calcula el coeficiente de correlación. ¿Es razonable suponer que existe una relación lineal entre el número de semanas y la ganancia de velocidad?

2. Calcula la recta de regresión.

3. ¿Qué velocidad de tecleado podemos esperar de una persona que hace 7 semanas que va a clase?

2.1. Introducción de los datos

En este apartado vamos a recordar como introducir datos de forma manual en R Commander. Otra posibilidad sería cargar datos desde un archivo, lo cual se ha visto anteriormente en este curso.

Comenzaremos abriendo R y cargando R Commander como ya sabemos hacer. El primer paso para la resolución de este ejercicio es introducir los datos como un nuevo conjunto de datos. Desde la ventana principal de R Commander, en el menú superior elegimos `Datos` → `Nuevo conjunto de datos`, tal y como se muestra en la Figura 1.

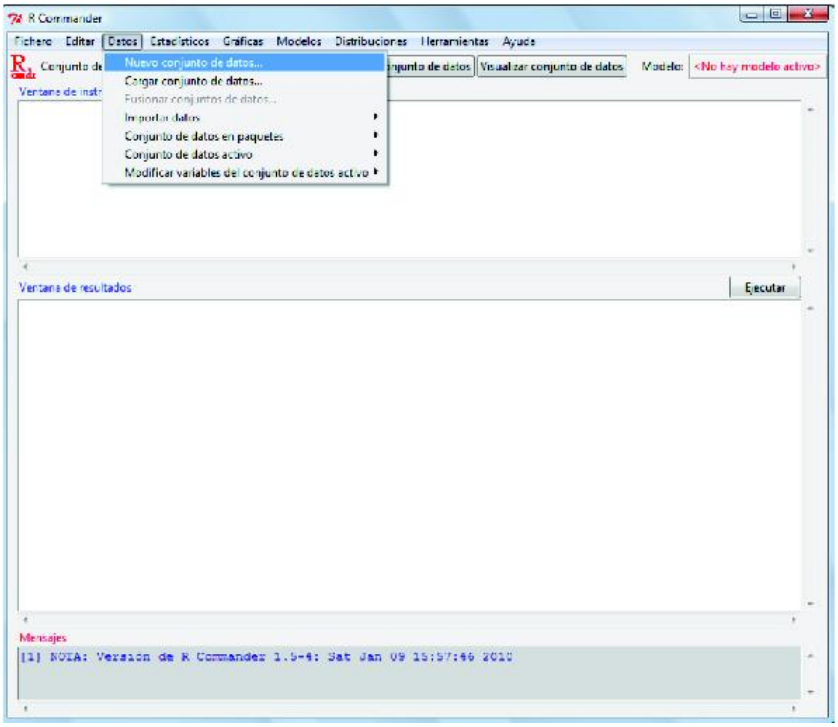
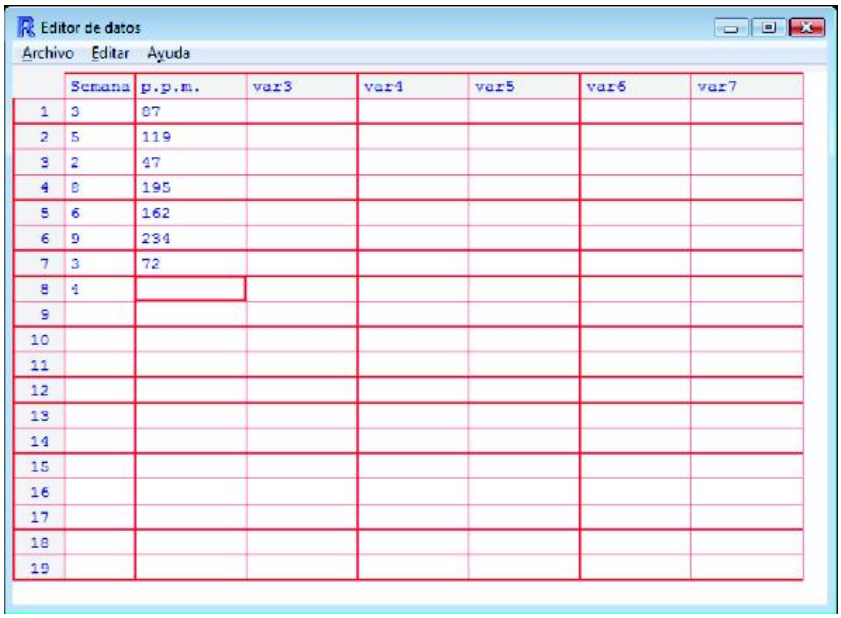


