EACS-Estadística descriptiva bivariante: Dos variables cuantitativas

Índice

1. Relación entre dos variables cuantitativas

Cuando analizamos conjuntamente dos variables numéricas (cuantitativas), X e Y, queremos saber si existe alguna relación entre ellas. Si existe relación, nos interesará averiguar si es aproximadamente lineal y, si lo es, cuantificar su intensidad y calcular los coeficientes de la recta que mejor la describa.

Los *diagramas de dispersión* permiten analizar visualmente el aspecto de la relación entre X e Y, el *coeficiente de correlación lineal de Pearson* proporciona una medida cuantitativa de la intensidad de una relación lineal y la *regresión lineal* proporciona los coeficientes de la recta que mejor se ajusta a los datos. En las siguientes secciones veremos cómo realizar estos análisis con R Commander.

1.1 Diagrama de dispersión

Los diagramas de dispersión representan los datos (x,y) de la pareja de variables numéricas (X,Y) como puntos en un plano. Por lo tanto, permiten obtener información visual sobre el tipo de relación existente entre estas variables. También sirven para detectar posibles datos atípicos para la relación. Por ejemplo, para las variables peso y altura de un grupo de personas esperaríamos que el gráfico muestre que las personas más altas suelen pesar más; además podríamos ver que una persona que refiere que pesa 50 kg y mide 195 cm posiblemente se represente alejada del resto de puntos de la muestra, aunque estos dos valores de peso y altura no sean atípicos por separado. Vamos a ilustrar estas ideas con los datos de *pulso*.

Abrimos R Commander y cargamos el conjunto de datos *pulsoNuevo* (que está en el fichero *pulsoNuevo.RData*) con la opción

Datos > Cargar conjunto de datos…

En el panel de *Mensajes* nos dirá que el conjunto de datos *pulsoNuevo* tiene 106 filas y 11 columnas.

Este conjunto de datos se ha construido, en una práctica anterior, eliminando en el fichero Pulso.RData cuatro casos con errores.

Nos interesa averiguar si hay relación lineal (aproximada) entre las variables *Peso* y *Altura*, y si la hay, querremos obtener la recta que mejor la represente. Conviene empezar siempre dibujando los datos.

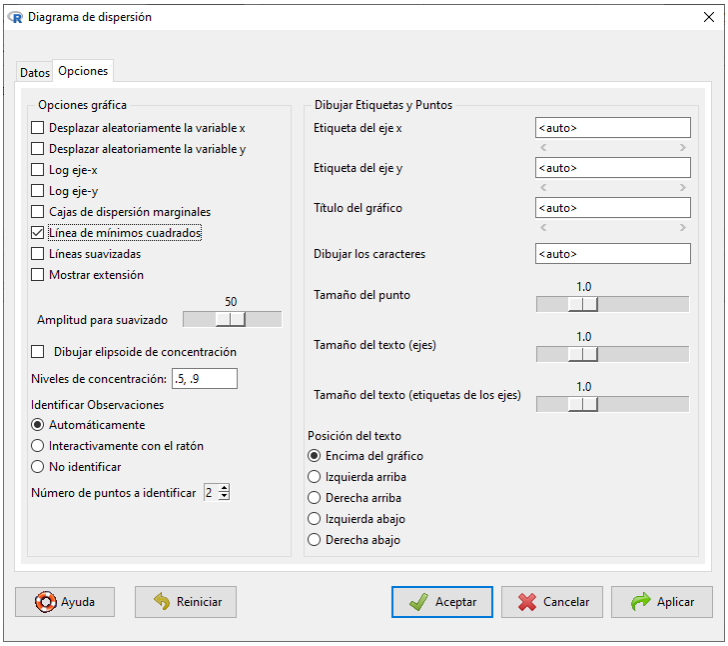
Para obtener el gráfico de dispersión de *Peso* en función de *Altura*, usamos el procedimiento R Commander

Gráficas > Diagrama de dispersión…

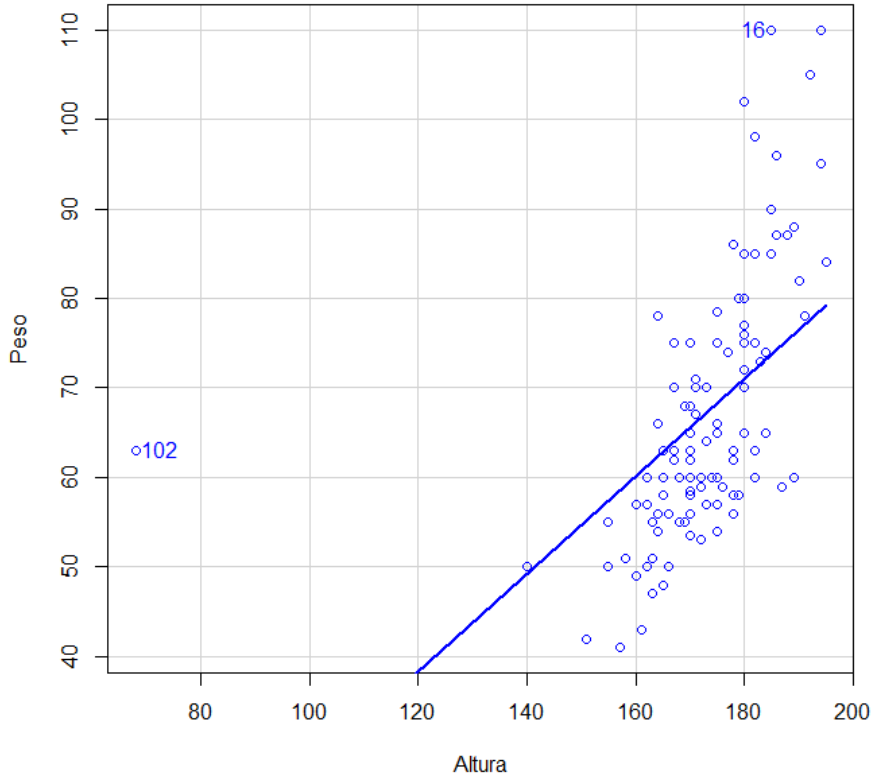
El cuadro de diálogo tiene dos pestañas. En la pestaña *Datos* seleccionamos *Altura* en la lista debajo de *variable x* y *Peso* en la lista debajo de *variable y*. En la pestaña *Opcione*s marcamos *Línea de mínimos cuadrados*, para que dibuje la recta que mejor se ajusta a la nube de puntos con este criterio (es la recta de regresión de Y sobre X).

Hay una opción que permite *Identificar Observaciones*. Si marcamos *Automáticamente*, aplica un algoritmo que identifica los puntos más alejados del centro de la nube de puntos. Le indicamos el número que puntos que deseamos que marque con su etiqueta en *Número de puntos a identificar*. Naturalmente, no todos esos datos identificados como atípicos son necesariamente datos erróneos.

