Práctica 3. Regresión lineal por mínimos cuadrados.

Universidad de Alcalá. Curso 2023-24.

Estadística (650008). Grado en biología sanitaria.

Actualizado: 2023-10-05

Presentación del problema.

Los datos de **este fichero** se refieren a una muestra de gatos domésticos. La tabla contiene las variables

- `Sex`: M/F (male/female).
- `Bwt`: peso del cuerpo en kg.
- `Hwt`: peso del corazón en g.

Puedes leer más sobre este conjunto de datos en este enlace:

https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/cats.html

Vamos a utilizar esos datos para explorar el posible grado de asociación entre dos variables relacionadas con la anatomía de los gatos: su peso corporal y el peso del corazón del gato.

Descarga y lectura de los datos.

- Crea un script de R y guardalo en un carpeta.
- Puedes leer el fichero usando el botón `Import dataset` que hay en el marco superior derecho de RStudio.
- Tampién puedes usar, claro, la función `read.table()` que se explicó en el vídeo de lectura de datos de la práctica 2. En ese caso, no olvides
  - Descarga el fichero en la misma carpeta que el script.
  - Fija la carpeta de trabajo en dicha carpeta.
  - Explora el fichero con un editor de texto como el *Bloc de Notas*, para ver su estructura.

```
cats = read.table(file = "Practica03-cats.csv", header = TRUE, sep = ";")
```

Ejercicio 1 Comprueba que la lectura ha sido correcta: visualiza la cabecera de la tabla, obtén un resumen de las variables y calcula las dimensiones de la tabla

Para comprobar que la lectura ha sido correcta vamos a ver las primeras lineas de la tabla

```
head(cats)
```

```
##   Sex Bwt Hwt
## 1   F 2.0 7.0
## 2   F 2.0 7.4
## 3   F 2.0 9.5
## 4   F 2.1 7.2
## 5   F 2.1 7.3
## 6   F 2.1 7.6
```

las dimensiones de la tabla y usamos summary para obtener una exploración inicial de las variables:

```
dim(cats)
```

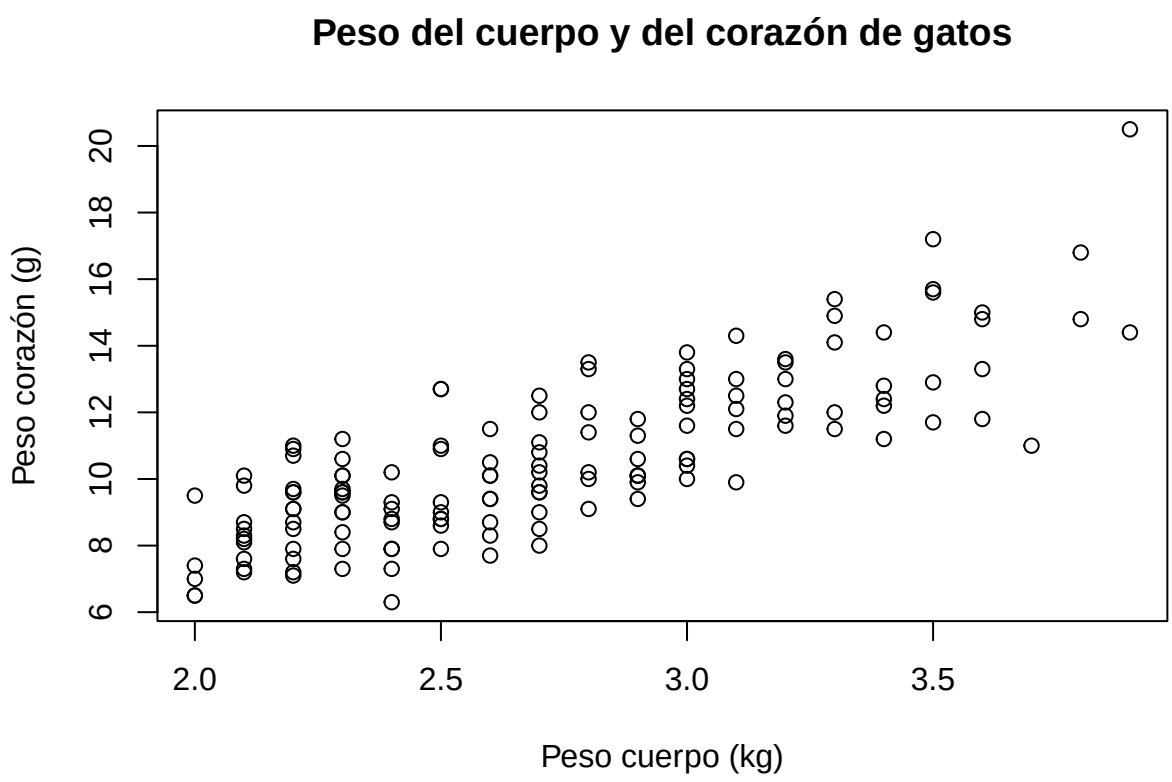
```
## [1] 144   3
```

```
summary(cats)
```

```
##      Sex                 Bwt             Hwt       
##  Length:144         Min.   :2.000   Min.   : 6.30  
##  Class :character   1st Qu.:2.300   1st Qu.: 8.95  
##  Mode  :character   Median :2.700   Median :10.10  
##                     Mean   :2.724   Mean   :10.63  
##                     3rd Qu.:3.025   3rd Qu.:12.12  
##                     Max.   :3.900   Max.   :20.50
```

