zsPráctica 3. Análisis de datos bidimensionales. Variables cuantitativas.

1. Introducción.

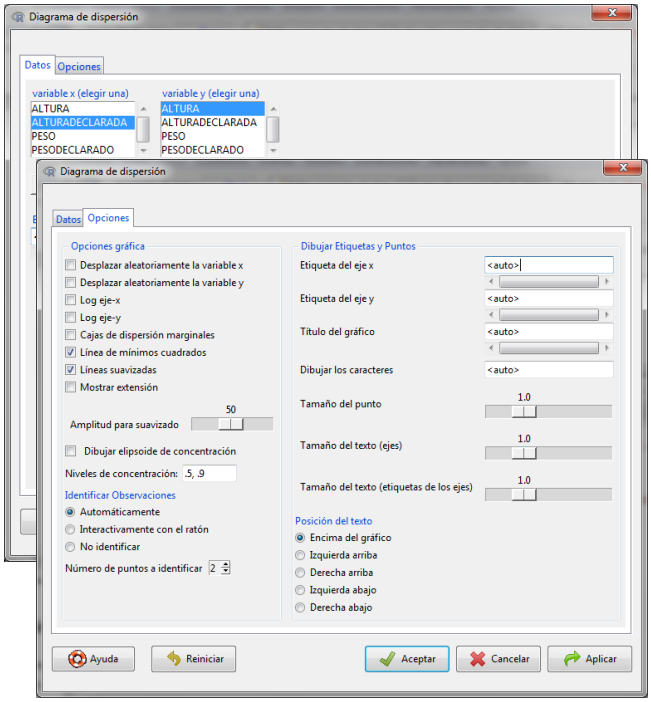
En esta práctica veremos el uso del análisis de regresión en la descripción de datos bidimensionales. El fichero que utilizaremos será el de Davis2 (Davis corregido) y, como ejemplo, comprobaremos hasta qué punto la altura real de las personas del estudio (Y) se ajusta a la altura que dichas persona declaran que tienen (X). Cuando queremos realizar una regresión, la X juega el papel de variable explicativa (causa), mientras que la Y juega el papel de variable que queremos predecir (efecto). En este caso tiene interés, cuando no sabemos la altura de una persona y se lo preguntamos a ella, cuál es la altura real que podemos 'deducir' de su respuesta. En este punto, también conviene decir que, en general, no deberíamos extrapolar nuestros resultados (recuérdese que los datos pertenecen a personas canadienses que practican deporte de modo regular) a cualquier población (por ejemplo españoles sedentarios), puesto que las características de las poblaciones pueden diferir y dar lugar a resultados erróneos.

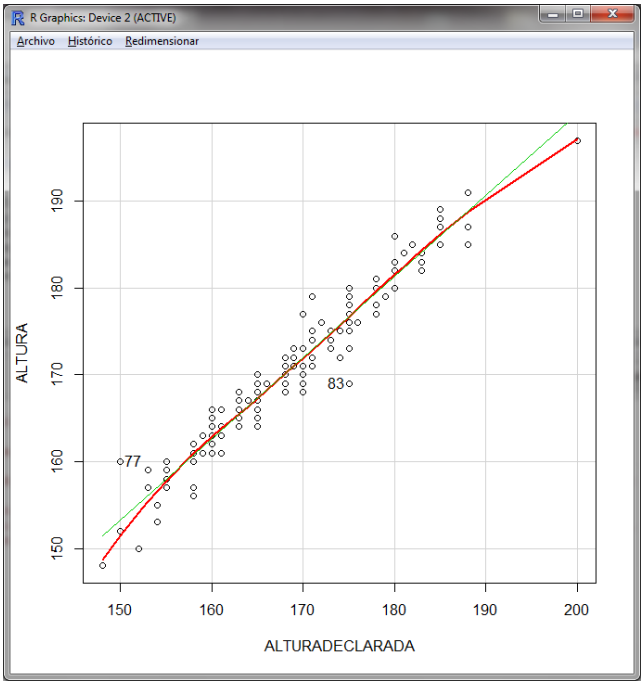
2. Dibujo de la gráfica de dispersión.

El primer paso es la representación gráfica de las dos variables. Para ello, vamos al menú **Gráficas >Diagrama de dispersión**, seleccionamos como X altura declarada, como Y altura, y en la pestaña **"Opciones"** pedimos que dibuje la línea de mínimos cuadrados.

Observamos que hay una relación positiva entre las dos variables (obviamente, a mayor altura declarada, mayor altura de la persona). En la pestaña **"Opciones",** podemos "identificar observaciones" automáticamente. En tal caso, quedan marcados dos datos, el 77 y el 83, como posibles "atípicos" (más adelante discutiremos la influencia de datos atípicos en la recta de regresión).

Nota: Asimismo, en la pestaña **"Opciones",** podemos activar la opción 'Líneas suavizadas'. Se adjuntará al dibujo una curva obtenida mediante suavizado local de los datos (suavizado loess, que es un ajuste polinómico local). Su forma global sirve para detectar posibles desviaciones de la linealidad en los datos.