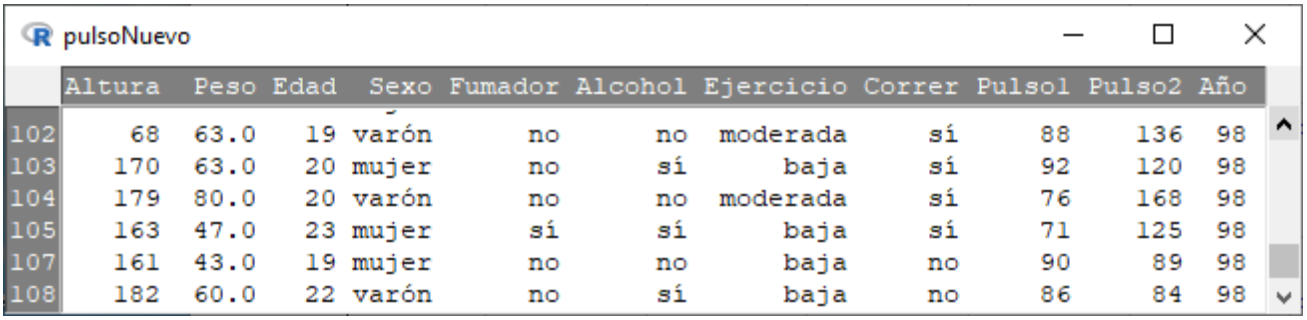
En este caso, marcaremos *Automáticamente* y pediremos que identifique dos puntos. La pestaña *Opciones* quedará:



El gráfico obtenido con estas opciones es el siguiente:

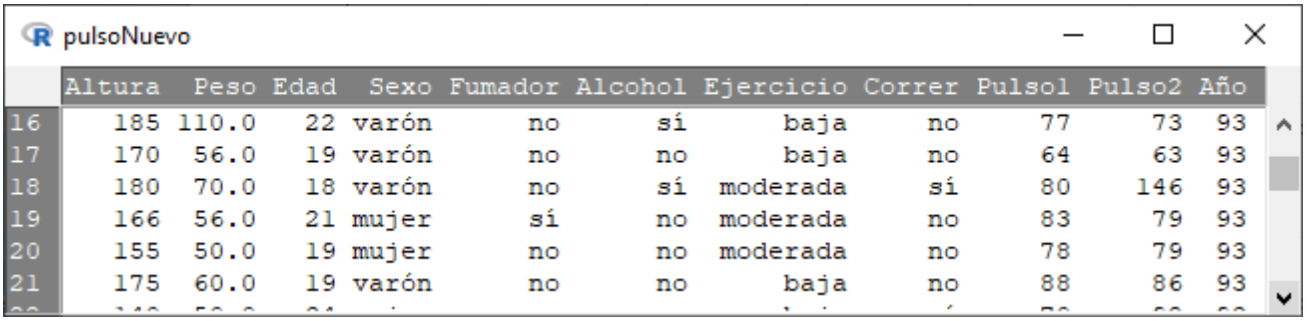


Los puntos aparecen bastante agrupados en la parte derecha del gráfico, mientras que en la parte izquierda se observa un caso que se aleja claramente del resto, el 102. El segundo caso identificado es el 16. Abriendo el *Visualizador* (puede verse más abajo el resultado) comprobamos que a la izquierda de los datos hay una columna en gris que contiene las etiquetas de cada caso. El caso 102 es un hombre con una altura de 68 centímetros y un peso de 63 kilos. El caso 16 es un hombre que mide 185 cm y pesa 110 kg.



pulsoNuevo

	Altura	Peso	Edad	Sexo	Fumador	Alcohol	Ejercicio	Correr	Pulso1	Pulso2	Año
102	68	63.0	19	varón	no	no	moderada	sí	88	136	98
103	170	63.0	20	mujer	no	sí	baja	sí	92	120	98
104	179	80.0	20	varón	no	no	moderada	sí	76	168	98
105	163	47.0	23	mujer	sí	sí	baja	sí	71	125	98
107	161	43.0	19	mujer	no	no	baja	no	90	89	98
108	182	60.0	22	varón	no	sí	baja	no	86	84	98



pulsoNuevo

	Altura	Peso	Edad	Sexo	Fumador	Alcohol	Ejercicio	Correr	Pulso1	Pulso2	Año
16	185	110.0	22	varón	no	sí	baja	no	77	73	93
17	170	56.0	19	varón	no	no	baja	no	64	63	93
18	180	70.0	18	varón	no	sí	moderada	sí	80	146	93
19	166	56.0	21	mujer	sí	no	moderada	no	83	79	93
20	155	50.0	19	mujer	no	no	moderada	no	78	79	93
21	175	60.0	19	varón	no	no	baja	no	88	86	93

Asumimos que se ha comprobado que el caso 102 es un error y decidimos eliminarlo del estudio, y que el caso 16 es correcto y, por lo tanto, lo mantendremos en nuestro estudio. El nuevo conjunto de datos, tras eliminar el caso 102, se llama *PulsoT4* y lo vamos a cargar en R Commander abriendo el fichero *PulsoT4.RData* desde la carpeta de la práctica. Con el *Visualizador* se puede comprobar que, efectivamente, en este fichero el caso 102 se ha eliminado.

COMENTARIO: Los pasos que tendríamos que haber seguido para eliminar este caso erróneo y obtener el fichero *PulsoT4.RData* son semejantes a los que se siguieron en el tema dedicado al análisis descriptivo univariante para obtener el fichero *pulsoNuevo.RData*. Primero se elimina el caso 102 con el *Editor*. El número de caso está en la primera columna, *rowname*. Puede observarse que el caso 102 ocupa la fila 98 del fichero. Tras eliminarlo, veremos en la ventana *Mensajes* que el fichero *pulsoNuevo* tiene ahora 105 filas. En esta versión de R Commander, cuando se elimina algún dato utilizando el editor se produce un cambio en el tipo de variable: las variables factor se convierten en variables carácter (character). Si queremos que R Commander las siga considerando variables de tipo factor, tendremos que cambiar su "tipo" utilizando el comando

Datos > Modificar variable del conjunto de datos activo > Convertir variables de carácter en factores …

Los cambios que hemos hecho en el fichero *pulsoNuevo* están en la memoria del ordenador, pero **no** en el fichero físico. Si queremos guardar este fichero lo tendremos que hacer explícitamente. Vamos a guardarlo con el nombre que nos interesa: *PulsoT4.* Crearemos ese conjunto de datos escribiendo en la ventana R Script

PulsoT4 = pulsoNuevo

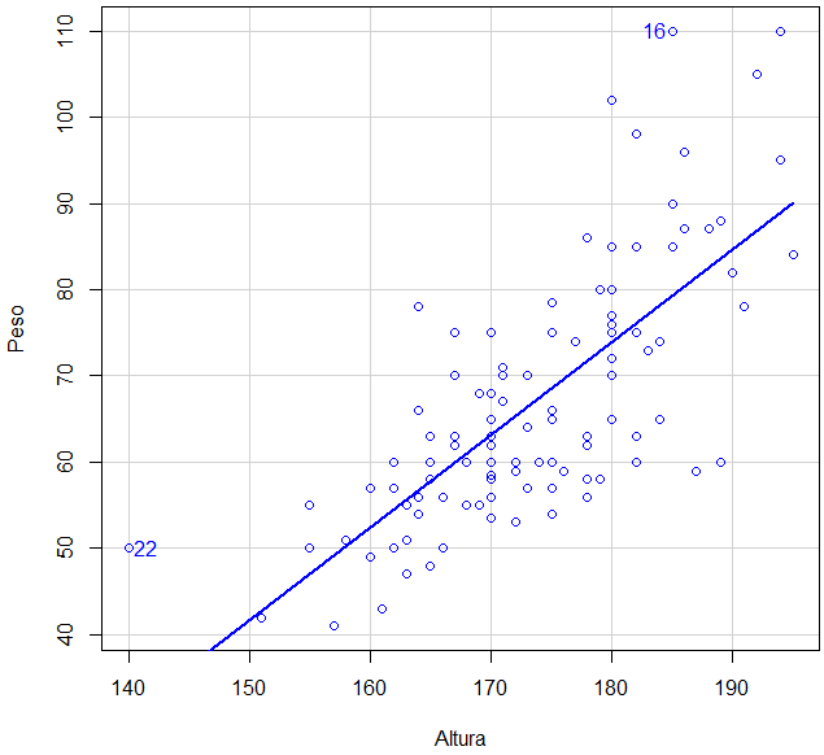
y ejecutando esa línea. Seleccionamos el conjunto de datos que acabamos de crear, *PulsoT4,* como conjunto de datos activo.

