Estadística Básica con R y R–Commander

(Versión Febrero 2008)

Autores:
A. J. Arriaza Gómez
F. Fernández Palacín
M. A. López Sánchez
M. Muñoz Márquez
S. Pérez Plaza
A. Sánchez Navas

UCA Universidad de Cádiz
Servicio de Publicaciones

Edita: Servicio de Publicaciones de la Universidad de Cádiz
C/ Dr. Marañón, 3
11002 Cádiz

http://www.uca.es/publicaciones

ISBN:

Depósito legal:

Estadística Básica con R y R-commander
(Versión Febrero 2008)
Autores: A. J. Arriaza Gómez, F. Fernández Palacín, M. A. López Sánchez, M. Muñoz Márquez, S. Pérez Plaza, A. Sánchez Navas

http://knuth.uca.es/ebrcmdr

Capítulo 3

Análisis Exploratorio de Datos multidimensional

Una vez estudiados los distintos caracteres de la matriz de datos de forma individual, resulta muy interesante realizar análisis conjuntos de grupos de ellos, de hecho, la mayoría de los análisis estadísticos tienen carácter multivariable. Los motivos para adoptar este enfoque son variados, aunque de nuevo la cuestión de la naturaleza de los caracteres y los objetivos del estudio serán determinantes a la hora de fijar las técnicas que se emplearán.

Aunque en posteriores entregas se tratarán técnicas multivariables muy potentes, los objetivos en este capítulo son mucho más modestos y se limitarán a un primer acercamiento de naturaleza descriptiva; empleándose para ello tanto medidas de relación entre caracteres como representaciones gráficas. En la mayoría de las ocasiones sólo se contemplarán dos caracteres de forma conjunta, realizándose, por tanto, un análisis bidimensional.

En este capítulo también se hará una primera incursión en el tema de la modelización. Un modelo estadístico relaciona mediante una o varias expresiones matemáticas a un grupo de caracteres, que ocasionalmente deben cumplir algunos requisitos. En este caso, se abordará un *modelo de ajuste bidimensional*, en el que se tratará de explicar el comportamiento de una variable *causa* a partir de otra que se denomina

efecto.

Siempre existe un cierto grado de tolerancia para asimilar caracteres de menor nivel de información a los de nivel superior, aunque existe una marca que no se debe transgredir, que es la de la ordenación. Así, podría justificarse el tratar una variable contada como variable de escala, pero nunca se podría asimilar un atributo a una variable ordenada.

1. Tipos de relaciones entre caracteres

En principio se podrían establecer tantos tipos de relación como los que resultarían de cruzar los diferentes caracteres definidos en el capítulo anterior. No obstante, el número de cruces sería demasiado elevado y muchos de ellos no tendrían interés práctico, por lo que se limitará el estudio a aquellos que habitualmente se encuentran en la práctica, que básicamente se corresponden con los que relacionan caracteres de la misma naturaleza. Se expondrán previamente algunas matizaciones y precauciones que conviene tener presente.

- En general funcionan mejor los cruces entre caracteres de la misma naturaleza. Ello se debe a que para realizar el análisis se debe especificar algún tipo de *disimilaridad* que establezca la diferencia, en función de los caracteres considerados, que existe entre cada par de individuos de la matriz de datos. Así, la disimilaridad entre dos individuos sobre los que se han medido dos variables de escala es habitualmente la distancia euclídea, que como se sabe posee buenas propiedades, mientras que si un carácter es de clase y el otro una variable de escala la disimilaridad que se elija tendrá, con toda seguridad, propiedades mucho más débiles.

- Como consecuencia de lo anterior cuando se incluyan en el mismo análisis caracteres de distinta naturaleza conviene, siempre que sea posible, asignarles roles distintos.

- La asignación de roles a variables de la misma naturaleza en ningún caso se soportará por motivos estadísticos, sino que dependerá exclusivamente del criterio del investigador.

A, B	B_1	$\cdots$	B_j	$\cdots$	B_s	
A_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1s}	$\mathbf{n_{1\cdot}}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{is}	$\mathbf{n_{i\cdot}}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	n_{r1}	$\cdots$	n_{rj}	$\cdots$	n_{rs}	$\mathbf{n_{r\cdot}}$
	$\mathbf{n_{\cdot 1}}$	$\cdots$	$\mathbf{n_{\cdot j}}$	$\cdots$	$\mathbf{n_{\cdot s}}$	$\mathbf{n}$

Tabla 3.1: Distribuciones conjuntas y marginales de (A, B)

- La investigación combinatoria, es decir aquella que considera todos los grupos posibles de variables, está fuertemente desaconsejada, aunque se trate, como es el caso, de un análisis de carácter exploratorio. La violación de este principio puede llevar a aceptar como válidas asociaciones meramente espúreas.

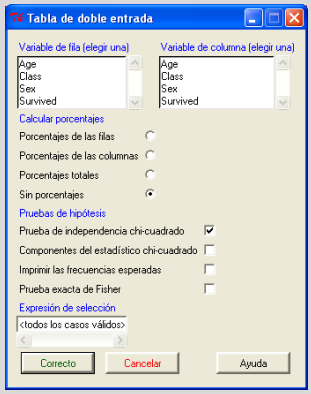
2. Análisis de relaciones entre dos atributos

Para relacionar dos atributos, tanto dicotómicos como politómicos, se construirá la tabla de frecuencias conjunta o *tabla de doble entrada*. Así, si se considera que el atributo A está conformado por las clases $A_1, A_2, \ldots, A_r$ y el atributo B por las clases $B_1, B_2, \ldots, B_s$, la información a tratar quedaría conformada por la tabla 3.1; donde n_{ij} representa la frecuencia absoluta del par (A_i, B_j), es decir el número de individuos que presentan de forma conjunta la clase A_i de A y la B_j de B. La última columna y la última fila de la tabla 3.1 representan las *distribuciones marginales* de A y B, respectivamente.

Cuando se consideran dos atributos dicotómicos se tendrá una tabla 2×2, que en ocasiones necesitará un tratamiento diferenciado. Mención aparte merece el caso en que uno o los dos atributos son del tipo *presencia-ausencia* de una cualidad.

Ejemplo 3.1

Como caso práctico para analizar la relación entre atributos se ha elegido el archivo de datos `titanic.dat`*, en el que aparecen las variables* `Class`*,* `Sex`*,* `Age` *y* `Survived`*, que aportan información, respectivamente, sobre la clase que ocupaba el pasajero, su sexo, edad y si sobrevivió o no al naufragio del famoso transatlántico. En concreto, se intentará establecer una posible asociación entre la supervivencia y la clase en la que viajaban los pasajeros del Titanic.*



En primer lugar se construirá la tabla de doble entrada con las variables seleccionadas. Con **Rcmdr** *esto se consigue desde* `Estadísticos`→`Tablas de contingencia`→`Tabla de doble entrada...`*, con lo que se abre la ventana de diálogo mostrada arriba, en la que se seleccionan los correspondientes atributos fila* `(Survived)` *y columna* `(Class)`*, además se eligen* `Porcentajes totales` *y se deja marcada la opción* `Prueba de independencia chi-cuadrado`*. Los resultados son:*

```
> .Table <-xtabs(~Survived+Class, data=Datos)

> .Table
Class

Survived    1st    2nd    3rd    Crew
No          122    167    528    673
Yes         203    118    178    212

> totPercents(. Table) # Percentage of Total

      1st     2nd     3rd     Crew     Total
No     5.5     7.6    24.0     30.6     67.7
Yes    9.2     5.4     8.1      9.6     32.3
Total 14.8    12.9    32.1     40.2     100.0

> .Test <- chisq.test(.Table, correct=FALSE)
> .Test
Pearson's Chi-squared test
data: .Table
X-squared=190.4011 ,df=3, p-value < 2.2e-16
```