Ejercicio 2 Visualiza la nube de puntos para determinar de forma visual si tiene sentido utilizar un modelo lineal para relacionar las variables `Bwt` y `Hwt`.

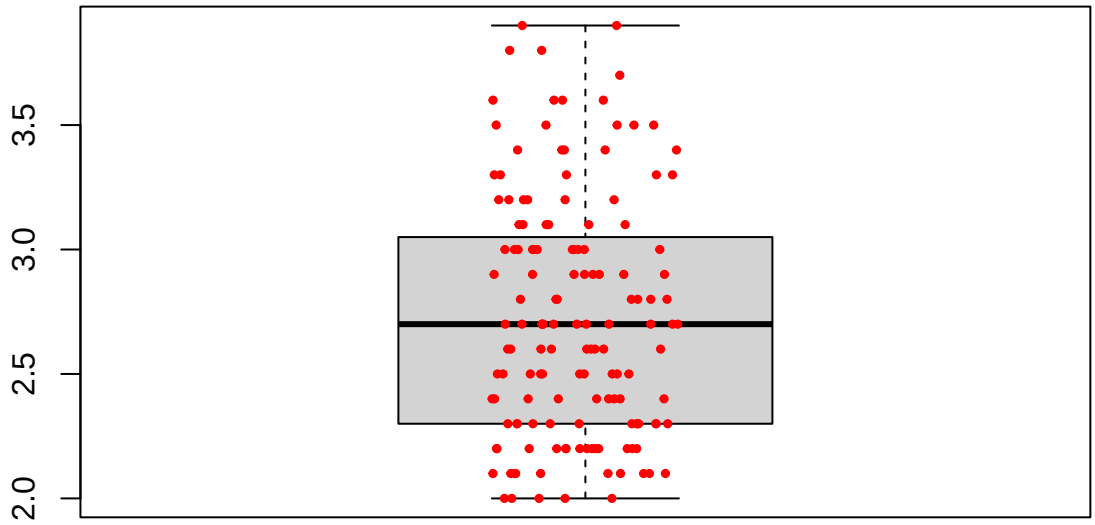
```
plot(cats$Bwt, cats$Hwt, main = "Peso del cuerpo y del corazón de gatos",
     xlab = "Peso cuerpo (kg)", ylab = "Peso corazón (g)"
     )
```