Para conservarlo en un dispositivo físico externo usamos la opción

Datos > Conjunto de datos activo > Guardar el conjunto de datos activo…

indicando la carpeta o dispositivo en la que queremos guardarlo físicamente y el nombre que le queremos dar, en este caso *PulsoT4.RData*.

Comprobamos que *PulsoT4* es el conjunto de datos activo y volvemos a obtener el gráfico de dispersión con *Altura* en el eje X y *Peso* en el Y. Pedimos en la pestaña *Opciones* que dibuje la *línea de mínimos cuadrados* y que identifique dos observaciones atípicas automáticamente.



Ahora se aprecia mejor que la relación entre *Peso* y *Altura* es aproximadamente lineal. Los dos casos identificados con el algoritmo son ahora el 16 y el 22. Decidimos mantener los dos datos en nuestro estudio.

Como resumen podríamos decir que **el gráfico de dispersión ha permitido identificar un error (el caso 102, un hombre con una altura de 68 centímetros y un peso de 63 kilos) que hemos eliminado del estudio. El grafico de dispersión de los otros 105 individuos de la muestra sugiere una clara relación lineal y creciente entre *Peso* y *Altura***.

Para ilustrar este comentario podríamos añadir en nuestro análisis el gráfico de dispersión, pero esta vez **sin identificar valores**, ya que hemos asumido que el resto de casos son correctos y los mantenemos en nuestro estudio.

1.2 Coeficiente de correlación

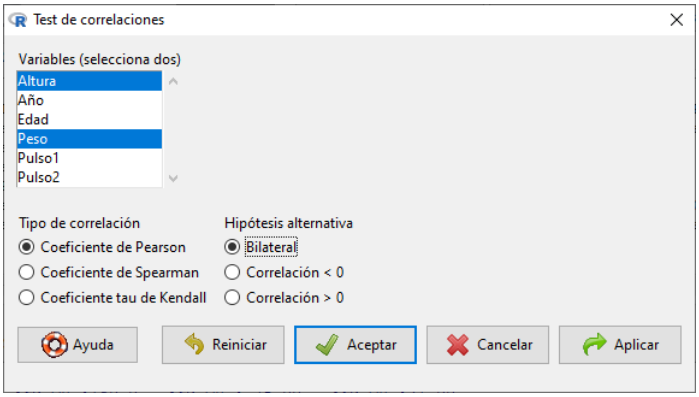
El diagrama de dispersión sugiere que hay una relación estadística lineal entre *Peso* y *Altura*. El coeficiente de correlación lineal de Pearson, *r*, permite cuantificar la intensidad de esta relación. El coeficiente *r* toma valores entre -1 y 1. Valores próximos a 1 en valor absoluto indican que la relación lineal es muy fuerte, mientras que valores próximos a 0 indican falta de relación lineal o que esta es irrelevante. Desde el punto de vista de la recta que mejor se ajusta a los datos, los valores positivos de *r* indican que la recta tiene pendiente positiva, y los negativos que la pendiente es negativa.

Convendría disponer de un criterio que nos diga cuando *r* es "suficientemente distinto" de 0 y podemos considerar que existe relación lineal entre las variables. El *test de correlación* proporciona un criterio para decidirlo. Podemos calcular el coeficiente de correlación y obtener los resultados del test de correlación con

Estadísticos > Resúmenes > Test de correlación…

En el cuadro de diálogo que aparece, seleccionamos las variables de interés, *Peso* y *Altura* (pulsando la tecla *Control* podemos seleccionar más de una variable).

Comprobamos que esté seleccionada la opción *Coeficiente de Pearson*, y en *Hipótesis alternativa*, la opción *Bilateral.*



Pulsando *Aceptar* se obtiene la salida

```
        Pearson's product-moment correlation

data:  Altura and Peso
t = 10.963, df = 103, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6309181 0.8113854
sample estimates:
      cor
0.7338433
```

El coeficiente de correlación de Pearson entre *Altura* y *Peso* es **r = 0.7338433**. En esta salida también se proporciona el p-valor *p* del test de correlación. En este caso nos dice que *p* es prácticamente nula (más pequeña que **2.2e-16**).

Para interpretar el test de correlación tenemos que considerar que el p-valor mide, en un rango de 0 a 1, la concordancia de los datos con la hipótesis nula (en este caso, que **no** hay correlación lineal entre las variables). Valores grandes de *p* indican que **no** hay correlación lineal y valores pequeños que **sí** la hay. Nosotros utilizaremos el p-valor como una regla empírica. Si *p* es menor que 0.05 diremos que hay correlación lineal entre las variables con un nivel de significación del 5%. Si es mayor, diremos que no hay correlación lineal con el nivel de significación habitual del 5%.

El comentario podría ser que **en esta muestra se ha obtenido un coeficiente de correlación de Pearson entre *Peso* y *Altura* de r = 0.734, que es positivo y relativamente grande; el p-valor del test de correlación es prácticamente cero. Esto indica que hay una relación estadística lineal creciente entre esas variables en la población muestreada, lo que concuerda con lo que hemos observado en la gráfica de dispersión. Por lo tanto, tiene sentido calcular la recta de regresión entre *Peso* y *Altura***.

NOTA 1.1: El coeficiente de correlación de Pearson está definido sólo para variables numéricas. Para medir la relación lineal entre variables ordinales, siempre que estas variables estén codificadas numéricamente, podemos utilizar el *coeficiente de correlación de rangos de Spearman*. El coeficiente de Spearman también se suele aplicar a variables numéricas cuando se sospecha que los datos tienen errores, ya que es una alternativa más robusta a errores que el coeficiente de Pearson. El coeficiente de Spearman se puede pedir desde la misma ventana con el comando

Estadísticos > Resúmenes > Test de correlación…

Si solicitamos el coeficiente de correlación de Spearman entre las variables *Altura* y *Peso*, obtendremos

```
        Spearman's rank correlation rho

data:  Altura and Peso
S = 50131, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7401456
```

1.3. Regresión lineal

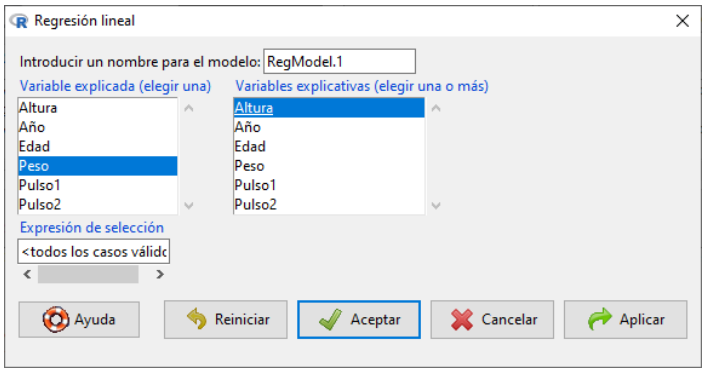
El diagrama de dispersión de las variables *Peso* y *Altura* sugiere una relación lineal y creciente. El coeficiente de correlación $r = 0.734$ es positivo y el p-valor del contraste de correlación es prácticamente nulo. Tiene sentido obtener la recta de regresión, que tendrá una expresión

$$[Peso] = a + b*Altura.$$

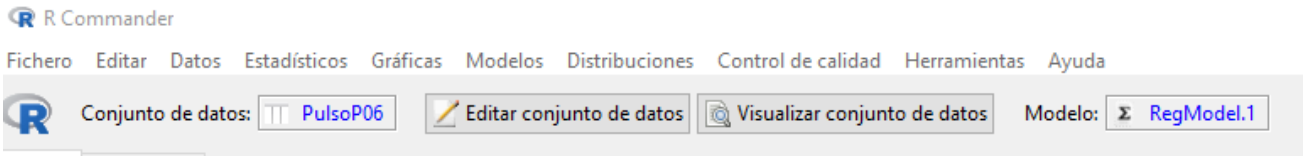
donde [*Peso*] indica el valor predicho para la variable *Peso* por la recta de regresión. Para calcular los coeficientes constantes ***a*** y ***b*** de esa recta de regresión usaremos el procedimiento

Estadísticos > Ajuste de modelos > Regresión lineal…

En el cuadro de diálogo se selecciona como variable explicada *Peso* y como explicativa *Altura*. El cuadro también proporciona un nombre por defecto (*RegModel.1*) al modelo que se va a construir. Se puede modificar ese nombre, si conviene. Esta posibilidad es útil cuando interesa comparar distintos modelos.