Figura 1: Menú para introducir datos

A continuación aparecerá una pantalla en la que se nos pedirá que demos un nombre al nuevo conjunto de datos. En este caso podemos llamarlo "Mecanografía", y después haremos click en `Aceptar`. Lo siguiente que aparece es una ventana, el editor de datos, con aspecto similar al de una hoja de cálculo, con diferentes celdas y con los títulos `var1`, `var2`, `var3`, etc. encabezando cada columna. Nosotros vamos a introducir nuestros datos en columnas. Podemos cambiar el nombre de cada columna (o variable) simplemente haciendo click sobre las palabras `var1`, `var2`, etc. De esta forma nos aparecerá una

nueva ventana en la que podemos elegir el nombre de la variable y el tipo. Para nuestro ejemplo, elegiremos los nombres "Semana" y "p.p.m.", siendo ambas variables del tipo `numeric`. Para guardas los cambios basta con salir de la pantalla haciendo click en la esquina superior derecha.

Tras realizar esas operaciones e introducir los datos, la ventana del editor de datos debería tener un aspecto similar al de la Figura 2.



	Semana	p.p.m.	var3	var4	var5	var6	var7
1	3	87					
2	5	119					
3	2	47					
4	8	195					
5	6	162					
6	9	234					
7	3	72					
8	4						
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							

Figura 2: Editor de datos para nuestro problema

Una vez introducidos los datos cerramos el editor de datos, volviendo así a la pantalla principal de R Commander. Observaremos que ahora, en la parte superior izquierda justo debajo del menú aparece: `Conjunto de datos:` Mecanografía. Si por algún motivo nos hemos equivocado al introducir los datos o deseamos realizar alguna modificación, pulsaremos en el botón `Editar conjunto de datos` que se encuentra a la derecha del nombre de nuestro conjunto de datos.

2.2. Diagrama de dispersión y coeficiente de correlación

Una vez hemos introducido nuestros datos, vamos a intentar responder a la primera cuestión del ejercicio. En primer lugar presentaremos el diagrama de dispersión para tener una primera impresión visual sobre la posible relación entre los datos. Desde la pantalla principal de R Commander seleccionamos `Gráficas` → `Diagrama de dispersión...` Una nueva pantalla con muchas opciones aparecerá ante nosotros (ver Figura 3).

De esas opciones sólo nos interesan unas cuantas de momento. Para empezar tenemos que escoger qué variable es la independiente (eje X) y cuál es la dependiente (eje Y). Para ello simplemente hacemos click sobre los nombres de las variables y automáticamente

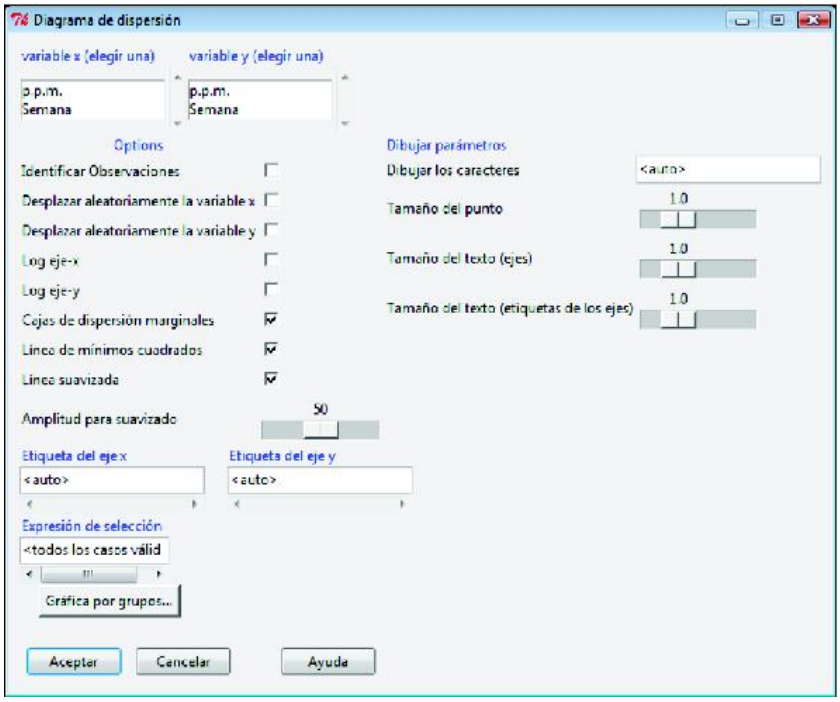


Figura 3: Opciones para la gráfica de dispersión

quedarán seleccionadas. En nuestro ejemplo la variable independiente es "Semana" y la variable dependiente es "p.p.m.".