R *además de proporcionar las tablas de valores absolutos y de porcentajes sobre el total, da información sobre el grado de relación entre los atributos, a través del coeficiente* χ^2*. De momento se considera sólo el valor del estadístico* $\chi^2 = 190{,}4$*. Este estadístico indica el grado de relación entre la clase que ocupaba el pasajero y si sobrevivió o no al naufragio; si* $\chi^2 = 0$ *indicaría una ausencia de relación y a medida que* χ^2 *crece la relación va en aumento.*

El estadístico no está acotado en un rango de valores que permita interpretar la intensidad de la relación, por lo que se debe recurrir a algún coeficiente derivado que esté acotado. Los más usuales son el coeficiente de contingencia y el coeficiente de Cramer, ambos acotados en el intervalo $[0, 1)$*. Se empleará en este caso el primero que viene dado por:*

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

donde n *es el tamaño muestral. En nuestro caso el coeficiente de contingencia vale* $0{,}28$*, lo que indica una cierta relación entre ambos atributos. Si se observa la tabla de doble entrada se ve que porcentualmente se salvaron más pasajeros de primera clase, mientras que los de tercera clase y la tripulación fueron los que más sufrieron las consecuencias del naufragio. Más adelante, se verá que se puede ser más contundente a la hora de concluir la existencia de relación utilizando los Contrastes de Hipótesis.*

Para poder visualizar la relación entre las variables puede ser muy útil la realización de un diagrama de barras de la variable supervivencia según la clase de los pasajeros. Para ello, se almacena en primer lugar la tabla de contingencia de las variables `Survived` *frente a* `Class`*, a la que se ha llamado* `Tabla`*, ejecutando en la ventana de instrucciones:*

```
>Tabla <-xtabs(~ Survived+Class, data=Datos)
```

A continuación se obtiene el diagrama de barras mediante las órdenes **R***:*

```
>barplot(Tabla, xlab=''Clase'', ylab=''Frecuencia'',
legend.text=c(''No superviviente'', ''Superviviente''),
beside=TRUE,col=cm.colors(2))
```

Observando el diagrama de barras de valores absolutos (figura 3.1), se aprecia que éste ofrece una visión que podría llevar a confusión, aparentando, por ejemplo, que el número de supervivientes de primera clase

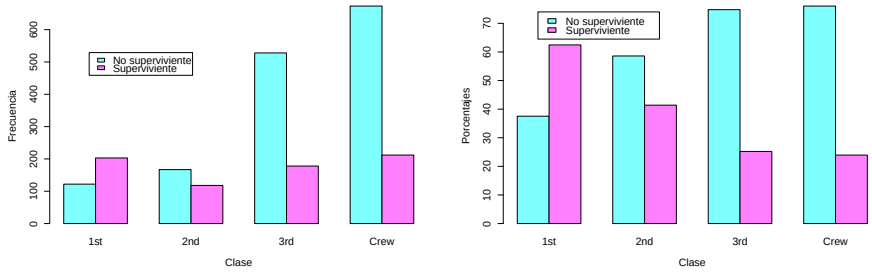


Figura 3.1: Diagramas de barras de la supervivencia

es prácticamente igual al número de supervientes de la tripulación. Ello se debe a que se han comparado las frecuencias absolutas de estos dos grupos, y mientras que en primera clase viajaban 325 individuos, los miembros de la tripulación eran 885. Una alternativa para apreciar la relación existente entre los dos atributos es construir el diagrama de barras de las frecuencias relativas, o porcentajes de supervivencia respecto a cada clase, en lugar de usar las frecuencias absolutas. Igual que antes, se debe almacenar previamente la tabla de porcentajes, lo que se consigue con las siguientes instrucciones **R***:*

```
>Tabaux <-colPercents(Tabla)
>Tablarel <-Tabaux[1:2][1:4]
```

`Tabaux` *contiene la tabla de porcentajes, los porcentajes totales y las frecuencias marginales. Para representar el diagrama de barras no son necesarias las dos últimas filas, por lo que se ha construido una nueva tabla denominada* `Tablarel` *con la información que interesa.*

Ahora se está en condiciones de construir el diagrama de barras; para ello se sustituye, en la secuencia de instrucciones usada para el diagrama de barras de valores absolutos, `Tabla` *por* `Tablarel` *(figura 3.1).*

Por último, se construirá un gráfico de mosaico, figura 3.2, con todos los atributos del fichero `Titanic`*. Para ello, se ejecuta la instrucción:*

```
>mosaicplot(Titanic, main=''Supervivientes del Titanic'',
color=c(''red'',''green''))
```

Se han seleccionado los colores verde para los supervivientes y rojo para los no supervivientes.

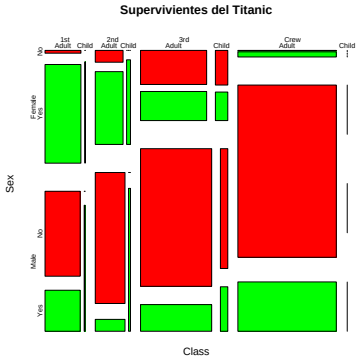


Figura 3.2: Gráfico de mosaico de los datos Titanic

R–Nota 3.1

Éste puede ser un buen momento para analizar someramente la sintaxis de las instrucciones **R**, *dado que en ocasiones, como ha ocurrido en este ejemplo, se necesita crear o editar una instrucción. Como el lector habrá podido comprobar, cada vez que se ha utilizado un procedimiento de* **Rcmdr**, *éste ha generado una o varias instrucciones* **R**; *en realidad,* **Rcmdr** *no es otra cosa que lo que se conoce como un* frontend *de* **R**, *es decir un forma más amigable de acceder a los recursos de* **R**.

Las instrucciones de **R** *pueden ser una expresión o una asignación. Una expresión se evalúa, se muestra su resultado y se descarta. Una asignación se evalúa obteniendo un nuevo objeto que se almacena con el nombre especificado.*

Concretamente, si se analiza la estructura de la instrucción:

```
>Tabla <-xtabs(~ Survived+Class, data=Datos)
```

se observa que se crea el objeto `Tabla`, *al que se le asigna* $(<-)$ *el resultado de la evaluación de la función* `xtabs`, *que genera una tabla de doble entrada con las variables* `Survived` *y* `Class` *del* `data.frame` *con nombre* `Datos`. *Si ahora se fija la atención en la instrucción:*

```
>barplot(Tabla, xlab=‘‘Clase’’, ylab=‘‘Frecuencia’’,
legend.text=c(‘‘No superviviente’’, ‘‘Superviviente’’),
beside=TRUE,col=cm.colors(2))
```

Ésta le indica a **R** *que cree un gráfico de barras,* `barplot`, *de la*

tabla de doble entrada `Tabla`*, siendo las etiquetas de los ejes,* `xlab e ylab, Clase y Frecuencia`*, que la leyenda de las clases,* `legend.text`*, sea* `No superviviente` *y* `Superviviente`*, que el tipo de barras sea pegada,* `beside=TRUE`*, y que utilice la gama de colores* `col=cm.colors(2)`*.*

R–Nota 3.2

En los diagramas de barras anteriores se usa el argumento `legend.text` *para incluir una leyenda de los datos, pero de esta forma la leyenda se dibuja en ocasiones sobre las barras. Para mejorar los resultados gráficos se pueden utilizar las siguientes instrucciones:*

1. *Escribir la orden del gráfico de barras sin* `legend.text`*:*

```
>barplot(Tablarel, xlab=‘‘Clase’’, ylab=‘‘Porcentajes’’,
beside=TRUE,col=cm.colors(2))
```

2. *Para localizar las coordenadas del gráfico en las que se desea insertar la leyenda se emplea la orden* `locator(n)`*, donde* `n` *es el número de puntos de los que se quiere averiguar las coordenadas, en nuestro caso* `n`$= 1$*.*