3. Cálculo del coeficiente de correlación.

Para el cálculo del coeficiente de correlación de Pearson entre dos variables, hay que usar la opción: **Estadísticos>Resúmenes>Test de correlación**. Seleccionando las variables *altura* y *altura declarada* en el cuadro de diálogo que aparece, y dejando el resto de opciones por defecto, obtendremos:

Pearson's product-moment correlation

data: ALTURA and ALTURADECLARADA

t = 60.029, df = 181, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:
0.9677122 0.9818710

sample estimates:
cor
0.9757937

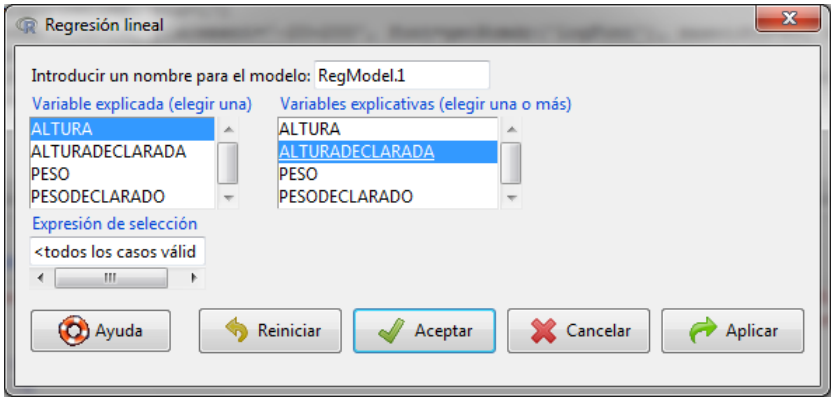
Sale información muy interesante, pero que no sabemos interpretar todavía. Nos tenemos que fijar en el final, donde se nos proporciona el coeficiente de correlación (señalado en negrita). Esto es, r=0.9758, lo que indica una correlación muy alta y de sentido positivo, como ya habíamos visto en el diagrama de dispersión.

Nota: Cuando el coeficiente de correlación es bajo, suelen aparecer dudas sobre si existe de verdad relación lineal entre las variables. En esos casos (y en general) si queremos saber si el valor del coeficiente de correlación que hemos obtenido es indicativo de la presencia o ausencia de relación (lineal) entre las variables, en la población de la que hemos extraído los datos, podemos hacer un contraste de hipótesis (de los que se hablará más adelante). En este contraste la hipótesis nula es la ausencia de relación lineal entre las variables. El valor que hay que mirar en el resumen anterior es el p-value (p-valor). Valores menores que 0.05 rechazarían la hipótesis nula (es decir, la ausencia de relación) y concluiríamos que las variables no son

independientes. En nuestro caso, el p-valor es menor que 2.2e-16, indicando que las variables no son independientes (existe por tanto una relación entre las variables, y dicha relación es lineal, como se ha visto en el gráfico de dispersión).

4. Valores de la recta de ajuste y cálculo del coeficiente de determinación.

Aunque en el gráfico de dispersión anterior ha salido dibujada la recta de ajuste, no conocemos sus coeficientes. Para calcularlos, vamos a **Estadísticos> Ajuste de modelos> Regresión lineal…**, y elegimos *Altura* como variable explicada y *Alturadeclarada* como variable explicativa.



De nuevo, en la salida de resultados hay muchas cosas, pero sólo vamos a fijarnos en lo señalado en negrita:

```
Call:
lm(formula = ALTURA ~ ALTURADECLARADA, data = Davis)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6567 -1.3023  0.2142  1.3433  6.7295

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      12.95340    2.62985   4.926 1.89e-06 ***
ALTURADECLARADA   0.93545    0.01558  60.029  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.99 on 181 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared:  0.9522,   Adjusted R-squared:  0.9519
F-statistic:  3603 on 1 and 181 DF,  p-value: < 2.2e-16
```

El valor que acompaña a Intercept es la ordenada en el origen, esto es, a*=12.95340, y el valor que acompaña a ALTURADECLARADA es la pendiente, esto es: b*=0.93545. Así que la recta de regresión es:

$$Alturapredicha = \mathbf{0.93545} \cdot Alturadeclarada + \mathbf{12.95340}.$$

Lo que se denomina Multiple R-squared es r^2, el coeficiente de determinación, que en este caso es muy alto (0.9522). El coeficiente de determinación es una medida de la bondad del ajuste y se suele interpretar en términos de variabilidad de la variable Y explicada a través el modelo. En nuestro caso

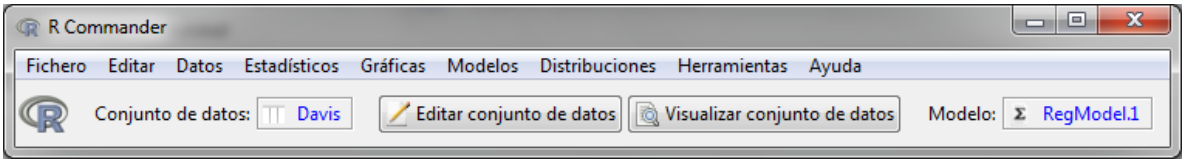
particular, el 95.22% de la variabilidad de la variable *Altura* viene explicada por la variable *Alturadeclarada* a través de este modelo de regresión simple.