Vemos que podemos obtener automáticamente la recta de regresión en la gráfica. De hecho, la opción `Línea de mínimos cuadrados` aparece seleccionada por defecto. En principio a nosotros sólo nos interesa el diagrama de dispersión, así que vamos a de-seleccionar todas las opciones seleccionadas por defecto, y pulsaremos en `Aceptar`. Aparecerá una nueva ventana con un diagrama de dispersión similar al que se muestra en la Figura 4[1]

El diagrama de dispersión nos muestra que la relación entre las dos variables puede ser considerada como lineal con pendiente positiva (o sea, que a mayor número de semanas de curso, mayor velocidad de tecleado, como nos dice el sentido común). Vamos a ver cuál es el coeficiente de correlación. En el menú de R Commander elegimos la opción `Estadísticos` → `Resúmenes` → `Matriz de correlaciones...`. En la ventana emergente debemos seleccionar ambas variables (Semana y p.p.m.) pinchando y arrastrando hacia abajo con el ratón o seleccionando ambas variables mientras mantenemos pulsada la tecla Ctrl. Seleccionamos `Coeficiente de Pearson` (está seleccionado por defecto) y pulsamos `Aceptar`.

[1]Observemos que en la ventana de instrucciones de R Commander ha aparecido lo siguiente: `scatterplot(p.p.m.~Semana, reg.line=FALSE, smooth=FALSE, labels=FALSE, boxplots=FALSE, span=0.5, data=Mecanografía)`. Esto nos puede dar una idea de los comandos de R para realizar este tipo de operaciones.

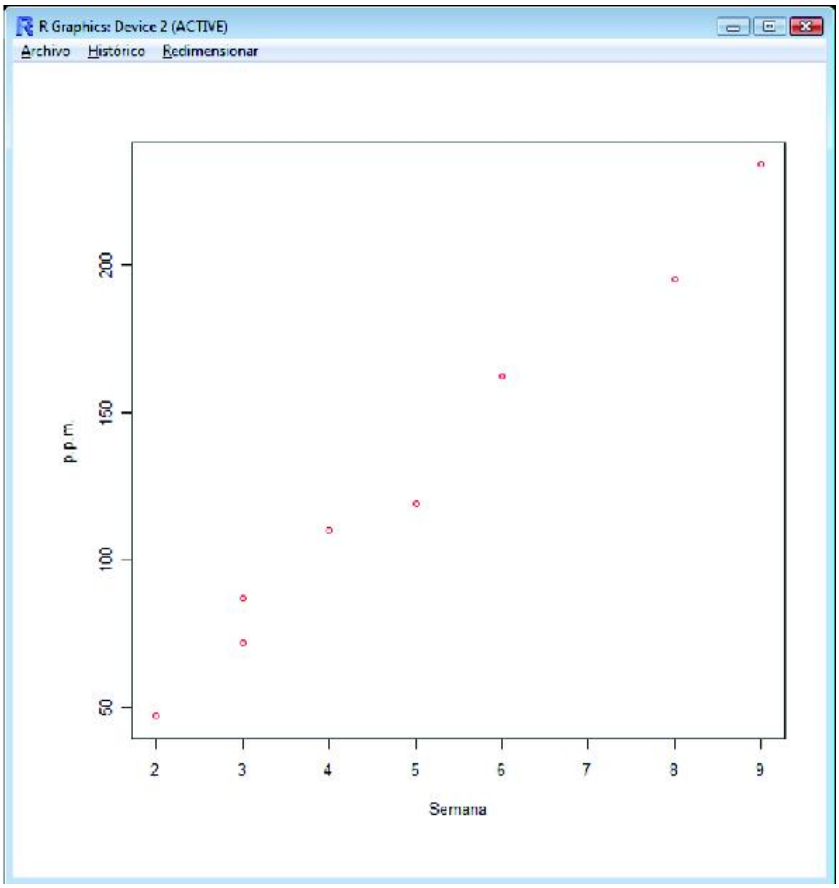


Figura 4: Gráfica de dispersión

En la ventana de resultados de R Commander habrá aparecido un comando de R y unos resultados numéricos similares a los presentados en la Tabla 2.