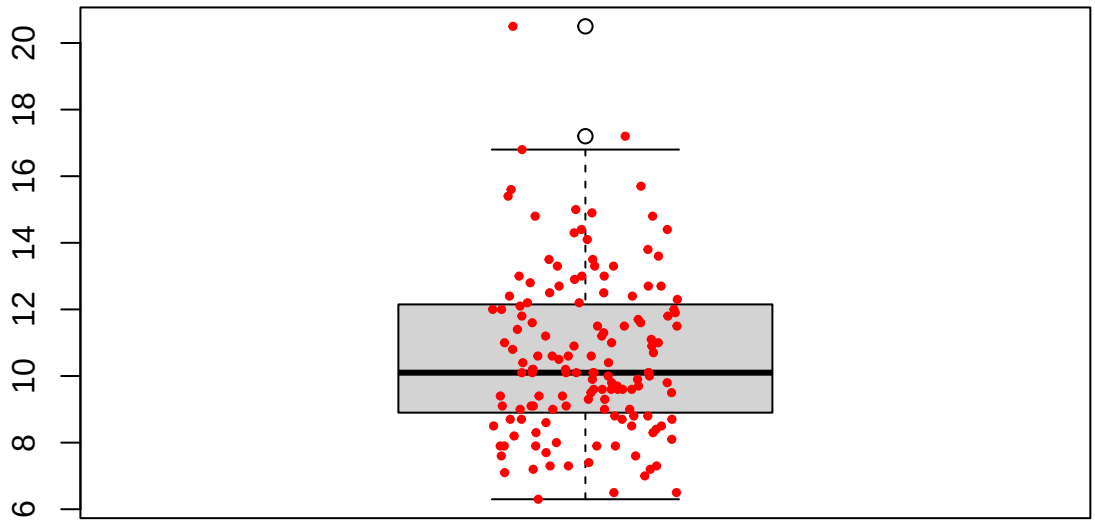
Ejercicio 3 Estudia la existencia de datos atípicos calculando con el boxplot de cada una de las variables.

Dibujamos diagramas de caja (boxplot) de las variables numéricas `Bwt` y `Hwt`, con los valores superpuestos.

```
bpWwt = boxplot(cats$Bwt)
stripchart(cats$Bwt,
           method = "jitter", add = TRUE, vertical = TRUE,
           pch=19, col= "red", cex=0.5)
```




```
bpHwt = boxplot(cats$Hwt)
stripchart(cats$Hwt,
           method = "jitter", add = TRUE, vertical = TRUE,
           pch=19, col= "red", cex=0.5)
```



Como ves en este último diagrama, hay valores atípicos. El tratamiento de esos valores atípicos es distinto en cada caso y no se pueden establecer reglas generales. A veces un valor atípico se debe simplemente a un error al anotar los datos y, en tal caso, lo más sensato es *eliminar* esa observación antes de seguir adelante con el análisis. Pero en otras ocasiones un valor atípico apunta a un fenómeno interesante y eliminarlo sería un error.

En este caso, son valores plausibles y vamos a conservarlos.

Ejercicio 4 Calcula los coeficientes de la recta de regresión. Usa la función `lm()` y guarda el resultado en la variable `modelo` para poder acceder a información extra Luego, usa la sintaxis `modelo$` para acceder a ella (`coefficients` contiene los coeficientes de la recta).

Recuerda que queremos estudiar la posible relación entre `Bwt` y `Hwt`. El diagrama que hiciste al principio del ejercicio muestra indicios de una correlación entre esas dos variables. Vamos a construir el modelo de regresión lineal correspondiente usando la función `lm()` y guardar su resultado en la variable `modelo`.

Fíjate en la sintaxis: a la izquierda de la `~` va la variable dependiente y a su derecha la dependiente

```
(modelo = lm(cats$Hwt ~ cats$Bwt))
```

```
##
## Call:
## lm(formula = cats$Hwt ~ cats$Bwt)
##
## Coefficients:
## (Intercept)     cats$Bwt
##     -0.3567       4.0341
```

y extraemos los coeficientes de la recta de regresión con:

```
modelo$coefficients
```

```
## (Intercept)    cats$Bwt
##  -0.3566624   4.0340627
```

Como ves, la pendiente de la recta es (aproximadamente) 4.034, mientras que la ordenada en el origen es -0.3567.