Nota: En el cuadro de diálogo anterior, dentro de la opción "expresión de selección", se podría haber seleccionado un subconjunto de los datos para realizar la regresión entre esas dos variables sólo para ese subconjunto de datos (por ejemplo, hubiésemos puesto SEXO=="F" si hubiésemos querido realizar la regresión sólo para las mujeres). Normalmente no utilizaremos esta opción para realizar regresiones para algunos subgrupos, pues tiene algún inconveniente.

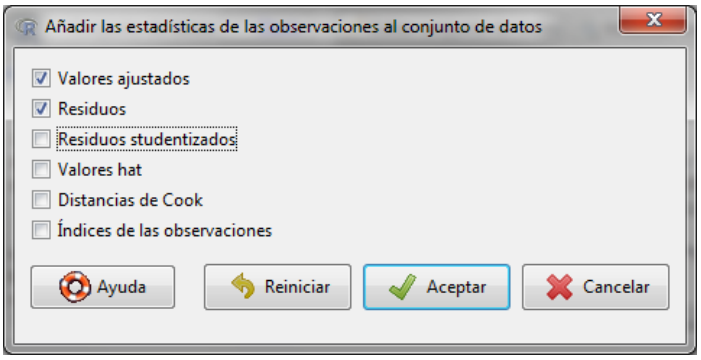
Ejercicio: Anota cuánto mides (que sería tu valor de *alturadeclarada*). Calcula el valor de altura que predice para ti la anterior recta de regresión (aunque puede no ser una buena predicción, ya que tú no eres un chico canadiense que practica deporte de forma regular).

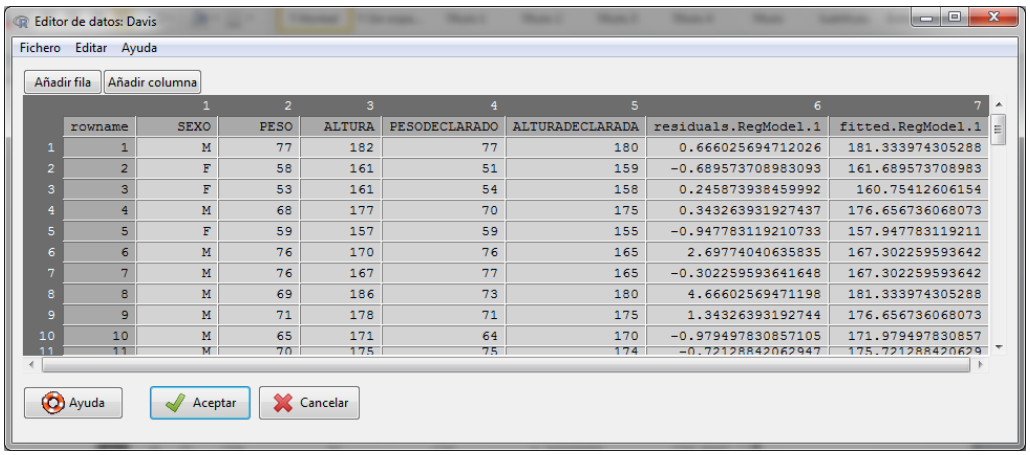
5. Valores ajustados.

Al haber ajustado un modelo a nuestros datos, el menú principal tiene el siguiente aspecto:



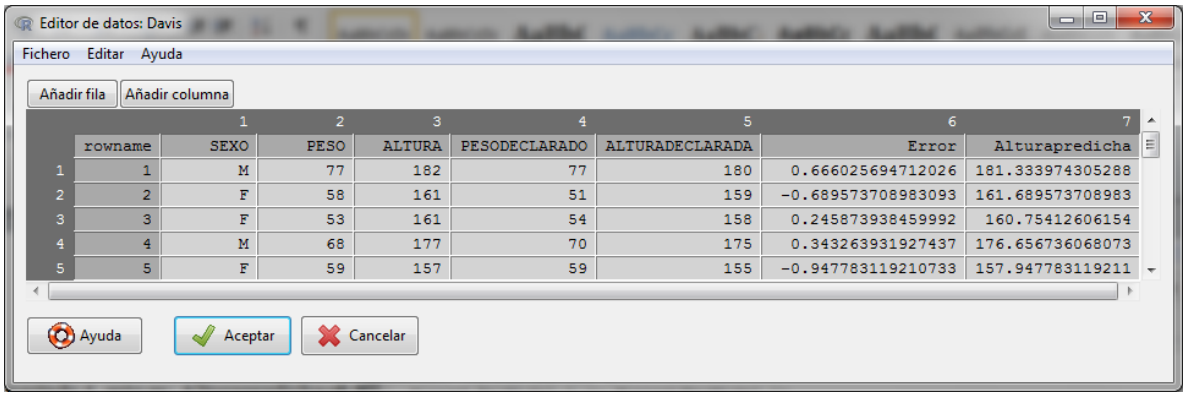
En el campo "Modelo" aparece RegModel.1 (es el nombre que R ha puesto por defecto al ajuste lineal entre las variables altura y alturadeclarada que hemos realizado anteriormente, pero en el cuadro de diálogo asociado a **Estadísticos> Ajuste de modelos> Regresión lineal…**, le podríamos haber puesto otro nombre al modelo). Lo importante es que con ese modelo podemos crear en nuestro fichero la columna de valores ajustados y la columna de los errores (residuos) del modelo. Para ello, vamos a **Modelos> Añadir las estadísticas de las observaciones a los datos…** y dejamos activadas las casillas correspondientes a "Valores ajustados" y "Residuos". Tras aceptar, al visualizar los datos con el editor veremos que se han añadido dos nuevas variables.