	p.p.m.	Semana
p.p.m	1.0000000	0.9919602
Semana	0.9919602	1.0000000

Cuadro 2: Matriz de correlación

En esa tabla observamos:

- La diagonal está formada por unos, ya que el coeficiente de correlación lineal de una variable consigo misma es uno.
- La matriz es simétrica, ya que el coeficiente de correlación lineal también lo es, es decir, el coeficiente de la variable X con la variable Y es idéntico al de la variable Y con la variable X.
- El coeficiente de correlación lineal entre ambas variables es positivo y muy cercano a uno, lo cual confirma la información proporcionada por el diagrama de dispersión,

donde observábamos una clara relación lineal entre ambas variables con relación positiva.

A la pregunta de si es razonable suponer que existe una relación lineal entre el número de semanas y la ganancia de velocidad, responderemos SÍ.

2.3. Cálculo de la recta de regresión

Para calcular la recta de regresión entre nuestras variables, seleccionaremos `Estadísticos` → `Ajuste de modelos` → `Regresión lineal...`, y elegiremos `p.p.m.` como "variable explicada" (variable dependiente) y `Semana` como "variable explicativa" (variable independiente). A continuación pulsamos `Aceptar` y vemos que aparecerá el siguiente texto en la ventana de resultados de R Commander.

```
> RegModel.1 <- lm(p.p.m.~Semana, data=Mecanografía)
> summary(RegModel.1)
Call:
lm(formula = p.p.m. ~ Semana, data = Mecanografía)

Residuals:
     Min       1Q   Median       3Q      Max
 -9.2500  -6.5114  -0.4091   7.4091   9.3864

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
Intercept     1.659       7.282     0.228     0.827
Semanas      25.318       1.319    19.200  1.29e-06  ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.747 on 6 degrees of freedom
Multiple R-squared: 0.984, Adjusted R-squared: 0.9813
F-statistic: 368.6 on 1 and 6 DF, p-value: 1.291e-06
```

Se han destacado los valores correspondientes a la ordenada en el origen (1.659) y a la pendiente (25.318) de la recta de regresión, que sería por tanto de la forma:

$$\text{Velocidad de tecleado (p.p.m.)} = 1{,}659 + 25{,}318 \cdot \text{N}^{\text{o}} \text{ de semanas de curso}$$

También se muestra el coeficiente de determinación $R^2 = 0{,}984$, que confirma la relación lineal entre ambas variables. Para mostrar la recta de regresión repetiríamos los pasos realizados para mostrar el diagrama de dispersión, y en la ventana de opciones marcaríamos `Línea de mínimos cuadrados`. El resultado sería el de la Figura 5.