Tras pulsar *Aceptar*, observa que en la casilla *Modelo* ha escrito el nombre *RegModel.1* (o el nombre que le hayamos dado, si lo hemos modificado). Este es el modelo activo que aparecerá en la ventana *Modelo*:



La salida de este procedimiento es

```
Call:
lm(formula = Peso ~ Altura, data = PulsoP06)
Residuals:
     Min       1Q   Median       3Q      Max
-23.6543  -7.1549  -0.8055   5.6952  30.6456

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -119.51465   17.03757  -7.015  2.5e-10 ***
Altura         1.07497    0.09805  10.963  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.05 on 103 degrees of freedom
Multiple R-squared:  0.5385,     Adjusted R-squared:  0.534
F-statistic: 120.2 on 1 and 103 DF,  p-value: < 2.2e-16
```

Esta salida proporciona mucha información. Hemos remarcado la más relevante. El término independiente ***a*** es el que se muestra a continuación de **(Intercept) : -119.51465**. El coeficiente ***b*** es el que multiplica a la variable *Altura* y es la pendiente de la recta de regresión. Es el que se muestra a continuación de **Altura : 1.07497**.

Se puede indicar que **la ecuación de la recta de regresión es,** redondeando a tres decimales

*[Peso] = -119.515 + 1.075*Altura.*

Este modelo predice que los pesos de dos personas de ese colectivo diferirán en 1.075 kilogramos por cada centímetro de diferencia en su altura.

El coeficiente de determinación, R^2 se muestra a continuación de **Multiple R-squared: 0.5385** y nos indica que **la recta de regresión sobre *Altura* explica el 54% de la variabilidad de la variable *Peso***.

La recta de regresión se puede utilizar para realizar predicciones aproximadas. En nuestro ejemplo, si queremos predecir el peso de una persona cuya altura es ***173*** centímetros, el modelo indica que esa persona pesará alrededor de los ***66.5*** kilogramos. Podemos comprobarlo utilizando R Commander como calculadora:

```
> -119.51465+1.07497*173
[1] 66.45516
```

Esta predicción está sujeta a error ya que con esta recta de regresión la variable *Altura* sólo explica el 54% de la variabilidad de *Peso* y queda un 46% sin explicar, que dependerá de otros factores. El peso observado para una persona que mide 173 cm podría ser bastante distinto de los 66.5 kg que indica predicción.

NOTA 1.2: En un modelo lineal con una variable explicativa el coeficiente de determinación R^2 es el cuadrado del coeficiente de correlación de Pearson *r*. Lo podemos comprobar en este caso utilizando R Commander como calculadora:

```
> 0.7338433^2
[1] 0.538526
```

1.4 Gráfica de residuos

El coeficiente de determinación R^2 entre X e Y se puede interpretar como la proporción de variabilidad de Y que explica la recta de regresión sobre X. Pero, aunque sea alto, la recta de regresión podría no ser un modelo adecuado de la relación entre X e Y. Esto ocurre, por ejemplo, cuando existe un modelo no lineal que explica mejor la relación entre las dos variables, o cuando hay alguna característica sospechosa, como la presencia de algún dato atípico y muy influyente, o también cuando hay cambios sistemáticos de variabilidad en torno a la recta. El *diagrama de dispersión* de los datos originales puede dar indicaciones sobre alguno de estos problemas, pero hay otro *diagrama de dispersión* en el que se aprecian mejor estos problemas: la *gráfica de residuos*.

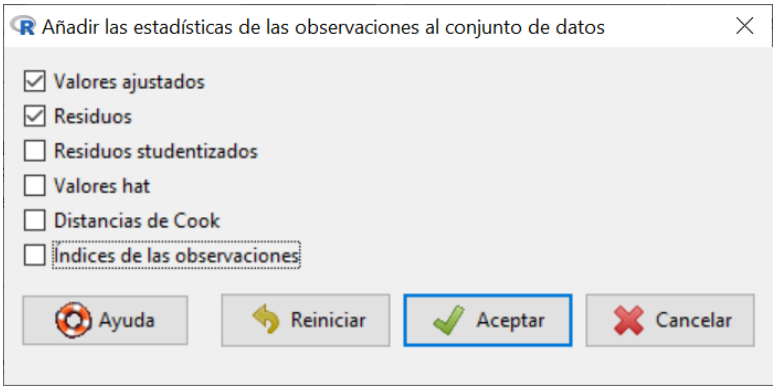
Los *residuos* son la diferencia entre los valores de ***y*** observados y los valores ***[y]*** que predice la recta de regresión para cada valor de ***x***. La *gráfica de residuos* representa en el eje X los *valores ajustados* ***[y]***, y en el eje Y los correspondientes *residuos* ***e = y - [y].***

Si el modelo es bueno, esta gráfica no debería mostrar ningún patrón. Deberíamos ver una nube de puntos repartida aleatoriamente en una franja horizontal en torno a Y = 0.

Para poder obtener el *diagrama de dispersión* de los *residuos* frente a los *datos ajustados*, debemos tener estos valores como variables en nuestro conjunto de datos. El procedimiento

Modelos > Añadir las estadísticas de las observaciones a los datos…

permite añadir *Valores ajustados* y *Residuos*, que son los valores ***[y]*** y ***e = y - [y]***, al conjunto de datos activo. Dejaremos seleccionadas solamente estas dos estadísticas en la ventana.

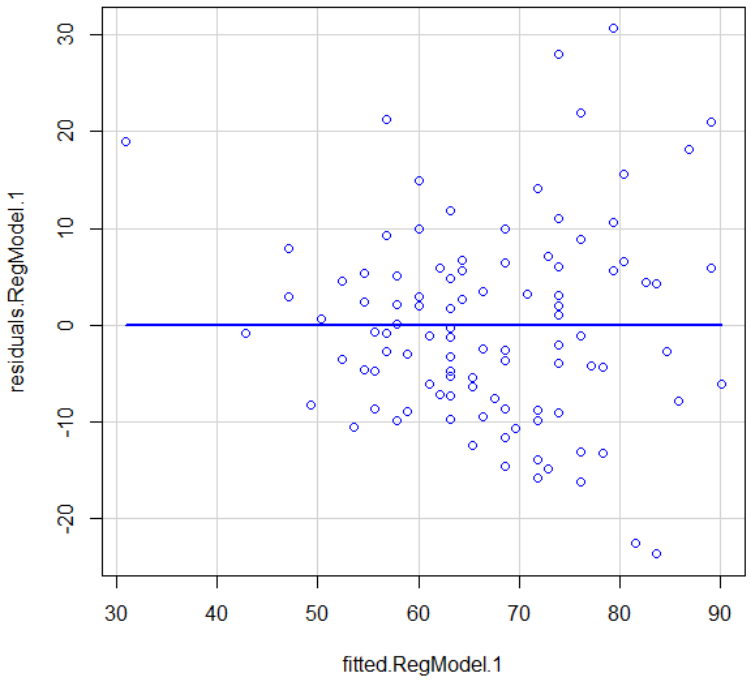


Si pulsamos *Aceptar*, al visualizar el archivo de datos veremos que se han creado dos nuevas variables, una con los **residuos**: *residuals.RegModel.1* y otra con los **valores ajustados**: *fitted.RegModel.1*.