	rowname	SEXO	PESO	ALTURA	PESODECLARADO	ALTURADECLARADA	residuals.RegModel1.1	fitted.RegModel1.1
1	1	M	77	182	77	180	0.666025694712026	181.333974305288
2	2	F	58	161	51	159	-0.689573708983093	161.689573708983
3	3	F	53	161	54	158	0.245873938459992	160.75412606154
4	4	M	68	177	70	175	0.343263931927437	176.656736068073
5	5	F	59	157	59	155	-0.947783119210733	157.947783119211
6	6	M	76	170	76	165	2.69774040635835	167.302259593642
7	7	M	76	167	77	165	-0.302259593641648	167.302259593642
8	8	M	69	186	73	180	4.66602569471198	181.333974305288
9	9	M	71	178	71	175	1.34326393192744	176.656736068073
10	10	M	65	171	64	170	-0.979497830857105	171.979497830857

La segunda de las variables (fitted.RegModel1.1) son los valores ajustados por la recta calculada en el apartado 4, esto es: Alturapredicha=**0.93545**·Alturadeclarada+**12.95340**. Si queremos, podemos cambiarle el nombre en el editor y llamarle "Alturapredicha". La primera de las variables (residual.RegMode1.1) son los errores del modelo, esto es: Altura-Alturapredicha. Podemos cambiarle también el nombre y llamarle "Error". Si tras hacer los cambios visualizamos el conjunto de datos, éste tiene el siguiente aspecto:



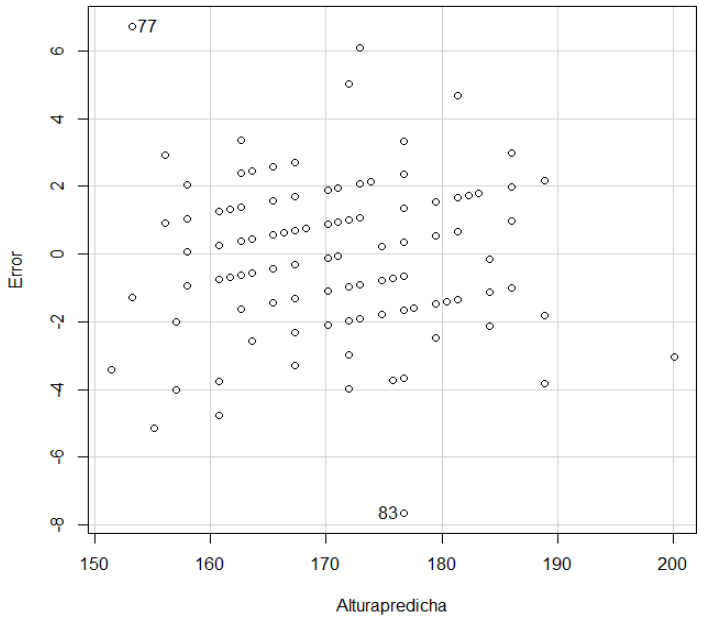
	rowname	SEXO	PESO	ALTURA	PESODECLARADO	ALTURADECLARADA	Error	Alturapredicha
1	1	M	77	182	77	180	0.666025694712026	181.333974305288
2	2	F	58	161	51	159	-0.689573708983093	161.689573708983
3	3	F	53	161	54	158	0.245873938459992	160.75412606154
4	4	M	68	177	70	175	0.343263931927437	176.656736068073
5	5	F	59	157	59	155	-0.947783119210733	157.947783119211

Por ejemplo, el primer individuo declaró que medía 180 cm (alturadeclarada). Según nuestro modelo, se calcula que mide 181.3340 (alturapredicha). En realidad esta persona mide 182 (altura), por lo que hemos cometido un error de 0.666 cm.

Ejercicio (continuación del anterior): Si vas a una farmacia, o te mides en casa, podrás calcular tu altura (tu dato para la variable altura). Altura-Alturapredicha, sería tu error de predicción.