3. *Una vez ejecutada la orden, se pincha en la gráfica anterior con el botón izquierdo del ratón en el lugar donde se desee insertar la leyenda y automáticamente aparecerán las coordenadas* `(x,y)` *del punto elegido.*

4. *Por último, se incluirá la leyenda en la posición elegida con la orden:*

```
legend(x,y,c(‘‘No superviviente’’,‘‘Superviviente’’),
fill=cm.colors(2))
```

   El argumento `fill` *sirve para indicarle los colores de las barras.*

3. Análisis de relaciones entre dos variables

Una vez analizada la relación entre dos atributos, se aborda el estudio de la relación entre dos variables medidas. Este estudio se hará a través de la construcción de una *función de ajuste*, que expresa matemáticamente cómo una de las variables denominada *causa* explica el comportamiento de la otra variable llamada *efecto*. A la variable causa se le conoce también con los nombres de *independiente*, *explicativa*, *exógena*, ..., mientras que la variable efecto es llamada también *dependiente*, *explicada*, *endógena*, ... Desde el punto de vista de la investigación que se esté realizando es fundamental la selección de las variables que entrarán en el análisis y la asignación de roles, causa-efecto, para cada una de ellas.

Es muy habitual confundir los conceptos de *ajuste* y de *regresión*, y aunque no es objeto de este manual entrar en temas teóricos en profundidad, si habría que aclarar que la idea de ajuste implica la selección de un modelo matemático que aproxime lo mejor posible la relación entre las variables, mientras que el concepto de regresión hace referencia a la idea de predecir mediante alguna regla, un valor de la variable dependiente para cada valor de la independiente. Dicho lo cual, y como suele ocurrir en muchos textos estadísticos, a partir de ahora se admitirá, y usará, de forma indistinta ambos conceptos.

Por otra parte, en la mayoría de las ocasiones la matriz de datos contiene varias variables numéricas y el investigador desea estudiar cómo se explica el comportamiento de una de ellas sobre la que tiene un especial interés (dependiente) a partir del conocimiento de un conjunto del resto de variables (independientes). En esta situación, el análisis dos a dos, en el que se consideraría la variable dependiente con cada una de las independientes es claramente ineficiente, siendo necesario la construcción de un modelo de ajuste múltiple que relacione de forma conjunta la variable dependiente con el conjunto de las independientes. La explicación para plantear este enfoque es que las variables independientes suelen estar relacionadas también entre ellas, es decir comparten información de los individuos que se están estudiando, de forma que si se hiciera el análisis dos a dos se estaría utilizando la misma información

de forma reiterada.

En lo sucesivo, se consideran sólo dos variables, la independiente (X) y la dependiente (Y), dando lugar a n parejas de valores $(x_i,\ y_i)$. Desde un punto de vista gráfico estos valores se pueden representar en un plano, siendo el conjunto de puntos la denominada *nube de puntos* o *diagrama de dispersión*. El objeto del ajuste es la obtención de una función que se adapte lo mejor posible a la nube de puntos.

$$Y^* = f(X)$$

El conocimiento previo que se puede tener de la relación Y/X junto con el análisis de la nube de puntos debe ofrecer las claves para la selección de la función f. En realidad seleccionar f es elegir una clase funcional que dependerá de unos parámetros que habrá que estimar. Es decir, se elige una recta $Y = a+bX$, una parábola $Y = a+bX+cX^2$, una función exponencial $Y = ab^X$, una función potencial $Y = aX^b$, una hipérbola $Y = a + \frac{b}{X}$, ...Se puede apreciar que mediante alguna transformación muchas de estas funciones se convierten en rectas.

Ejemplo 3.2

- *La clase funcional exponencial $Y = ab^X$ aplicando una transformación logarítmica se linealiza, $logY = loga + Xlogb$.*
- *La clase funcional hiperbólica $Y = a + \frac{b}{X}$ también se convierte en una recta transformando $X' = \frac{1}{X}$.*

Cuando antes se ha escrito «la selección de un modelo matemático que aproxime lo "mejor posible" la relación entre las variables» o la «obtención de una curva que se adapte lo "mejor posible" a la nube de puntos», en realidad se estaba indicando la necesidad de establecer un criterio de ajuste que minimice las diferencias entre la curva de ajuste y la nube de puntos. El criterio más generalizado es el de los *mínimos cuadrados*, que establece que la suma de las distancias al cuadrado entre los valores observados de la variable Y, es decir los y_i, y las predicciones

que se obtienen de ésta a partir de la función de ajuste, $y_i^* = f(x_i) \forall i$, sea mínima. La aplicación de este criterio permite la estimación de los parámetros del modelo y la determinación de forma unívoca de la función de ajuste.

La figura 3.3 ilustra lo dicho para el caso lineal $Y = a + bX$, donde a representa el punto de corte de la recta con el eje Y y b el incremento–decremento de Y para un incremento unitario de X.

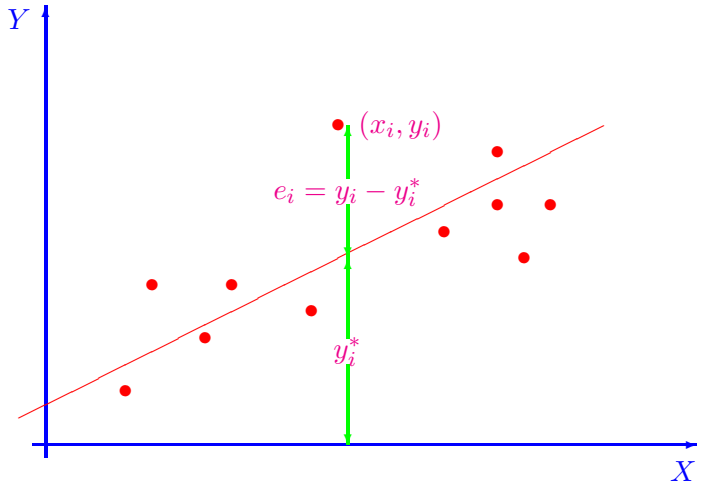


Figura 3.3: Recta de ajuste

- **Predicciones.** Una de las utilidades más importantes del ajuste es la de realizar predicciones de la variable explicada para distintos valores de la variable explicativa. En realidad, se trata de sustituir en el ajuste los valores de X para obtener los correspondientes valores de Y. Cuando se sustituyen los valores de X que se han empleado para calcular la función de ajuste, $x_1, x_2, \ldots, x_n$ se obtienen los correspondientes valores ajustados por el modelo, $y_1^*, y_2^*, \ldots, y_n^*$, mientras que si se asigna a X cualquier valor factible para esta variable, el valor que se obtiene para Y es una predicción. Obsérvese que la diferencia entre los valores observados de Y, y_i, y sus correspondientes valores ajustados, y_i^*, son los errores del ajuste $e_i = y_i - y_i^*$. Los puntos ajustados (x_i, y_i^*) pertenecen

a la recta de ajuste y los y_i^* tienen menos varianza que los y_i, de hecho, se puede demostrar para una gran cantidad de modelos, en particular para el lineal, que la varianza de Y es igual a la de Y^* más la varianza del error, $S_Y^2 = S_{Y^*}^2 + S_e^2$.

Las predicciones para valores de X distintos a los empleados en el ajuste se denominan interpolaciones cuando dichos valores se encuentran dentro del rango de valores de ajuste para X, y extrapolaciones cuando se encuentran fuera de dicho rango. La validez estadística de las interpolaciones es mayor que las de las extrapolaciones, de hecho la calidad de la predicción decrece cuando aumenta la distancia al centro de gravedad de la nube de puntos, $(\bar{x}, \bar{y})$.