Finalmente, obtendremos la *gráfica de residuos* con el *Diagrama de dispersión*

Gráficas > Diagrama de dispersión…

poniendo *fitted.RegModel.1* en el eje X y *residuals.RegModel.1* en el eje Y. También le pedimos que dibuje la línea de mínimos cuadrados. Obtenemos



En este caso **la gráfica de residuos muestra una debilidad de este modelo lineal: los residuos del modelo (esto es, los errores de predicción) muestran una cierta estructura de "embudo". Esto indica que los errores al predecir *Peso* con la recta de regresión en función de *Altura* tienden a ser mayores cuanto mayores son estos valores predichos**.

NOTA 1.3: La recta de mínimos cuadrados de los residuos frente a los valores ajustados, en la gráfica de residuos, siempre va a ser Y = 0. La recta de regresión ya ha "extraído" toda la información de X que sirve para predecir Y con un modelo lineal y no deja ninguna información en los residuos.

La debilidad del modelo puede ser debida a que no hemos considerado algunos factores que influyen en la relación. En este caso particular, podemos sospechar que la relación entre *Peso* y *Altura* no será la misma para hombres que para mujeres. En la sección 2 de esta práctica veremos cómo analizar la relación entre las dos variables en cada uno de los subgrupos de interés que, en este caso, son los hombres y las mujeres de la muestra.

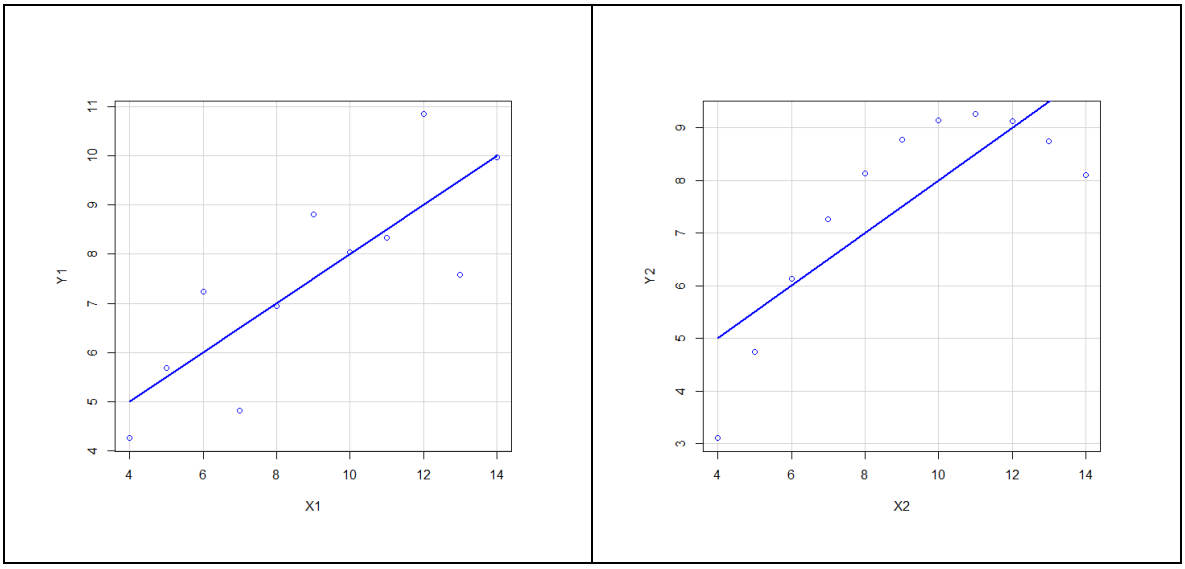
Antes, veremos, como ejercicio, otro ejemplo de cómo utilizar la *gráfica de residuos* para comprobar si un modelo es adecuado o no, cuando hay un modelo no lineal que proporcionaría un mejor ajuste a los datos.

Ejercicio 1.1: Carga el fichero de datos *Anscombe.RData*. Construye la rectas de regresión para las variables (X1,Y1), considerando Y1 como variable explicada, y llama al modelo *Anscombe1*. Haz lo mismo para las variables (X2,Y2), considerando Y2 como variable explicada, y llama al modelo *Anscombe2*. Puedes comprobar que los modelos obtenidos son prácticamente iguales

```
[Y1] = 3.0001 + 0.5001*X1
[Y2] = 3.0010 + 0.5000*X2
```

El porcentaje de variabilidad explicada también es prácticamente el mismo (66.65% y 66.62%). Con estos datos, podríamos concluir que los dos modelos son igual de buenos para describir los dos conjuntos de datos. Pero antes de llegar a una conclusión hay que examinar **siempre** los gráficos de dispersión.

Ejercicio 1.2: Comprueba que *Anscombe* es el conjunto de datos activo. Dibuja el diagrama de dispersión para los dos pares de variables, (X1,Y1) y (X2,Y2), y pide que dibuje la recta de mínimos cuadrados. Comprueba que se obtiene

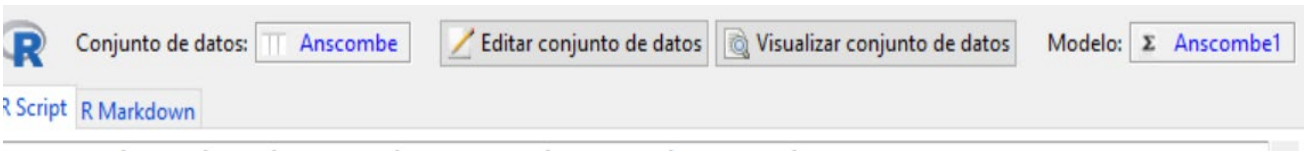


En el primer caso, la recta de regresión puede representar adecuadamente la relación estadística entre Y1 y X1. Los puntos parecen repartirse aleatoriamente en torno a esa recta. En el segundo caso, parece que una parábola sería más adecuada que una recta para describir la relación entre Y2 y X2. Esta impresión se confirma si obtenemos las *gráficas de residuos* de los dos modelos.

Ejercicio 1.3: Obtén las gráficas de residuos de estos dos modelos.

Comprobamos primero que el *Conjunto de datos* activo sea *Anscombe*.

Para obtener las estadísticas del primer modelo tenemos que comprobar que el *Modelo* activo sea *Anscombe1*.