6. Análisis de residuos.

Conviene representar gráficamente mediante un diagrama de dispersión (**Gráficas> Diagrama de dispersión**) los valores de los residuos (sería la Y del diagrama) frente a los valores ajustados (sería la X del diagrama). Lo ideal es que dichos residuos formasen una 'nube de puntos' (esto es, correlación cercana a 0), pues esto significaría que mediante nuestro modelo hemos eliminado toda la asociación en nuestros datos. Esta representación gráfica también sirve para ver si existen observaciones atípicas que podrían haber llegado a distorsionar nuestros resultados.

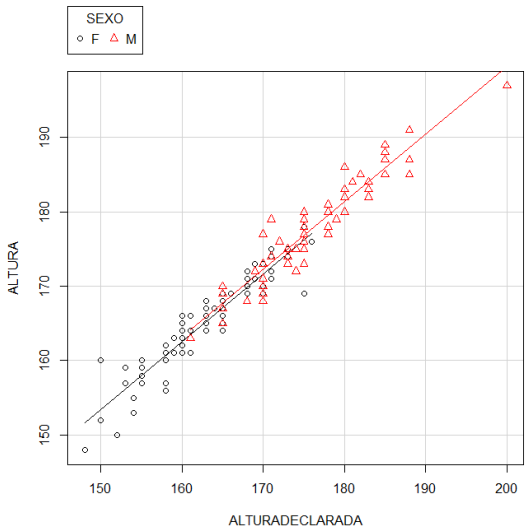


Vemos que la gráfica resulta bastante satisfactoria. Al igual que antes, si en la pestaña **"Opciones"** hemos activado "Identificar observaciones" automáticamente, notamos que el error más extremo se produce para el dato 83 (con una equivocación de casi -8 centímetros). Como se ha comentado con anterioridad, dicho dato no se separa excesivamente de la nube de puntos.

Nota: En el caso de regresión lineal, el gráfico de los residuos frente a la variable explicativa proporcionaría la misma información. Pero con vistas a modelos más generales, es mejor colocar en el eje de las X los valores ajustados.

7. ¿Merece la pena separar la población según los niveles de un factor?.

Para observar si merece la pena el estudio para los diferentes niveles de un factor (por ejemplo, para hombres y mujeres por separado), en el cuadro de diálogo asociado a **Gráficas> Diagrama de dispersión**, se seleccionaría la variable sexo pulsando el botón "**Gráfica por grupos**". De esta manera, se ajustarían rectas diferentes para hombres y mujeres. Obtenemos el siguiente resultado, en el que se observa cómo no existe prácticamente diferencia entre las dos rectas. Las opciones elegidas en la pestaña "Opciones" has sido "Línea de mínimos cuadrados" y "No identificar" observaciones.



Nota 1: En este caso la diferencia no parece sustancial, pero en algunas ocasiones, el realizar análisis separados por grupos heterogéneos puede mejorar apreciablemente los resultados de la regresión.

Nota 2: Aun cuando en el diagrama de dispersión las rectas pueden dibujarse separadas para sexos, si se quisieran calcular coeficientes de correlación y rectas de regresión por separado para cada sexo, tendríamos que utilizar la opción **Datos> Conjunto de datos activo> Filtrar el conjunto de datos activo**, tal y como se hizo en la primera práctica. De esta forma, tendríamos separados los datos correspondientes a los hombres y a las mujeres en dos ficheros distintos (ficheros Davismasc y Davisfem de la práctica 1).

Otra forma sería seleccionar el subconjunto deseado en el cuadro de diálogo de "**Regresión lineal**", como explicamos más arriba. Pero así tendríamos el inconveniente de que no podríamos luego pedir valores ajustados y errores en "Añadir estadísticas de las observaciones a los datos".

8. Análisis de observaciones influyentes.

En el análisis de residuos de los datos anteriores, hemos observado un dato algo extremo (un chico que declaraba una altura 'normal', pero con un error de predicción de -8 cm). Normalmente, dicho dato no afectará mucho al cómputo global de la recta de regresión, salvo en casos de un error disparatado (de hecho, puedes repetir los cálculos eliminando a ese chico del fichero de datos, y observarías que la recta y el R^2 no experimentan muchos cambios). Pero a veces son peligrosos datos con un X extremo que tienen un valor Y que no sigue la pauta de los demás datos.

Para ver un ejemplo, abrimos el fichero tabaco.rda. Este fichero lo utilizamos en la práctica 1, y contiene datos relativos al consumo de alcohol y tabaco en condados de Gran Bretaña. Se cree que el consumo de tabaco es un posible predictor para el consumo de alcohol.

- Dibuja el diagrama de dispersión de alcohol (Y) frente a tabaco (X), junto con su recta de ajuste (¿observas algo?)
- Calcula la recta de regresión. Comprueba que el R^2 no es nada bueno. Teniendo tan pocas observaciones, una de las posibles causas podría ser un 'residuo' extremo. Obtén los residuos del modelo (cuando vayas a 'Añadir las estadísticas de las observaciones a los datos', pide también que saque las 'Distancias de Cook', luego diremos para qué sirve). ¿Notas algún dato extremo en el análisis de residuos?
- Vuelve a realizar el análisis eliminando el dato más extremo del fichero (explicamos cómo eliminar datos en la práctica 2). Observa cómo cambia la recta de regresión, a la par que mejora el R^2.