Recuerda que hemos escrito la ecuación genérica de la recta de regresión así:

$$y = b_0 + b_1 \cdot x$$

siendo x la variable predictora (en nuestro caso `Bwt`) e y la variable respuesta (en nuestro caso `Hwt`). Pudes guarlar los coeficientes en variables de forma separada

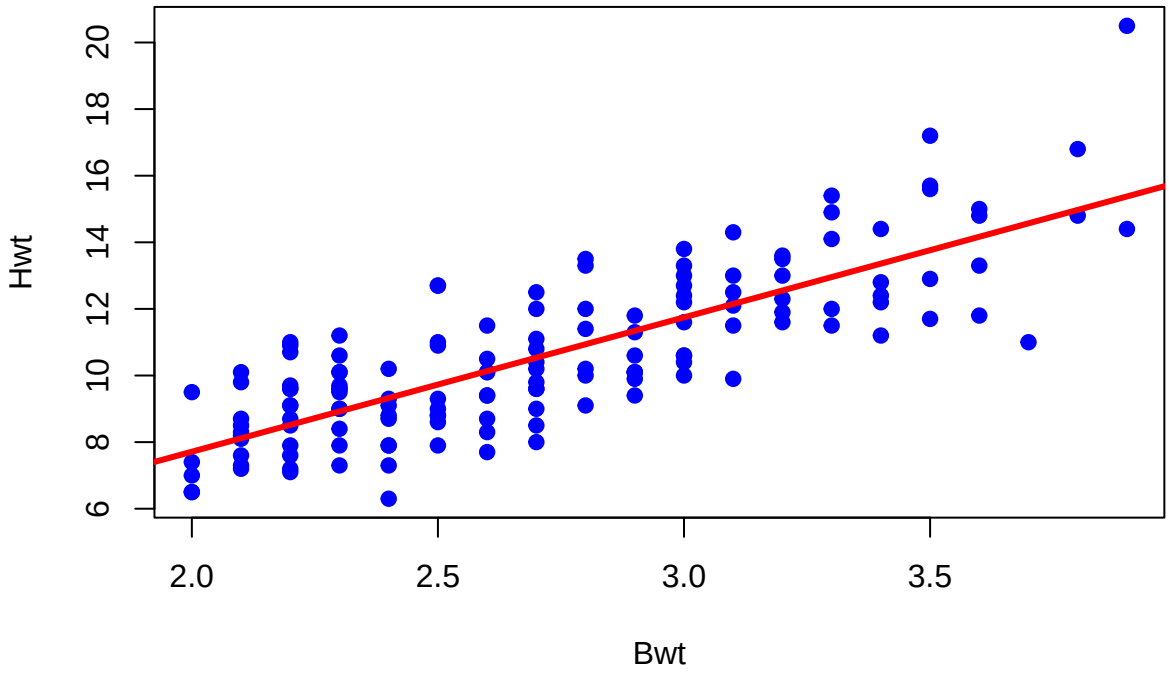
```
b0 = modelo$coefficients[1]
b1 = modelo$coefficients[2]
```

La recta de regresión de `Hwt` frente a `Bwt` es, aproximadamente:

$$Hwt = -0.357 + 4.034 \cdot Bwt$$

Puedes representar la recta de regresión al diagrama de dispersión usando la función `abline()`

```
plot(cats$Bwt, cats$Hwt, xlab="Bwt", ylab="Hwt", col="blue", pch=19)
abline(modelo, col="red", lwd=3)
```




```
modelo$coefficients[1] + modelo$coefficients[2] * cats$Bwt
```

```
modelo$fitted.values
```

Ejercicio 5 Calcula el peso del corazón predicho para individuos con pesos corporales de 2.25, 2.75, 3.25, 3.75 kg

Empieza por asegurarte de no extrapolar, es decir, de que los valores del peso corporal están en el recorrido de la variable.

```
range(cats$Bwt)
```

```
## [1] 2.0 3.9
```

Una vez comprobado que pesos corporales cuyo peso del corazón queremos predecir están en el recorrido de la variable independiente, Basta con hacer

```
modelo$coefficients[1] + modelo$coefficients[2] * c(2.25, 2.75, 3.25, 3.75)
```

Ejercicio 6 Calcula e interpreta el coeficiente de correlación. A continuación calculamos el coeficiente de correlación r de este modelo.

```
cor(cats$Bwt, cats$Hwt)
```

```
## [1] 0.8041274
```

Teniendo en cuenta ese valor, el tipo de datos de que se trata y a la vista del gráfico de dispersión, podemos considerar que el ajuste de la recta de regresión a los datos es razonablemente bueno.