Usamos el procedimiento

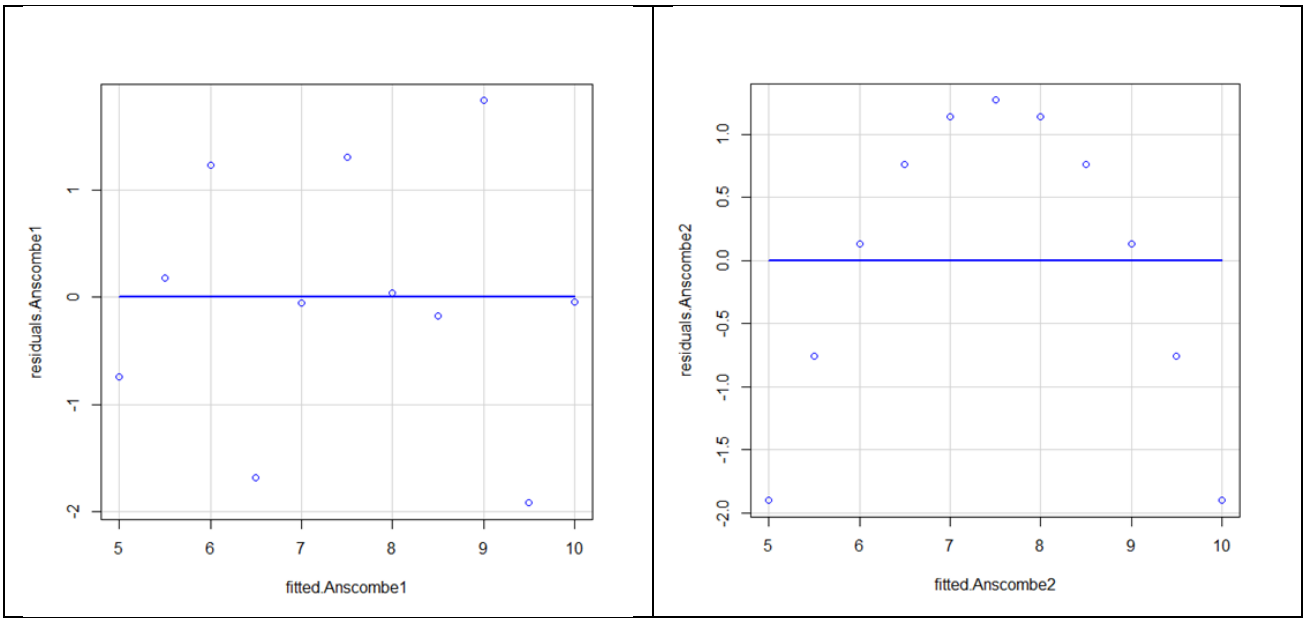
Modelos > Añadir las estadísticas de las observaciones a los datos…

para añadir las estadísticas *Valores ajustados* y *Residuos* del modelo *Anscombe1* al conjunto de datos *Anscombe*. Después obtendremos la gráfica de residuos, como hemos hecho en el ejemplo anterior, utilizando

Gráficas > Diagrama de dispersión…

poniendo *fitted.Anscombe1* en el eje X y *residuals.Anscombe1* en el eje Y. También pediremos que dibuje la línea de mínimos cuadrados. Repetiremos el mismo procedimiento, pero ahora para el modelo *Anscombe2*.

Las gráficas de residuos de los dos conjuntos tienen este aspecto:



En la primera gráfica no observamos ninguna estructura aparente. Los residuos se distribuyen aleatoriamente en torno al eje Y = 0.

En la segunda gráfica los residuos muestran una estructura (en este caso parabólica) que indica que existe un modelo mejor que el modelo lineal para explicar la relación entre Y2 y X2.

2. Relación entre dos variables numéricas, por grupos

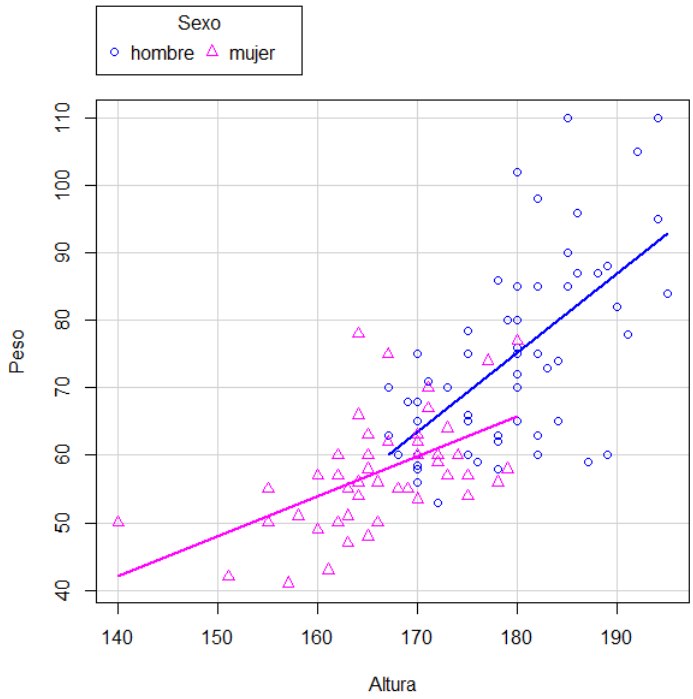
En muchas ocasiones interesa estudiar si la relación entre dos variables numéricas es la misma para los subgrupos de población definidos por una tercera variable de tipo categórico. Por ejemplo, puede interesar comprobar si la relación entre la altura y el peso es similar entre los hombres y las mujeres.

Para hacer una gráfica de dispersión distinguiendo entre hombres y mujeres usaremos el mismo procedimiento

Gráficas > Diagrama de dispersión…

pero ahora, en la pestaña *Datos*, tras seleccionar *Altura* como variable X, y *Peso* como variable Y, pulsaremos el botón *Gráfica por grupos…* En la ventana que se abre, seleccionamos la variable *Sexo*, dejamos marcada la casilla *Dibujar líneas por grupos* y pulsamos *Aceptar*.

En la pestaña *Opciones* seleccionamos *Línea de mínimos cuadrados*. Tras *Aceptar*, obtenemos el gráfico:

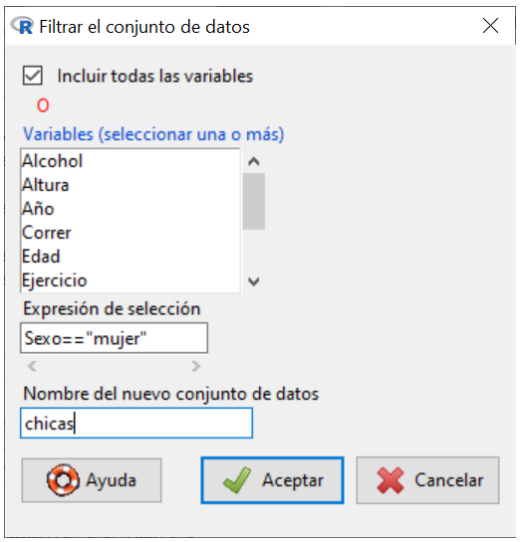


Se observa una relación aproximadamente lineal y creciente tanto en los datos de hombres (círculos azules) como en los datos de mujeres (triángulos rojos), pero la recta de regresión para los datos de las mujeres tiene una pendiente menor que la recta de regresión para los datos de los hombres. **El gráfico indica que no parece adecuado utilizar la misma recta de regresión para predecir en los dos grupos.** Por lo tanto, tendremos que calcular el coeficiente de correlación y obtener una recta de regresión por separado para el subgrupo de los hombres y para el subgrupo de las mujeres.