Nota 1: El problema estaba en el dato de Irlanda del Norte. Su consumo de tabaco es de los más altos (X extrema), mientras que su consumo de alcohol es de los más bajos (Y que no sigue la pauta del resto de los condados). Al eliminar ese dato, la regresión ha mejorado sustancialmente (la situación se agrava por el pequeño número de datos que tenemos).

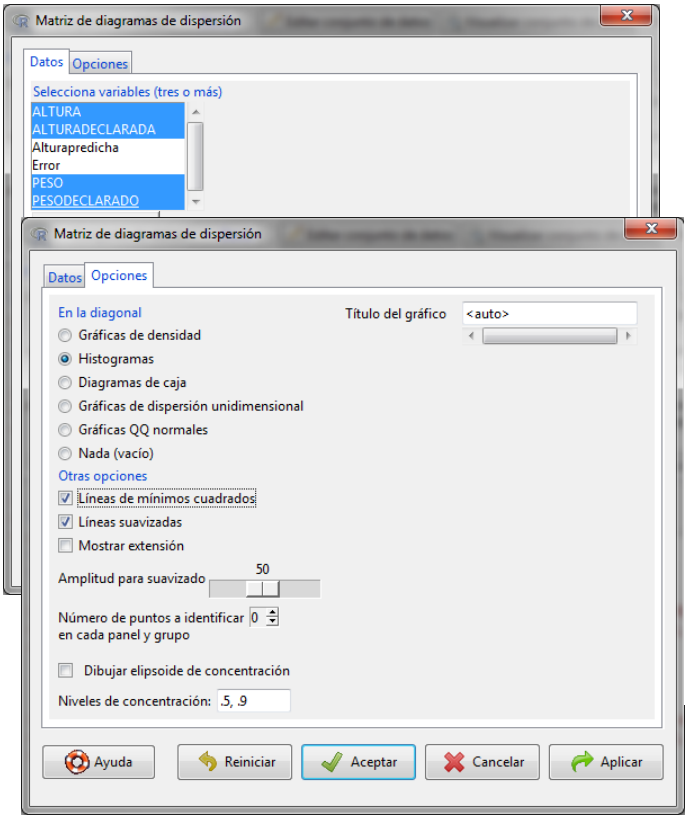
Nota 2: Una herramienta para observar el efecto de observaciones influyentes, aparte del gráfico de residuos, es la variable 'distancias de Cook'. Esta variable mide la diferencia entre los parámetros del modelo que se obtiene incluyendo la observación i-ésima y sin incluirla. En general, se debería 'sospechar' de observaciones cuya distancia de Cook fuese superior a la unidad.

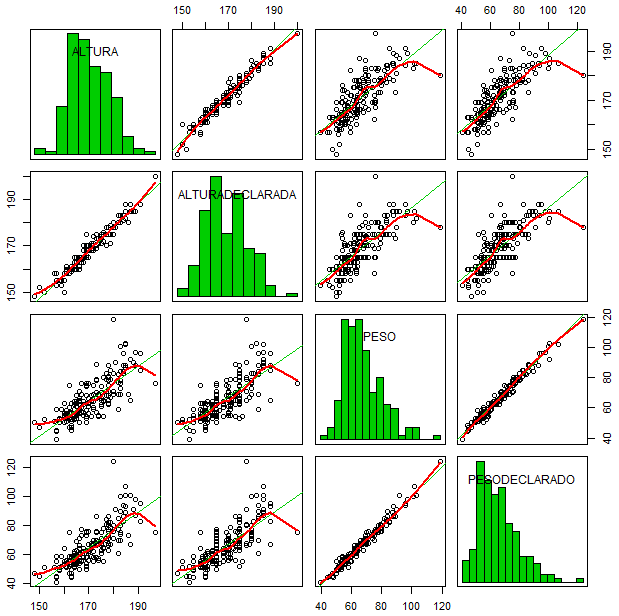
9. Matriz de diagramas de dispersión.

Cuando se tiene un gran número de variables, suele resultar de utilidad representarlas por parejas, para ver relaciones de causalidad más o menos fuertes entre ellas. Ello se puede hacer mediante la opción: **Gráficas> Matriz de diagramas de dispersión.** En el cuadro de diálogo que aparece, las variables que

queramos estudiar se eligen con el ratón y la tecla <Ctrl>. Si en la pestaña "Opciones" activamos "Histogramas", "Líneas de mínimos cuadrados" y "Líneas suavizadas", se obtendrá el gráfico que figura a continuación.

En el gráfico, podemos apreciar las parejas de variables con comportamientos más cercanos a la linealidad. Altura frente a Altura declarada (y su análoga Peso-Peso declarado) presentan una relación lineal muy fuerte. También se observa que hay relación lineal entre altura y peso, aunque no tan pronunciada como en los casos anteriores.