- **Análisis de bondad del ajuste.** El ajuste no estaría totalmente resuelto si no viniera acompañado de una medida de su bondad, es decir, de un valor, a ser posible acotado en un intervalo, que exprese en qué porcentaje la variable dependiente se explica por la independiente a través del ajuste realizado. Si el ajuste fuera perfecto todos los valores observados se situarían sobre la nube de puntos y los residuos y su varianza se anularían, mientras que en el extremo contrario sería la variable ajustada la que tendría varianza nula.

  La medida que sintetiza lo expresado en el párrafo anterior es el *coeficiente de determinación*, $R^2 = \frac{S_{Y^*}}{S_Y^2}$ que, como puede verse, toma valores en $[0, 1]$; interpretándose que la variable Y se explica en un $100 * R^2\,\%$ por la variable X, mientras que el resto, es decir el $100 * (1 - R^2)\,\%$, se explicaría por una parte a través de una mejora de la función de ajuste, por otra incorporando, si es factible, información nueva (otras variables, con lo que se tendría un modelo de regresión múltiple) y por la variabilidad intrínseca de los datos.

  Para el caso de ajuste lineal existe un coeficiente específico de bondad de ajuste denominado *coeficiente de correlación lineal* r, que toma valores en el intervalo $[-1, 1]$ y que además de medir la intensidad de la relación indica si ésta es de tipo directo, cuando X crece Y crece, o inverso, cuando X crece Y decrece. Se verifica que $r^2 = R^2$.

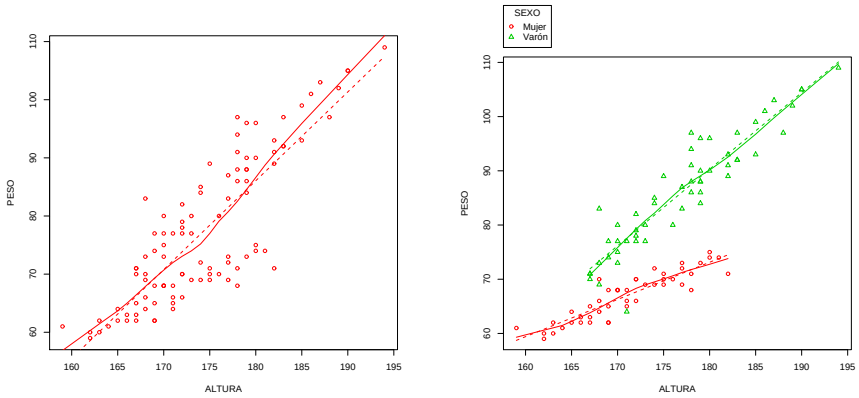


Figura 3.4: Diagramas de dispersión peso-altura

- **Análisis de residuos del modelo.** Conviene examinar, tanto desde un punto de vista numérico como sobre todo gráfico, los residuos que genera el ajuste, es decir las diferencias entre los valores observados, Y, y los ajustados por la función de ajuste, Y^*. En particular, resulta de especial interés el análisis de los residuos extremos y de las gráficas de los residuos frente a valores de X, indexados o frente a las predicciones. También es interesante el análisis de puntos influyentes, entendiendo esto como aquellos puntos que tienen un sobrepeso en la construcción de la función de ajuste. Estos puntos van a estar localizados en los extremos de la nube de puntos, ver ejemplo 3.3.

- **Mejora del modelo.** Para terminar, conviene indicar que reemplazar una función de ajuste por otra más sofisticada, con más parámetros y más compleja, sólo se justifica si la mejora en términos de R^2 es alta, pues en otro caso se complica la interpretación del modelo sin apenas recompensa.

Ejemplo 3.3

Para ilustrar los conceptos sobre el ajuste lineal se procederá a analizar la relación entre `peso` *y* `altura` *del fichero de datos* `peso_altura.dat`*, en*

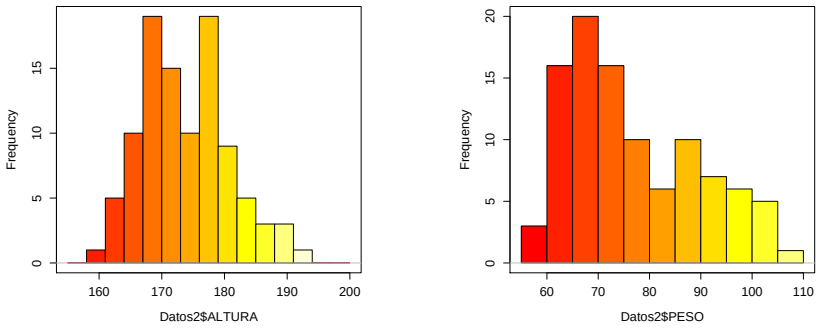


Figura 3.5: Histogramas de peso y altura

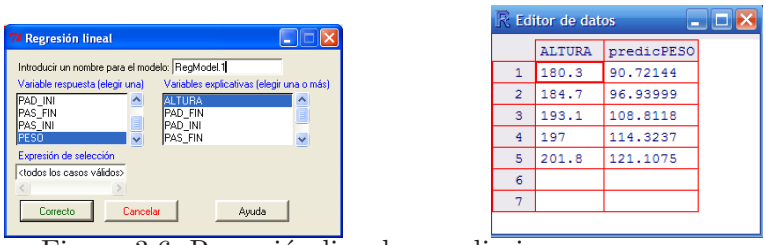


Figura 3.6: Regresión lineal y predicciones

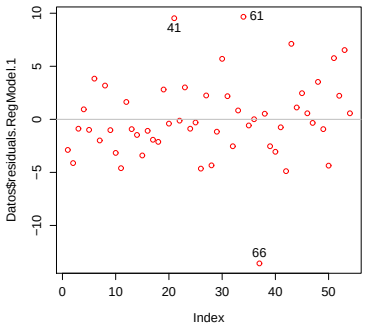


Figura 3.7: Residuos indexados

el que aparecen, entre otras variables, el sexo, peso y altura de un grupo de personas. Como se ha indicado anteriormente es necesario establecer qué variable será la explicada y cuál la explicativa. Dado que se trata de un ejemplo y que no se cuenta con elementos adicionales para avalar la decisión, se decide explicar el `peso` *en función de la* `altura`*.*

1. *Histogramas. Antes de abordar el análisis bidimensional propiamente dicho, se representarán los histogramas de las variables* `peso` *y* `altura`*, operando para ello tal y como se indicó en el capítulo anterior. Al objeto de fijar el número de clases de los histogramas y los colores, se retocan las instrucciones* **R** *que genera* **Rcmdr***, cambiando en ambos casos las opciones del número de intervalos (breaks) y los colores (col) y se vuelven a ejecutar, con lo que se obtiene las figuras en 3.5. Las instrucciones retocadas son respectivamente:*

```
>Hist(Datos$ALTURA, scale=‘‘frequency’’, breaks=seq(155,200,3),
col=heat.colors(13))
>Hist(Datos$PESO, scale=‘‘frequency’’, breaks=seq(55,110,5),
col=heat.colors(12))
```

   Una primera visión de los histogramas permite detectar una bimodalidad tanto en la variable `peso` *como en la* `altura`*, aunque ello es un indicio claro de mezcla de poblaciones, se continuará con los siguientes pasos del ajuste con todos los datos, en un ejercicio básicamente didáctico, en busca de establecer la relación que justifique el* `peso` *en función de la* `altura`*.*

2. *Diagrama de dispersión. Al objeto de decidir el tipo de función de ajuste que se utilizará, se representa el diagrama de dispersión. En* **Rcmdr** *se seleccionan las opciones* `Gráficas`→ `Diagrama de dispersión...`*, para las variables mencionadas. Por defecto aparece marcada la opción* `línea suavizada`*, que ofrece una regresión a los puntos y que da una idea de la clase funcional más eficiente bajo el criterio de mínimos cuadrados.*

   A la vista de la figura 3.4 se observa la existencia de relación entre las dos variables. La línea de regresión suavizada y la línea discontinua de ajuste lineal, sugieren que los ajustes más eficientes son tipo lineal y posiblemente parabólico o potencial. No obstante, la escala de representación de las variables podría ser un factor

distorsionador que podría llevar a pensar, erróneamente, que las variables mantienen un grado de relación lineal mayor del que realmente existe. Para confirmar la existencia de una alta correlación se calculará el coeficiente de correlación lineal de Pearson.