2.1 Coeficiente de correlación y regresión lineal, por grupos

El procedimiento que hemos utilizado para calcular el coeficiente de correlación de Pearson y el test de correlación no permite un cálculo por grupos. Si queremos realizar el test de correlación, por separado, para los datos de hombres y de mujeres, tendremos que construir un conjunto de datos de hombres y otro de mujeres. Vamos a crear, en primer lugar, un conjunto de datos solo con los datos de mujeres. Esto se puede hacer utilizando la opción de *Filtrar,* que se encuentra en

Datos > Conjunto de datos activo > Filtrar el conjunto de datos activo…



En el cuadro de diálogo hemos escrito la expresión para seleccionar sólo a las mujeres: *Sexo=="mujer"*. También hemos escrito el nombre que le queremos dar al nuevo conjunto de datos (en este ejemplo, *chicas*). Tras *Aceptar*, el *Conjunto de datos* activo será *chicas*, y en el panel de mensajes tendremos

```
NOTA: El conjunto de datos chicas tiene 48 filas y 11 columnas.
```

Tras comprobar que *chicas* es el conjunto de datos activo, podemos calcular el coeficiente de correlación para este grupo con el procedimiento habitual

Estadísticos > Resúmenes >Test de correlación…

Seleccionamos (usando la tecla *Control*) *Altura* y *Peso*. Tras aceptar, en la ventana de salida obtendremos

```
        Pearson's product-moment correlation

data:  Altura and Peso
t = 4.3023, df = 46, p-value = 0.00008734
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2966732 0.7115005
sample estimates:
      cor
0.5356548
```

La correlación es **r = 0.5356548** y el p-valor es **p-value = 0.00008734** < 0.05. Rechazamos que el coeficiente de correlación lineal sea 0 en la población muestreada. Por lo tanto tiene sentido obtener la recta de regresión. Tras comprobar que *chicas* sigue siendo el conjunto de datos activo, usamos el procedimiento habitual para obtener la recta de regresión

Estadísticos > Ajuste de modelos > Regresión lineal…

Seleccionamos *Peso* como variable explicada y *Altura* como variable explicativa. Vamos a llamar a este modelo *chicasRR.1*. La salida que produce es

```
Call:
lm(formula = Peso ~ Altura, data = chicas)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5162  -4.9696  -0.8297   3.3394  21.7040

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) -40.9992    22.9398  -1.787    0.0805 .
Altura        0.5933     0.1379   4.302 0.0000873 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.154 on 46 degrees of freedom
Multiple R-squared:  0.2869,     Adjusted R-squared:  0.2714
F-statistic: 18.51 on 1 and 46 DF,  p-value: 0.00008734
```

Si consideramos sólo los datos de las mujeres, el coeficiente ***b*** es **0.5933** (por cada centímetro de altura el modelo predice un incremento de 0.5933 kg) y el término independiente ***a*** es **-40.9992**. La ecuación de la recta de regresión es

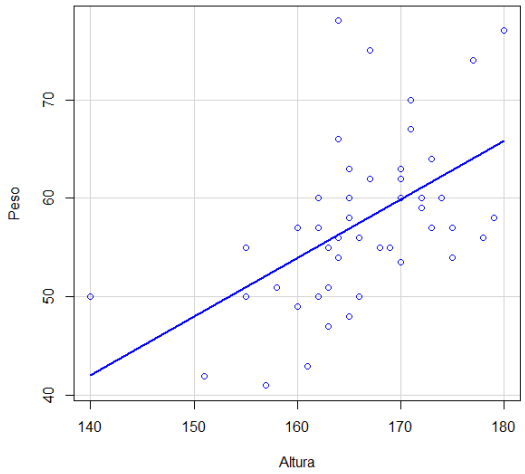
*[Peso] = **-40.9992** + **0.5933** *Altura.*

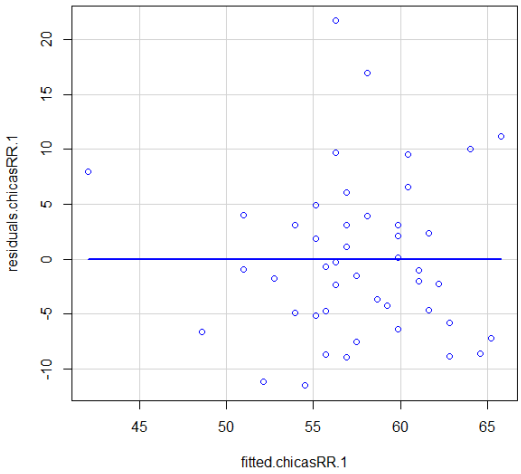
El coeficiente de determinación es $\mathbf{R^2 = 0.2869}$. Este modelo explica un 29% de la variabilidad de *Peso*.

Tras comprobar que el *Conjunto de datos* sigue siendo *chicas* y que el *Modelo* es *chicasRR.1*, le pedimos que guarde los valores ajustados y los residuos con la opción

Modelos > Añadir las estadísticas de las observaciones a los datos…

Los gráficos de dispersión (*Peso* frente a *Altura*) y de residuos (residuos frente a valores ajustados) que podemos obtener para este conjunto de datos son:





Si queremos realizar este análisis para los datos de los hombres tendremos que repetir estos procedimientos, pero con la precaución de **volver al fichero original** *PulsoT4* (que es el que contiene todos los casos, hombres y mujeres) ya que el fichero *chicas* sólo contiene datos de mujeres. Después filtramos con la condición *Sexo=="hombre"* e indicamos el nombre del nuevo conjunto de datos, *chicos.* Al aceptar pondrá en el panel de Mensajes

```
NOTA: El conjunto de datos chicos tiene 57 filas y 11 columnas.
```

El contraste de correlación, para el conjunto de datos *chicos*, proporciona esta salida:

```
        Pearson's product-moment correlation

data:  Altura and Peso
t = 5.7287, df = 55, p-value = 0.0000004395
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4172016 0.7520832
sample estimates:
      cor
0.6113138
```

Con la opción *Regresión Lineal*, para el conjunto *chicos*, obtenemos la salida:

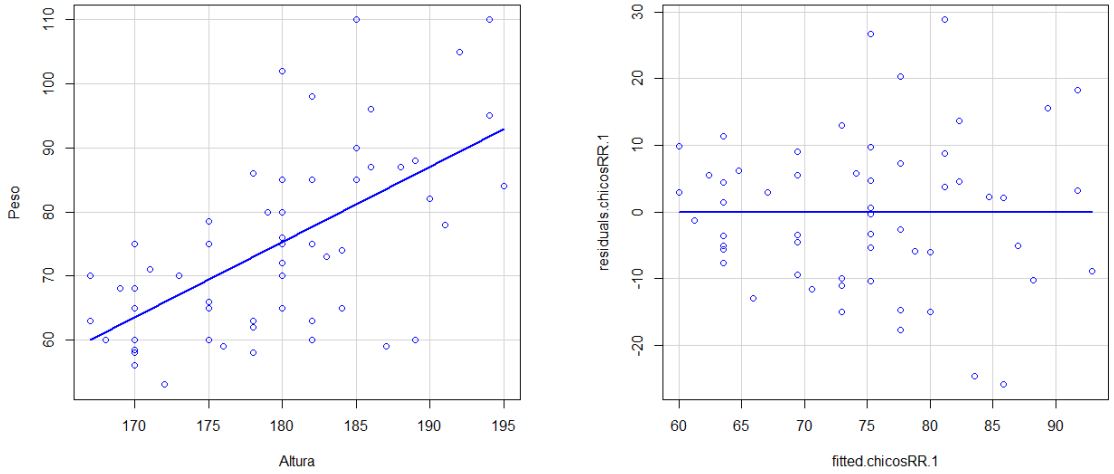
```
Call:
lm(formula = Peso ~ Altura, data = chicos)

Residuals:
     Min       1Q   Median       3Q      Max
-25.8691  -7.5826   0.6876   6.2444  28.8228

Coefficients:
             Estimate Std. Error t value   Pr(>|t|)
(Intercept) -135.8230    36.8079  -3.690   0.000516 ***
Altura         1.1730     0.2048   5.729 0.00000044 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.43 on 55 degrees of freedom
Multiple R-squared:  0.3737,     Adjusted R-squared:  0.3623
F-statistic: 32.82 on 1 and 55 DF,  p-value: 0.0000004395
```

Si consideramos sólo los datos de los hombres, el coeficiente ***b*** es **1.1730** (por cada centímetro de altura el modelo predice un incremento de 1.1730 kg) y el término independiente ***a*** es **-135.8230**. La ecuación de la recta de regresión es

*[Peso]= **-135.8230** + **1.1730** *Altura.*

El coeficiente de determinación es $\mathbf{R^2 = 0.3737}$. Por lo tanto, este modelo explica un 38% de la variabilidad de *Peso*.