10. Alternativas al ajuste lineal.

Hasta ahora, hemos considerado observaciones bidimensionales (x_i, y_i), y tratado de explicar la Y en función de la X, considerando que los datos 'siguen' una línea recta. Es lo más habitual, pero no siempre un modelo lineal explica bien la realidad. Un modelo general de ajuste trata de explicar la variable Y a través de una función de la variable X, es decir, $Y^*=f(X)$.

En el modelo lineal se considera $f(X) = a + bX$, pero se podrían considera otros modelos, tales como:

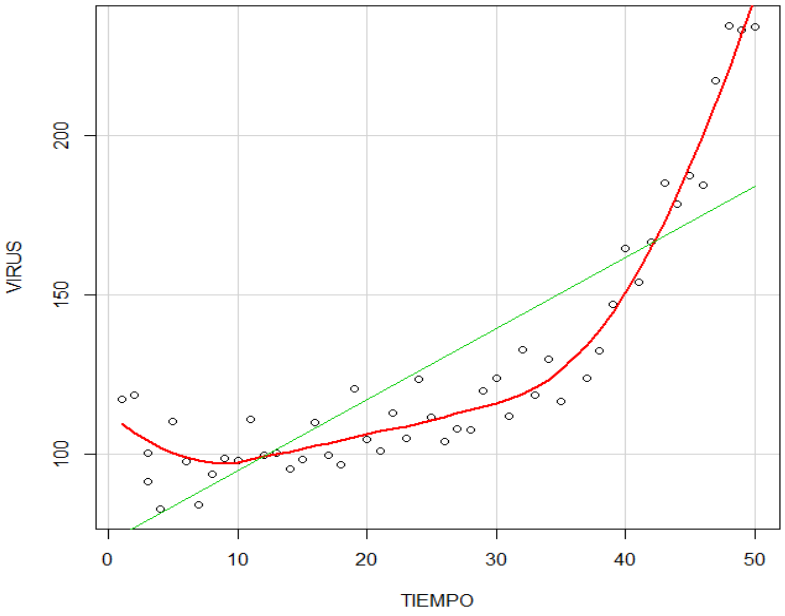
- Una parábola: $Y^* = f(X) = a + bX + cX^2$
- Una función exponencial: $Y^* = f(X) = a\ b^X$
- Una función potencial: $Y^* = f(X) = a\ X^b$

- Una hipérbola: $Y^* = f(X) = a + b/X$

Como en la recta de regresión, los parámetros del modelo se 'elegirían' (ajustarían) por el método de mínimos cuadrados. Todos los modelos indicados arriba, tras una transformación adecuada, se podrían estudiar mediante un modelo lineal (véase nota al final).

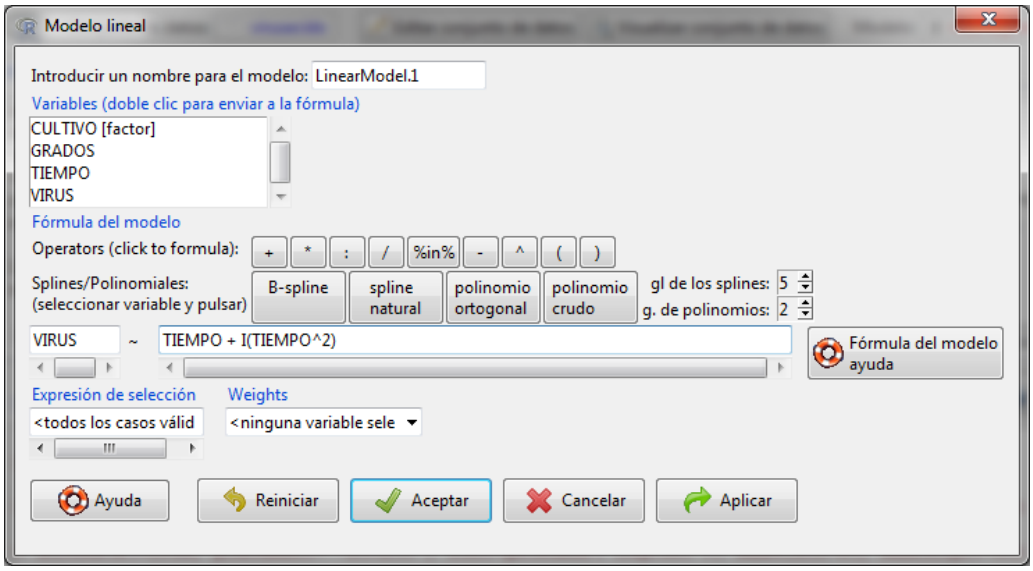
Ejemplo: Abre el fichero virus.rda. En este fichero tenemos unos datos de laboratorio referentes a la reproducción de un virus en un cultivo. La variable "tiempo" representa el tiempo que lleva ese cultivo, la variable "virus" el número de virus presentes en el cultivo y la variable "cultivo" el tipo de cultivo (ácido, básico, neutro). Queremos predecir el número de virus en función del tiempo, pero centrándonos en el cultivo ácido.