3. *Análisis de la correlación. Se selecciona la secuencia de opciones* `Estadísticos`→`Resúmenes`→`Test de correlación`, *eligiéndose en el cuadro de diálogo las variables que interesan. La salida que ofrece* **Rcmdr** *es:*

```
> cor.test(Datos$ALTURA, Datos$PESO, alternative=‘‘two.sided’’,
method=‘‘pearson’’)
    Pearson’s product-moment correlation
data: Datos$ALTURA and Datos$PESO
t = 15.8396, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.7818060 0.8952982
sample estimates:
cor
0.8480039
```

El coeficiente de correlación es positivo y relativamente alto, $r = 0,848$, *lo que indica que existe relación directa entre las variables. En cuanto a la intensidad, el coeficiente de determinación* $R^2 = r^2 = 0,719$ *implica que un* $28\,\%$ *de la variación de* Y *no se explica por* X *a través de la recta de ajuste.*

En este momento, y si no se hubiera detectado la bimodalidad en el histograma, habría que plantearse la posibilidad de mejorar la función de ajuste utilizando una clase funcional que se adaptara mejor a la nube de puntos; en el diagrama de dispersión se ha visto que la regresión suavizada sugería la posibilidad de un crecimiento de tipo parabólico o potencial. Pero como ya se ha comentado antes, la bimodalidad del histograma parece indicar la confusión de dos poblaciones. En efecto, se están considerando conjuntamente los dos sexos, hombre y mujer, cuando los patrones de relación peso–altura no tienen porqué coincidir y de hecho no lo hacen. Si se observa atentamente el diagrama de dispersión se puede entrever la existencia de dos poblaciones, para confirmarlo se representará el diagrama de dispersión pero diferenciando los individuos de ambos sexos.

4. *Análisis por grupo. En* **Rcmdr** *se eligen las opciones* `Gráficas→ Diagrama de dispersión...`, *seleccionando en la ventana de diálogo la opción* `Gráfica por grupos...` *la variable* `sexo`. *La visualización del gráfico 3.4 es muy elocuente, las dos líneas de ajuste se acomodan mucho mejor a sus respectivos grupos y la regresión suavizada, al contrario de lo que ocurría antes, no presenta desviaciones claras de la linealidad. Por lo que procede ajustar de forma diferenciada las variables* `peso-altura` *para cada sexo.*

   Para dividir el conjunto de datos según la variable `SEXO`, *se procede en* **Rcmdr** *desde* `Datos→Datos activos→ Filtrar los datos activos...` *tomando como expresión de selección* `SEXO==''Mujer''` *para la muestra femenina y* `SEXO==''Varón''` *para la masculina.* **R** *crea nuevos conjuntos de datos con los nombres que se le hayan indicado en el correspondiente apartado de la opción de filtrado. En este caso se han denominado* `Peso_Altura_Mujer` *y* `Peso_Altura_Varon`, *respectivamente.*

   Para analizar cada grupo de sexo, se elige como juego de datos activos el que interese y se calcula su coeficiente de correlación de Pearson. Se observa como la correlación para las mujeres es de $0,897$, *mientras que para los hombres llega hasta* $0,928$, *con* R^2 *iguales, respectivamente a* $0,804$ *y* $0,861$, *mucho más altas que las que se tenían para el ajuste conjunto.*

```
> cor.test(Peso_Altura_Mujer$ALTURA, Peso_Altura_Mujer$PESO,
alternative=''two.sided'', method=''pearson'')
    Pearson's product-moment correlation
data: Peso_Altura_Mujer$ALTURA and Peso_Altura_Mujer$PESO
t = 13.4879, df = 44, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.8208994 0.9422066
sample estimates:
cor
0.8973532
```

```
> cor.test(Peso_Altura_Varon$ALTURA, Peso_Altura_Varon$PESO,
alternative=‘‘two.sided’’, method=‘‘pearson’’)
    Pearson’s product-moment correlation
data: Peso_Altura_Varon$ALTURA and Peso_Altura_Varon$PESO
t = 13.0335, df = 52, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.8793910 0.9580797
sample estimates:
cor
0.9285171
```

5. *Recta de ajuste. Se obtendrá ahora una de las dos rectas de ajuste del peso en función de la altura, concretamente se ha elegido el subgrupo de los hombres. Una vez elegido el conjunto de datos activo correspondiente a los hombres, se selecciona* `Estadísticos→ Ajuste de modelos→Regresión lineal...`*, y en la ventana de la figura 3.6, se elige* `PESO` *como variable explicada y* `ALTURA` *como variable explicativa.*

```
> RegModel.1 <- lm(PESO~ ALTURA, data=Peso_Altura_Varon)
> summary(RegModel.1)
Call:
lm(formula = PESO ~ ALTURA, data = Peso_Altura_Varon)
Residuals:

Min        1Q        Median        3Q        Max
-13.578    -2.091    -0.491        2.213     9.662

Coefficients:

            Estimate        Std. Error     t value      Pr(> |t|)
(Intercept)    -164.09760      13.89222     -11.81     2.43e-16 ***
ALTURA          1.41331         0.07837      18.03     < 2e-16 ***

Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
Residual standard error: 3.937 on 52 degrees of freedom
Multiple R-Squared: 0.8621, Adjusted R-squared: 0.8595
F-statistic: 325.2 on 1 and 52 DF, p-value: < 2,2e − 16
```

A la vista de los resultados se sabe que la recta de regresión es `Y=-164,09760 +1,41331X`*. Si sólo se quisieran obtener los coeficientes de la recta éstos se pueden obtener con las órdenes:*

```
> RegModel.1 <- lm(PESO~ ALTURA, data=Peso_Altura_Varon)
> coef(RegModel.1)
(Intercept) ALTURA
-164.097600 1.413306
```

6. *Valores ajustados y predicciones. Para obtener los valores ajustados por el modelo se selecciona* `Modelos→ Añadir las estadísticas de las observaciones a los datos...` *y se marcan las opciones deseadas, en este caso* `Valores ajustados` *y* `residuos`. **R** *añade al conjunto de datos activos dos nuevas columnas llamadas* `fitted`*.RegModel.1 y* `residuals`*.RegModel.1 con los correspondientes valores ajustados y residuos del modelo activo.*

   Al realizar las estadísticas descriptivas de Y, Y^* y e, seleccionando las opciones media y desviación típica en resúmenes numéricos, se tiene:

```
> numSummary(Hombres[,c(''fitted.RegModel.1'', ''PESO'',
''residuals.RegModel.1'')], statistics=c(''mean'', ''sd''))


                       mean             sd           n
fitted.RegModel.1     8.624074e+01     9.753284     54
PESO                  8.624074e+01     10.504150    54
residuals.RegModel.1  -3.781456e-17    3.900081     54
```

   y efectivamente se comprueba que $S_Y^2 = S_{Y^*}^2 + S_e^2$, ya que $10,504^2 = 9,753^2 + 3,9^2$; pudiéndose calcular el coeficiente de determinación como $R^2 = \frac{9,753^2}{10,504^2} = 0,8621$.