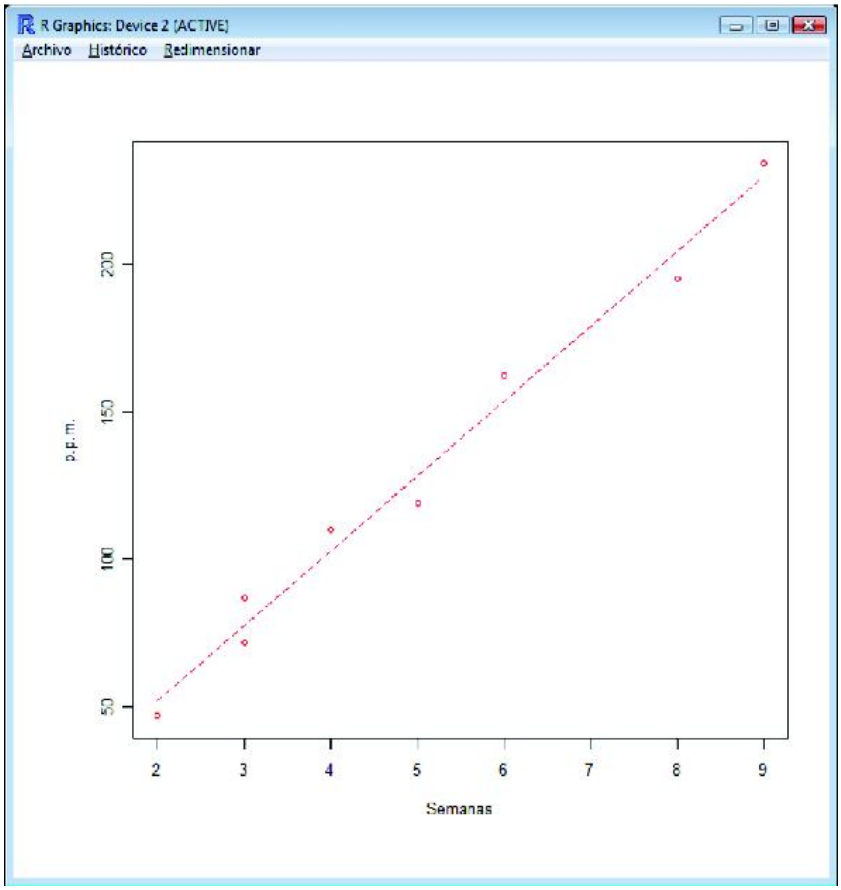


Figura 5: Recta de regresión

La interpretación de los parámetros obtenidos sería la siguiente: La ordenada en el origen no tiene ninguna interpretación con sentido, ya que correspondería a la velocidad media de tecleado de los alumnos antes de comenzar el curso. Sin embargo, un valor tan bajo (menos de 2 pulsaciones por minuto) no tiene ningún sentido en la realidad. La pendiente de la recta sí que nos da una información útil: por cada semana de clase se tiene una ganancia de velocidad de aproximadamente 25 p.p.m. En cuanto al coeficiente de determinación, nos confirma el excelente ajuste a un modelo lineal de estas dos variables.

2.4. Realizando predicciones

La última cuestión de nuestro ejemplo es calcular la velocidad de tecleado de un alumno que lleve 7 semanas de curso. Esto lo podemos hacer de 3 formas diferentes:

1. Directamente en la ventana de instrucciones de R Commander introduciendo manualmente los valores de la ecuación; es decir, tecleando `1.659 + 25.318*7`, seleccionando esa expresión y presionando `Ejecutar`. El resultado aparecerá en la ventana de resultados:

```
> 1.659 + 25.318*7
[1] 178.885
```

2. Usando el comando `predict` en la ventana de commandos de R Commander. Con nuestro conjunto de datos activo y cargando nuestro modelo lineal como se explicó anteriormente, escribimos `predict(RegModel.1,data.frame(Semana=7))` (en caso de que nuestro modelo se llame `RegModel.1`), lo seleccionamos y pulsamos `Ejecutar`. El resultado aparecerá en la ventana de resultados.

3. Cargando el paquete RcmdrPlugin.HH desde la ventana princial de R. Al cargar este paquete puede dar la sensación de que todos nuestros datos han sido borrados de R Commander, pero si vamos a `Modelos` → `Seleccionar modelo activo` podremos elegir de nuevo el modelo calculado anteriormente. Tras cargar este nuevo paquete, habrá aparecido una nueva opción bajo el menú `Modelos` llamado `Prediction Intervals... HH`. Seleccionamos esta opción y aparecerá una nueva ventana como la de la Figura 6. Bajo la variable "Semana" escribiremos "7" y seleccionaremos la opción `point estimate only`, pulsando después `Aceptar`. El resultado aparecerá en la ventana de resultados de R Commander.