- En primer lugar, realizamos un filtrado de datos, para quedarnos con los correspondientes al cultivo ácido.
- En segundo lugar, dibujamos la gráfica de dispersión.



Podemos observar (hemos dibujado la recta de ajuste y la línea suavizada) que los datos se 'comban' y que la recta no es un modelo del todo satisfactorio. En este caso, por su forma, parecería más apropiado un ajuste parabólico.

Para realizar el ajuste parabólico, iremos a **Estadisticos> Ajuste de modelos> Modelo lineal** y completaremos el cuadro de diálogo que aparece de la siguiente forma.



- En "Fórmula de modelo" se coloca a la izquierda la variable Y (virus). Se podría indicar una transformación (por ejemplo, log(virus) indicaría el logaritmo del virus).
- A la derecha, la función de X que se quiere utilizar. Se utiliza + para indicar cada uno de los términos lineales del modelo.
- I(TIEMPO^2) indica que queremos incluir la potencia de grado 2 de la variable tiempo en el modelo (el término independiente se incluye por defecto, no hace falta ponerlo).

Los resultados que se obtendrían serían los siguientes:

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

El programa nos devuelve los coeficientes del modelo, esto es, tendríamos **Viruspredicho** = 115.552345 -2.901809 * **TIEMPO** + 0.101647* **TIEMPO^2**. El R^2, como en el ajuste lineal, nos daría una medida de la bondad del ajuste. Un R^2 próximo a 1 indicaría un buen ajuste.

Notas:

- El modelo se elegiría según la forma de la nube de datos. En caso de duda, se pueden estimar diferentes modelos y comparar su R^2.
- **Precaución**: El modelo cuadrático siempre ajustará mejor que el modelo lineal (es más general). Sin embargo, sólo es recomendable cuando la forma gráfica de los datos indiquen una desviación de la linealidad y cuando el R^2 aumente de modo apreciable con respecto al R^2 del modelo lineal. Para terminar, es necesario realizar también un análisis de residuos frente a valores ajustados.

- **Transformaciones para conseguir linealidad**: Se puede observar que, mediante una transformación apropiada, los modelos indicados al principio se convertirían en modelos lineales. Por ejemplo, si en el modelo exponencial tomamos logaritmos, tendríamos:

  $$\log(Y^*) = \log(a) + X \log(b)$$

  Mediante dicha transformación, pasaríamos al modelo lineal que ya conocemos (esto es, si quisiéramos utilizar este ajuste con nuestros virus, en fórmula del modelo pondríamos a la izquierda **log(VIRUS)** y a la derecha **TIEMPO**). En el modelo hiperbólico, el modelo lineal se consigue definiendo X' = 1/X, con lo que tendríamos el Y* = a + b X' (en la fórmula del modelo, pondríamos a la izquierda **VIRUS** y a la derecha **I(TIEMPO^(-1))**).

  Las transformaciones anteriores se podrían realizar sobre el fichero de datos, y luego utilizar la opción de Regresión Lineal con las variables transformadas.

***Ejercicio*:** Considera otro de los posibles tipos de cultivo (básico o neutro). Indica si sería conveniente un ajuste lineal. Realiza el ajuste lineal y el ajuste cuadrático y comenta si merece la pena este último.

11. Actividad propuesta para entrega voluntaria.

Elige un conjunto de datos que quieras estudiar (o, alternativamente, utiliza algún fichero de datos de los manejados en clase). De los datos elegidos, toma dos variables de tipo continuo que pienses que puedan estar relacionadas y realiza un ajuste lineal. En concreto:

- Indica cuál eliges como X, cuál como Y, y por qué.
- Realiza un breve resumen descriptivo de la variable Y.
- Comprueba si el ajuste lineal es apropiado.
- Realiza el ajuste y calcula las predicciones para dos valores concretos de la variable X: uno dentro del rango de datos y otro fuera de ellos.
- Realiza un análisis de residuos y de valores influyentes.
- Indica si parece necesario un modelo diferente al lineal.
- Piensa en una variable cualitativa que pueda influir en los resultados de la regresión. Comprueba si merecería la pena realizar el análisis por grupos.

12. Ejercicios propuestos.

1) Se quiere analizar si existe alguna relación lineal entre la altura y el peso de las personas. Usando los datos que se encuentran en el fichero Davis2.rda contesta a las siguientes cuestiones.

   - A través del diagrama de dispersión, el coeficiente de correlación, el coeficiente de determinación y la recta de regresión, ¿sería razonable hablar de una relación lineal entre las dos variables?
   - ¿Qué previsión darías para el peso de una persona, sabiendo que su altura es 175 cm?
   - Realiza el mismo estudio anterior sólo para las chicas y haz la previsión del nuevo peso suponiendo que una nueva chica mide 175 cm.
   - Realiza el mismo estudio anterior sólo para los chicos y haz la previsión del nuevo peso suponiendo que un nuevo chico mide 175 cm.
   - ¿Qué diferencias observas entre los tres estudios?