   Para realizar predicciones para cualquier valor de X, se necesita crear previamente un nuevo conjunto de datos, que en este caso se ha llamado `pred` *y que contendrá una variable cuyo nombre se hace coincidir con el nombre de la variable independiente del modelo:*

```
>pred<-data.frame(ALTURA=c(180.3,184.7,193.1,197.0,201.8))
```

   Se incluyen en el fichero `pred` *los valores 180.3, 184.7, 193.1, 197.0 y 201.8 cms. Seguidamente se asigna a la variable* `predicPESO` *las predicciones que genera el modelo con la orden* `predict` *para los valores de la variable* `ALTURA` *del* `data.frame pred`*:*

```
> predicPESO <-predict(nombreModelo,pred)
```

   Por último se añade la variable `predicPESO` *al conjunto de datos* `pred`*:*

```
>pred<-data.frame(pred,predicPESO)
```

El nuevo conjunto de datos se puede ver en la figura 3.6. Puesto que el rango de valores de la altura es $(167, 194)$*, se estarían realizando tres interpolaciones y dos extrapolaciones para los valores* 197,0 *y* $201,8$*; además, puesto que* $\bar{x} = 177,1$*, la predicción más fiable corresponde al valor* 180,3 *y la menos al valor* 201,8*.*

7. *Análisis de Residuos. Para obtener los residuos, tanto absolutos como estudentizados, se selecciona de nuevo* `Modelos`→ `Añadir las estadísticas de las observaciones a los datos...` *y se marcan las opciones correspondientes, generándose por parte de* **R** *dos nuevas columnas en el fichero de datos activos, denominadas* `residuals.(RegModel.1)` *y* `rstudent.(RegModel.1)`*, donde RegModel.1 hace referencia al modelo usado.*

   Aunque en este capítulo se está abordando la regresión desde un punto de vista descriptivo y por tanto no se exigen condiciones a los datos, resulta interesante hacer una diagnosis de los residuos que detecte básicamente problemas de mala elección del modelo, existencia de otras variables relevantes, presencia de valores atípicos,... Para ello se suelen utilizar algunas representaciones gráficas, entre las que destacan la de `Residuos indexados` *y la de* `Residuos frente a ajustados`*. De su observación se pueden extraer valiosas conclusiones.*

   - ***Residuos indexados****. Detecta sobre todo problemas relacionados con la influencia que valores previos de la variable X ejercen sobre los posteriores. Ocurre sobre todo cuando la variable independiente es el tiempo, desde el punto de vista estadístico se dice que existe un problema de autocorrelación y la solución pasa por enfocar el tema desde la óptica de las* series temporales*. El gráfico de los residuos indexados se obtiene desde* `Gráficas`→`Gráfica secuencial...` *seleccionando la variable* `residuals.RegModel.1`*, la opción* `Identificar puntos con el ratón` *y por último elegir la representación por* `puntos`*. En este caso, la figura 3.7 presenta una distribución de residuos*

sin ninguna relación y no se obtiene mayor anormalidad que la existencia de los candidatos a valores atípicos.

- ***Residuos estudentizados frente a valores ajustados.*** *Es probablemente el gráfico que proporciona más información sobre la calidad del ajuste realizado, informando sobre la falta de linealidad de la relación, la presencia de valores atípicos, la existencia de terceras variables que aportarían información relevante sobre Y, etc.*

  Usando las opciones `Gráficas→Diagrama de dispersión...`, *tomando* `fitted.RegModel.1` *como variable explicativa y* `rstudent.RegModel.1` *como explicada, se obtiene la figura 3.8. En el que, al igual que en el gráfico de residuos indexados, sólo destaca la presencia de los candidatos a valores atípicos.*

- ***Obtención de valores influyentes.*** *Se buscan ahora valores especialmente determinantes a la hora de estimar los parámetros del modelo. Normalmente estos valores van a coincidir con valores extremos para una de las dos variables. Uno de los criterios para detectar estos valores influyentes se basa en el cálculo de la* distancia de Cook. *La distancia de Cook para la observación i-ésima calcula la diferencia entre los parámetros del modelo que se obtiene incluyendo la observación i-ésima y sin incluirla. En general se deben tener en cuenta aquellas observaciones cuya distancia de Cook sea mayor que 1. La figura 3.8, se genera a través de* `Gráficas→ Gráfica secuencial...` *y se puede apreciar que los valores más influyentes coinciden con las observaciones 41, 61 y 66.*

  Otra forma de ver la influencia de una observación es a través de su potencial, que estima el peso de cada observación a la hora de realizar predicciones. Los potenciales se obtienen como los elementos de la diagonal principal de la matriz de Hat, $H = X(X'X)^{-1}X'$. En la figura 3.9 se tienen la representación indexada de los potenciales Hat, realizada a partir de la misma opción gráfica anterior. Los puntos influyentes serían aquellos que superaran el doble del cociente entre el número de variables regresoras más uno y el número de observaciones. En este caso el valor de referencia es $0,074$ y los

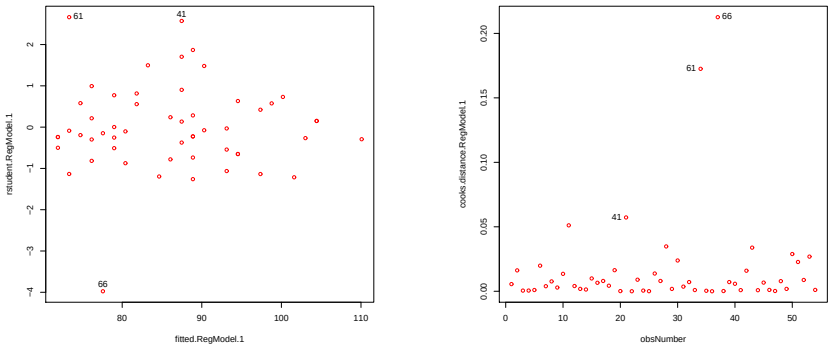


Figura 3.8: Residuos estundentizados frente a Y^* y distancias de Cook

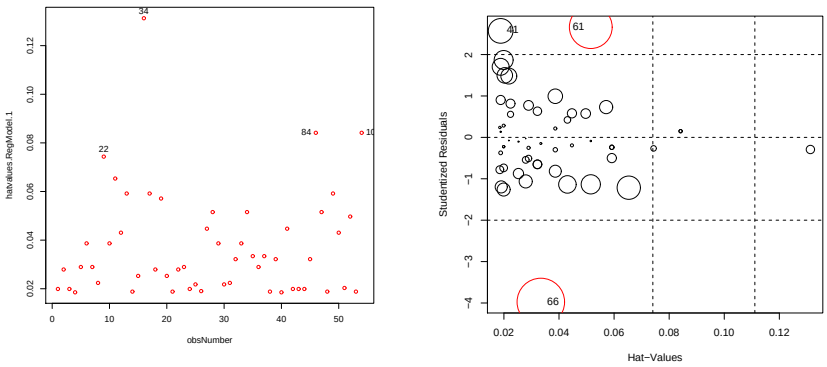


Figura 3.9: Potenciales Hat y puntos influyentes

puntos que superan esta cota son el 32, el 34, el 84 y el 100.