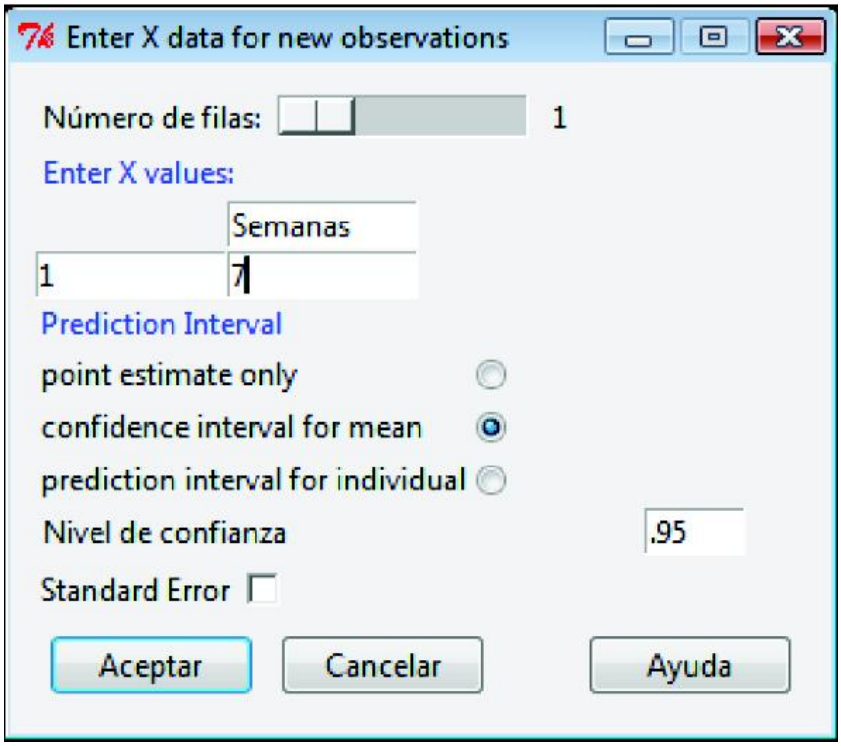


Figura 6: Ventana para hacer predicciones

A la pregunta inicial de qué velocidad de tecleado podemos esperar de un alumno que lleve 7 semanas de curso, responderemos que unas 179 p.p.m.

3. Ejemplo de ajuste exponencial

Consideramos el siguiente problema:

Hidrólisis del éster La hidrólisis de un cierto éster tiene lugar en medio ácido según un proceso cinético de primer orden. Partiendo de una concentración inicial desconocida del éster, se han medido las concentraciones del mismo a diferentes tiempos, obteniéndose los siguientes resultados (fichero `ester.txt`:

T (mn)	3	4	10	15	20	30	40	50	60	75	90
conc 10^{-3}(M)	25.5	23.4	18.2	14.2	11	6.7	4.1	2.5	1.5	0.7	0.4

3.1. Introducimos los datos y representamos la nube de puntos asociada

Vamos a crear un conjunto de datos (data.frame) llamado ester que contenga los datos del problema. Podemos introducir los datos manualmente en dos columnas que llamaremos `t` y `conc` o podemos importarlos del fichero ester.txt, incluido en el fichero datos_para_importar.zip (descargable en el Aula Virtual).
Una vez los datos introducidos, empezamos por realizar una nube de puntos de la variable `conc` respecto al tiempo `t`, usando el comando `Gráficas >Diagrama de dispersión.`

Claramente la relación entre estas dos variables no es lineal sino que presenta una forma exponencial decreciente.

3.2. Añadimos una variable `logconc` al conjunto de datos y representamos la nube asociada

En la ventana de instrucciones de R-Commander, añadimos una columna al conjunto ester:

```
ester$logconc=log(ester$conc)
```

Podemos ahora representar el diagrama de dispersión de `logconc` respecto a `t`. Estas variables presentan claramente una relación lineal, lo que confirma que la relación entre `conc` y `t` es exponencial.

3.3. Obtenemos la recta de regresión de `logconc` respecto a `t`

Podemos ahora llevar a cabo una regresión lineal de `logconc` respecto a la variable explicativa t tal como está explicado en el párrafo 2.3. Obtenemos la recta

```
logconc=3.3653-0.0486 t
```

3.4. Volvemos al modelo inicial

Sólo nos queda deshacer la transformación logarítmica para obtener el modelo inicial: tenemos

$$exp(logconc) = conc = exp(3{,}3653 - 0{,}0486t)$$

Calculando $exp(3{,}3653) = 28{,}912$ con R, deducimos que nuestro modelo ajustado es:

$$C = 28{,}912exp^{-0{,}0486t}.$$

4. Ejercicios propuestos

Tras la explicación paso a paso de cómo llevar a cabo una regresión lineal con R Commander, se proponen a continuación una serie de ejercicios para que los alumnos practiquen con ellos.