El diagrama de dispersión en torno a la recta de regresión y la gráfica de residuos para *chicos* son:



Un comentario que relacione los dos modelos podría decir que **la relación entre *Peso* y *Altura* es aproximadamente lineal y creciente, tanto en el grupo de los hombres (con un coeficiente de correlación r = 0.611) como en el de las mujeres (con un coeficiente de correlación r = 0.536), los dos significativamente distintos de 0, con el nivel habitual del 5%. La recta de regresión para los hombres es**

*[Peso] = **-135.8230** + **1.1730** *Altura.*

que explica un 38% de la variabilidad de *Peso*. La recta de regresión para las mujeres es

*[Peso] = **-40.9992** + **0.5933** *Altura.*

y explica un 29% de la variabilidad de *Peso*. El modelo predice un incremento de pesos de 1.1730 kg para los hombres y de 0.5933 kg para las mujeres por cada centímetro de aumento en altura.

NOTA 2.1: Si no necesitamos los test de correlación ni las gráficas de residuos y sólo necesitamos las rectas de regresión, no es necesario filtrar. El procedimiento

Estadísticos > Ajuste de modelos > Regresión lineal…

tiene la opción *Expresión de selección*. Si se aplicara al conjunto de datos original *PulsoT4*, con la expresión de selección *Sexo=="mujer"*, se obtiene el modelo lineal solo para las mujeres. Con la expresión de selección *Sexo=="hombre"*, se obtiene el modelo lineal solo con los casos de los hombres de esta muestra.

3. Ejercicios

1. Un investigador piensa que cuanto mayor sea el nivel del colesterol, mayor será el nivel de triglicéridos en sangre. En el archivo *colesteroltriglicéridos.RData* se encuentran los niveles en sangre de colesterol y triglicéridos (en mg/dl) para personas sin aparentes enfermedades coronarias y para personas con estrechamiento de las arterias coronarias.

   a) A través del diagrama de dispersión, identifica el dato más atípico y elimínalo del conjunto de datos.

   b) Tras eliminar del estudio ese dato y usando los diagramas de dispersión y el coeficiente de correlación indica si los datos apoyan (o no) lo que piensa el investigador si consideramos todos los datos. ¿Tiene sentido calcular la recta de regresión de triglicéridos en función de colesterol?

   c) Obtén la recta de regresión de triglicéridos en función de colesterol. Indica qué porcentaje de variabilidad explica esta recta de regresión.

   d) Considerando sólo las personas sin aparentes enfermedades coronarias, indica si los datos confirman lo que piensa el investigador también para este grupo, obtén la recta de regresión e indica qué porcentaje de variabilidad explicaría.

2. En un ensayo clínico realizado para estudiar el posible efecto hipotensor de un fármaco, se evalúa la tensión arterial diastólica (TAD) en condiciones basales y tras 4 semanas de tratamiento, en un total de 14 pacientes hipertensos. Los datos del ensayo se encuentran en el archivo *TAD.RData.* ¿Existe relación lineal entre la TAD basal y la que se observa tras el tratamiento? ¿Cuál es el valor esperado de TAD tras el tratamiento, en un paciente que presentó una TAD basal de 103 mm de Hg?

3. Se han obtenido importantes ventajas del hecho de enseñar a los diabéticos a medir su propia glucosa en sangre. Se está investigando una nueva técnica, menos costosa que el procedimiento habitual, para que el diabético pueda medir con facilidad su propio nivel de glucosa en sangre. Si se puede probar que esta nueva técnica mide con precisión el nivel de glucosa en sangre, se generalizará su uso. El archivo *glucosa.RData* contiene los niveles de glucosa en sangre de diferentes pacientes diabéticos, obtenidos usando la nueva técnica y usando el método tradicional. Los datos están en milimoles por litro. Analiza la relación entre estas variables usando el diagrama de dispersión, el coeficiente de correlación y la recta de regresión. ¿Sería razonable generalizar el nuevo método? ¿Qué previsión darías para el nivel de glucosa en sangre medido en laboratorio, si el nivel medido con la nueva técnica fuese de 4 mmol/litro?

4. Se quiere analizar si existe alguna relación entre la tensión sistólica y la edad en personas del sexo femenino. Para ello, se seleccionaron aleatoriamente 36 mujeres de una población homogénea y se midió su edad, en años, y su presión sistólica, en mmHg. Los datos se encuentran en el archivo *presionsistolica.RData*. 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 la presión sistólica de una nueva mujer cuya edad fuese de 35 años?

5. El fichero *Anscombe.RData* contiene cuatro pares de variables, (X1,Y1), (X2,Y2), (X3,Y3) y (X4,Y4). ¿Tiene sentido obtener la recta de regresión de Y1 sobre X1? Para contestar, obtén el gráfico de dispersión, el coeficiente de correlación, la recta de regresión y el gráfico de residuos y comenta lo que obtienes. Repite el análisis para los otros tres pares de variables. Estos datos también están en la pestaña Anscombe del fichero Excel *EACS-tema-2.xlsx*.

6. La pestaña Palmos en el fichero Excel *EACS-tema-2.xlsx* contiene las medidas del palmo derecho (PD) y del palmo izquierdo (PI) de 40 estudiantes. Importa estos datos a R Commander. Calcula la correlación lineal entre PD y PI. ¿Es significativamente distinta de 0? ¿Es razonable obtener una recta de regresión para predecir PD en función del PI? ¿Qué porcentaje de la variabilidad de PD explicaría esta recta de regresión? ¿Qué valor de PD predice el modelo para un PI = 202? Obtén el diagrama de dispersión de los puntos en torno a la recta de regresión y el gráfico de residuos, y discute si la recta de regresión que has obtenido es un modelo adecuado para predecir PD en función de PI.

7. La pestaña BT en el fichero Excel *EACS-tema-2.xlsx* contiene las medidas de los pliegues del bíceps y del tríceps de 205 estudiantes. Lee estos datos desde R Commander y analiza si hay una relación lineal entre estas dos variables utilizando las herramientas gráficas y estadísticas que hemos utilizado en esta práctica.

8. Estudia gráfica y numéricamente la relación entre las pulsaciones antes del experimento (*Pulso1*) y después del experimento (*Pulso2*) en el fichero *PulsoT4.RData*, siguiendo los siguientes pasos:

8.1 Obtén la gráfica de dispersión de *Pulso2* frente a *Pulso1*, dibujando la recta de regresión y pidiéndole que identifique automáticamente tres atípicos. Copia el gráfico y coméntalo.

8.2 Obtén la misma gráfica, pero ahora distinguiendo entre los que han corrido y los que no (variable *Correr*), pidiéndole que dibuje la recta de regresión para cada grupo y que identifique automáticamente tres atípicos. Copia el gráfico y coméntalo.

8.3 Analiza ahora los datos de los que no han corrido, filtrando con los valores de la variable Correr que toman valor no. Para el análisis vas a seguir los siguientes pasos.

a) Obtén la gráfica de dispersión de *Pulso2* frente a *Pulso1*, dibujando la recta de regresión, pero sin pedir que identifique. Copia el gráfico y coméntalo.

b) Calcula el coeficiente de correlación y decide, con el test de correlación, si ese coeficiente indica, efectivamente, que hay una relación lineal entre las variables en la población muestreada.

c) Obtén la recta de regresión, indica qué porcentaje de la variabilidad de Pulso2 se puede explicar con el valor de Pulso1 e indica cuál sería el valor ajustado por la recta de regresión para unas pulsaciones iniciales de 60 lpm.

d) Analiza si este modelo es adecuado utilizando la gráfica de residuos.

8.4 Repite lo mismo para los que sí han corrido.

8.5 Compara el comportamiento de la recta de regresión en los dos grupos.