Por último, la gráfica de potenciales hat frente a residuos estudentizados, donde cada observación está identificada por un círculo cuyo diámetro es proporcional a su distancia de cook, sintetiza toda la información a tener en cuenta a la hora de identificar los puntos influyentes. La gráfica ha sido creada desde `Modelos→Gráficas→ Gráfica de influencia` *y refleja de nuevo que los valores a considerar son el 61 y el 66, ver figura 3.9.*

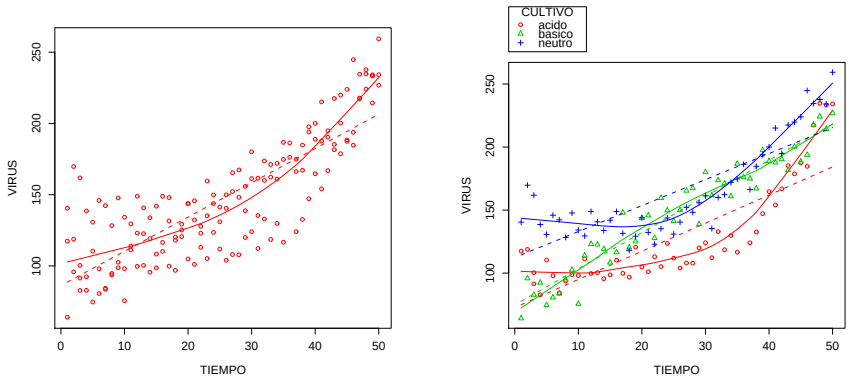


Figura 3.10: Dispersión y dispersión según cultivo

R–Nota 3.3

Supóngase un conjunto de datos del cual se desea obtener un modelo para un subconjunto de estos datos. Por ejemplo en los datos `peso_altura` *se quiere hacer un modelo para los datos femeninos, se selecciona* `Estadísticos→Ajuste de modelos→Regresión lineal...` *y en la ventana de diálogo aparecerá la opción* `Expresión de selección` *donde se puede elegir el subconjunto deseado, en este caso* `SEXO=='Mujer'`*. El problema surge si se quiere añadir, por ejemplo, la columna de valores ajustados seleccionando* `Modelos→ Añadir estadísticas de las observaciones a los datos...`*, esto se debe a que el conjunto de datos activos no se corresponde con el modelo activo, para solucionar esto, sólo se debe hacer en primer lugar el filtrado de los datos para el subconjunto y seguidamente aplicar el modelo.*

Ejemplo 3.4

Para ilustrar la realización de un ajuste de tipo polinomial, se consideran los datos del fichero `reproduccion_vir.dat` *en el que se muestran el* `número de virus` *reproducidos en función del* `tiempo (minutos)` *y de la* `temperatura (grados)`*, según el* `tipo de cultivo (ácido,`

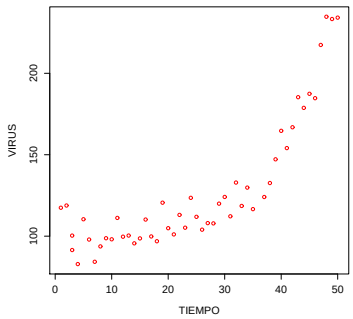


Figura 3.11: Diagrama de dispersión del cultivo ácido

básico o neutro). *Se está interesado en ver como influye el tiempo en el número de virus.*

Se realiza en primer lugar el diagrama de dispersión de la variable `número de virus` *frente al* `tiempo`*. La observación de la figura 3.10 revela para el conjunto de datos una disposición no lineal, aunque la evidente variabilidad presente en cualquier rango de valores del tiempo hace presuponer que el factor* `tipo de cultivo` *debería tenerse en cuenta (figura 3.10).*

Si se rehace el gráfico para cada uno de los subgrupos que determina la variable cultivo, se observa que los cultivos de `tipo básico` *tienen un comportamiento aproximadamente lineal, mientras los de* `tipo neutro` *y* `ácido` *no lo tienen.*

El estudio se centrará en el cultivo ácido, para ello se filtran los datos (se almacenan como `reproduccion_vir_acido`*) y se representan de nuevo. El diagrama de dispersión, figura 3.11, parece sugerir un comportamiento de tipo parabólico.*

Para realizar el ajuste parabólico se selecciona `Estadísticos`→ `Ajuste de modelos`→`Modelo lineal...`*, tomando como fórmula del modelo VIRUS∼ 1+ TIEMPO+ I(TIEMPOˆ2) (figura 3.12). Los resultados obtenidos son:*

```
> LinearModel.3 <- lm(VIRUS ~ 1 + TIEMPO +I( TIEMPO^2),
data=acido)
summary(LinearModel.1)
 Call:
lm(formula = VIRUS ~ 1 + TIEMPO + I(TIEMPO^2), data = acido)

Residuals:

Min       1Q       Median     3Q     Max
-23.295   -6.140   1.510      6.491  24.271

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 115.552345 4.917038 23.500 < 2e-16 ***
TIEMPO -2.901809 0.455127 -6.376 7.25e-08 ***
I(TIEMPO^2) 0.101647 0.008731 11.642 1.89e-15 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 11.73 on 47 degrees of freedom
Multiple R-Squared: 0.9179, Adjusted R-squared: 0.9144
F-statistic: 262.8 on 2 and 47 DF, p-value: < 2.2e-16
```

Se concluye que el tiempo explica casi el 92 % *del número de virus a través del ajuste parabólico estimado.*

Después de representar el gráfico de dispersión de la variable `VIRUS` *frente al* `TIEMPO` *(de los datos* `reproduccion_vir_acido`*) (figura 3.11) es posible representar en la misma ventana la parábola del modelo (figura 3.12) mediante las instrucciones:*

```
> x<- seq(0,50)
> y<- 115,552345 − 2,901809*x + 0,101647*x^2
> lines(x,y,col=''green'')
```

Llegados a este punto, se podría plantear si los datos se ajustarían mejor a un polinomio de grado tres. Aunque no existen evidencias en el gráfico de dispersión, se procederá a realizar este ajuste por motivos básicamente pedagógicos.

Al ser un modelo más general que el parabólico se producirá una mejora del ajuste, aunque la cuestión es si esta mejora es lo suficientemente importante para justificar la mayor complejidad del modelo.

Para realizar el ajuste de grado tres, se selecciona `Estadísticos`→ `Ajuste de modelos`→`Modelo lineal...`*, tomando como fórmula del modelo VIRUS*∼ *1+ TIEMPO+ I(TIEMPO^2)+I(TIEMPO^3) (figura 3.13).*

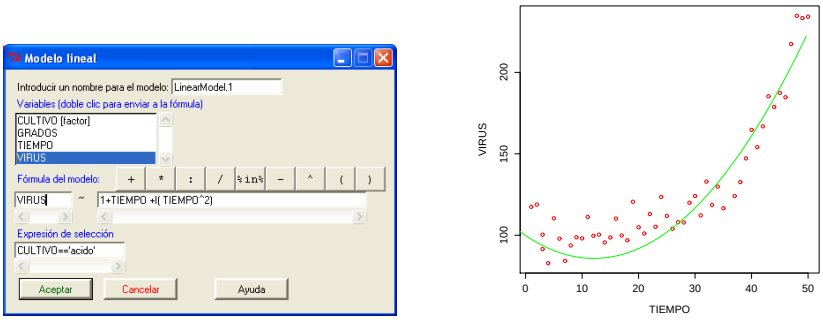


Figura 3.12: Opciones y representación del modelo parabólico

```
> summary(LinearModel.2)
 Call:
lm(formula = VIRUS ~ 1 + TIEMPO + I(TIEMPO^2) + I(TIEMPO^3),
data = Virus_acido)
Residuals:

Min        1Q       Median      3Q     Max
-21.1995   -5.1259  -0.1860    7.1273  21.0148

Coefficients:

                 Estimate       Std. Error       t value       Pr(> |t|)
(Intercept)      98.1018701       5.6855078       17.255       < 2e-16 ***
TIEMPO           1.1938655        0.9905237        1.205       0.2343
I(TIEMPO^2)      -0.1006612       0.0457034       -2.202       0.0327 *
I(TIEMPO^3)      0.0026659        0.0005944        4.485       4.83e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 9.892 on 46 degrees of freedom
Multiple R-Squared: 0.9429, Adjusted R-squared: 0.9392
F-statistic: 253.2 on 3 and 46 DF, p-value: < 2.2e-16
```

El coeficiente de determinación es igual a 0,9429, con una mejora de un 2 %, lo que no parece justificar la adopción de este modelo más complejo. Igual que antes es posible representar el ajuste cúbico como puede observarse en la figura 3.13.