Problema 1

Los datos de la Tabla 3 son el resultado de un estudio del efecto de la temperatura de cristalización primaria (en °C) de una solución sobre el contenido en fósforo (g/L):

T (°C)	-6	-3	0	3	6	9	12	15	20	25
Concentración (g/L)	2.0	2.8	3.9	4.2	5.8	6.2	7.5	8.2	9.3	10.9

Cuadro 3: Contenido en P vs. temperatura de cristalización

Representa la gráfica de dispersión y ajusta los datos a un modelo lineal. Escribe la ecuación de la recta de regresión resultante del ajuste así como el coeficiente de determinación R^2. ¿Podemos inferir de los resultados cuál sería el contenido en fósforo para una temperatura de cristalización de 35°C?

Problema 2

Ajusta los datos de la Tabla 4 mediante regresión lineal por una función lineal del tipo $y = ax + b$, por una función potencial $y = ax^b$, y por una exponencial $y = ae^{bx}$.

X	1	2	4	8	16	32	64
Y	2	4	7	11	16	19	21

Cuadro 4: Datos del Problema 2

Para resolver este problema realiza la conversión de las variable mediante R Commander. Por ejemplo, para un ajuste potencial (y considerando que el conjunto de datos se llame "prob2" y las variables se llamen "x", "y"), podemos definir nuevas variables usando la ventana de instrucciones de R Commander de la siguiente manera: `prob2$logy=log(prob2$y)`, `prob2$logx=log(prob2$x)`. Estas nuevas variables "logx",

"logy" aparecerán automáticamente calculadas en nuestro conjunto de datos, y así podremos hacer la regresión respecto a ellas.

Presenta las gráficas de dispersión con la recta de regresión, la ecuación obtenida y el coeficiente de determinación en cada caso. ¿Qué modelo de curva es el más adecuado?

Problema 3 En el archivo "anscombe.txt" se encuentran los datos de Anscombe (1973), "Graphs in statistical analysis", *American Statistician*, **27**, pp 17-21. Cada par de columnas representa un par de variables x_i, y_i (así, las columnas 1 y 2 serían las variables x_1, y_1 respectivamente; las columnas 3 y 4 serían las variables x_2, y_2,y así sucesivamente). Calcula los coefientes de correlación y las rectas de regresión, (mostrando sus correspondientes gráficos de dispersión y las líneas de regresión) para los cuatro pares de variables. ¿Podemos explicar alguno de esos conjuntos de datos mediante un modelo lineal?

Problema 4 En el área de trabajo datos_primera_sesión.Rdata, se encuentra el conjunto de datos `cemento` que contiene las columnas `días` y `resistencia` (ver más información en la práctica "Manejo básico de la línea de comandos de R").

1. Representar la nube de puntos de la resistencia respecto a los días.
2. Añadir dos nuevos columnas al conjunto `cemento`, la primera llamada `logresist` será igual al logaritmo de `resistencia` y la segunda llamada `invt` será igual a `1/t`.
3. Representar la nube de puntos de `logresist` respecto a `invt`.
4. Ajustar una recta de regresión de `logresist` respecto a `invt`.
5. Volver al modelo inicial y obtener un ajuste de la forma $resistencia = Rexp^{-k/dias}$. ¿Qué interpretación física tiene el coeficiente R

Problema 5

En el área de trabajo datos_primera_sesión.Rdata, se encuentra el conjunto de datos `cemento` que contiene las columnas `duración` y `intervalo` (ver más información en la práctica "Manejo básico de la línea de comandos de R"). *Un geyser es un nacimiento de agua hirviente que de vez en cuando se vuelve inestable y expulsa agua y vapor. El geyser .*^{o}*ld Faithful.*^{en} *el parque de Yellowstone en Wyoming es probablemente el más famoso del mundo. Los visitantes del parque se acercan al emplazamiento del geyser intentando no tener que esperar demasiado para verlo estallar. Los servicios del Parque colocan un cartel donde se anuncia la próxima erupción. Es por lo tanto de interés estudiar los intervalos de tiempo entre dos erupciones conjuntamente con la duración de cada erupción. En el fichero* `geyser.txt` *están los datos correspondientes a la duración de 222 erupciones asi como el intervalo de tiempo hasta la siguiente erupción, durante los meses de agosto 1978 y agosto 1979. Las unidades de medición son mn.*

1. ¿Cuál sería la variable respuesta y cuál la variable explicativa?

2. Representa la gráfica de dispersión.

3. Ajusta los datos a un modelo lineal. Escribe la ecuación de la recta de regresión resultante del ajuste así como el coeficiente de determinación R^2.

4. ¿Cuál sería el principal uso de este modelo?