R–Nota 3.4

Para realizar un ajuste polinomial con ***Rcmdr*** *se selecciona la opción*

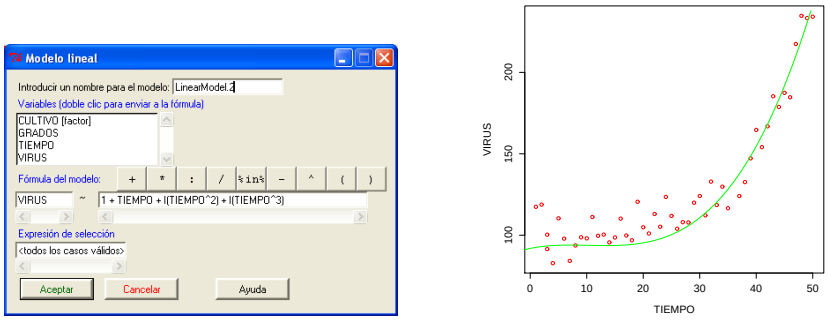


Figura 3.13: Opciones y representación del modelo cúbico

`Estadísticos`→`Ajustes de modelos`→`Modelo lineal...` *y en la ventana de diálogo se escribe la expresión del modelo deseado:*

- *Para indicar un modelo lineal con término independiente se escriben cualquiera de las dos fórmulas siguientes:*

$$Y \sim X$$

$$Y \sim 1 + X$$

- *Si se desea omitir el término independiente en un modelo lineal se utiliza una de las fórmulas siguientes:*

$$Y \sim -1 + X$$

$$Y \sim 0 + X$$

- *En general para un modelo polinomial con término independiente se escribe:*

$$Y \sim X + I(X^2) + I(X^3) + \cdots + I(X^n) \text{ o bien}$$

$$Y \sim 1 + X + I(X^2) + I(X^3) + \cdots + I(X^n)$$

y con un -1 *ó* 0 *para un modelo sin término independiente.*

Si se quiere observar la notación que utiliza **R** *para formular estos modelos, véase el apéndice C.*

4. Ejercicios

3.1 Para los datos del fichero `peso_altura.dat`, analice el comportamiento del peso en función de la altura para el grupo de las mujeres.

3.2 La tabla 3.2 muestra una serie histórica sobre el olivar español que recoge la superficie, rendimiento y producción, durante el periodo 1965-1979, donde:
X = Superficie en miles de Ha.
Y = Rendimiento en Qm/Ha.
Z = Producción en miles de Tm.

Se pide:

a) El diagrama de dispersión de las variables X e Y.

b) Las medidas más representativas para cada una de las variables, indicando su representatividad.

c) El estudio de la relación entre las variables XY, XZ e YZ.

3.3 La siguiente tabla muestra la relación existente entre la lluvia caída, en l/m^2, en el periodo octubre–mayo y la producción obtenida en kilogramos por olivo.

X	300	400	500	600	700
Y	13	26	40	57	64
Y	24	21	31	45	69
Y	17	17	38	51	57
Y	11	26	34	58	76
Y	20	30	27	44	74

donde X representa la lluvia e Y la producción.

a) Represente el diagrama de dispersión.

b) Indique si existe alguna tendencia.

c) Cuantifique y comente la relación existente entre las dos variables.

Año	X	Y	Z
1965	73,6	69,8	8,5
1966	98,1	62,5	6
1967	99,8	98,5	8,7
1968	107,7	102,5	6
1969	107,7	97,4	3,7
1970	122	113,8	8,9
1971	127	118	7,9
1972	138,1	128,1	10,1
1973	152,1	145,8	6,8
1974	144,8	139,8	5
1975	160,7	152,9	11,1
1976	150,2	143,4	9,8
1977	152,1	146	9,5
1978	167,3	162,1	10,8
1979	165	160,2	10

Tabla 3.2: Datos ejercicio 3.2

3.4 Dada la siguiente tabla de doble entrada con valores porcentuales:

$Y \setminus X$	2	3	4
0	0,22	0,13	0,04
1	0,16	0,11	0,05
2	0,08	0,16	0,05

a) Obtenga la distribución marginal de X. Calcule su media, moda y mediana.

b) Calcule la media de Y cuando X toma el valor 3.

c) Estudie la dependencia de las variables X e Y.

3.5 A un grupo de estudiantes se les preguntó por el tiempo que tardan en llegar desde su hogar hasta la facultad, X (minutos), el tiempo que le dedican diariamente al estudio, Y (horas), y las calificaciones

obtenidas en la asignatura de Estadística, Z, obteniéndose las siguientes respuestas:

(40, 4, 4), (45, 3, 3), (30, 4, 5), (40, 4, 5), (80, 2, 5), (20, 3, 5)
(10, 1,5, 6), (10, 4, 6), (20, 4, 6), (45, 3, 3), (20, 4, 4), (30, 4, 7)
(30, 3, 7), (20, 4, 6), (30, 1, 6), (10, 5, 5), (15, 5, 5), (20, 6, 5)
(20, 3, 7), (20, 4, 5), (20, 5, 6), (60, 2, 3), (60, 5, 5)

a) Obtenga el diagrama de dispersión correspondiente al tiempo dedicado al estudio y las calificaciones obtenidas en Estadística.

b) ¿Se aprecia alguna tendencia?

c) Estudie las relaciones existentes entre XY, XZ e YZ.

3.6 Al mismo grupo del ejercicio anterior se le ha pedido que escriba un dígito al azar entre 0 y 9 así como el número de hermanos que tiene, obteniéndose los siguientes pares de valores:

(7, 4), (0, 1), (2, 1), (2, 0), (9, 4), (7, 4), (6, 3), (8, 5)
(7, 3), (3, 2), (7, 3), (2, 1), (7, 4),(7, 3), (8, 4), (8, 5)
(5, 3), (3, 1), (4, 2), (4, 2), (5, 3), (2, 0), (4, 2)

¿Existe alguna relación entre las variables?, ¿de qué tipo?

3.7 Se examinan 300 alumnos de una asignatura y durante el examen se les pregunta por el tiempo que han dedicado a su preparación (menos de una hora, entre una hora y tres, más de tres), obteniéndose la siguiente tabla de calificaciones según el tiempo de estudio:

Nota \ Horas Estudio	< 1	$1 - 3$	> 3
Suspenso	43	32	10
Aprobado	31	48	81
Notable	7	13	20
Sobresaliente	3	4	8

¿Están relacionadas las calificaciones con las horas de estudio?

3.8 Dada la distribución:

X	1	1,5	2	2,5	3	3,75	4,5	5
Y	1	1,5	2,95	5,65	8,8	15	25	32

a) Elija la mejor clase funcional para ajustar Y/X y estime sus parámetros.

b) Establezca la bondad del ajuste.

c) Calcule la previsión para Y cuando $X = 7$. Analice dicha previsión.

3.9 Dada la distribución:

X	$2,5$	$3,75$	5	$7,5$	10	$12,5$	20
Y	8	14	$23,75$	40	62	90	165

a) Utilice una ecuación del tipo aX^b para ajustar Y/X.

b) Dé una medida de la bondad del ajuste.

3.10 Dada la distribución:

X	1	$1,5$	2	3	4	5	6	7
Y	1	$1,75$	$2,65$	$4,7$	7	$9,5$	12	15

a) Ajuste Y/X utilizando una función del tipo aX^b.

b) Analice la bondad del ajuste.

3.11 Dada la distribución:

X	5	6	8	10	13	18	20
Y	$1,5$	$1,25$	$0,93$	$0,7$	$0,46$	$0,23$	$0,15$

a) Estime los parámetros de la clase funcional $ab^{-0,2X}$ para ajustar Y/X.

b) Estudie la bondad del ajuste.