Ejercicio 7 Calcula e interpreta el coeficiente de determinación. Vale

```
cor(cats$Bwt, cats$Hwt)^2
```

```
## [1] 0.6466209
```

Indica que el modelo explica aproximadamente el 64.66% de la variabilidad del peso del corazón de los gatos a través de la variabilidad del peso de su cuerpo. El resto (de la variabilidad) se explica con otros factores no considerados en el modelo. Podrían añadirse variables, es lo que se conoce como `regresión lineal múltiple`, una herramienta algo más avanzada.

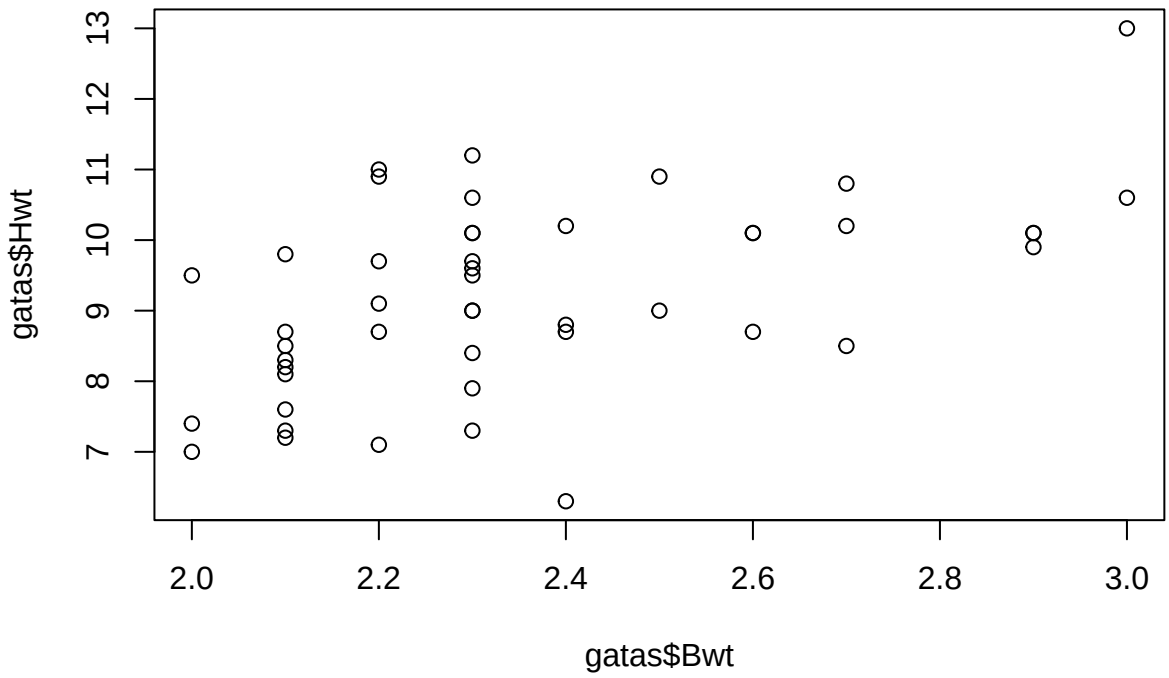
Ejercicio 8 PARA HACER EN CASA. Realiza los apartados del ejercicio anterior sólo para las gatas (seguro que ya lo has pensado: es sencillo reutilizar el código que ya generado).

Tendrás que empezar seleccionando las filas de la tabla que se refieren a las hembras, por ejemplo, teclea:

```
gatas = cats[cats$Sex == "F", ]
```

Sin embargo, ahora

```
plot(gatas$Bwt, gatas$Hwt)
```



Parece que la relación lineal entre las variables es bastante débil, de hecho,

```
cor(gatas$Bwt, gatas$Hwt)
```

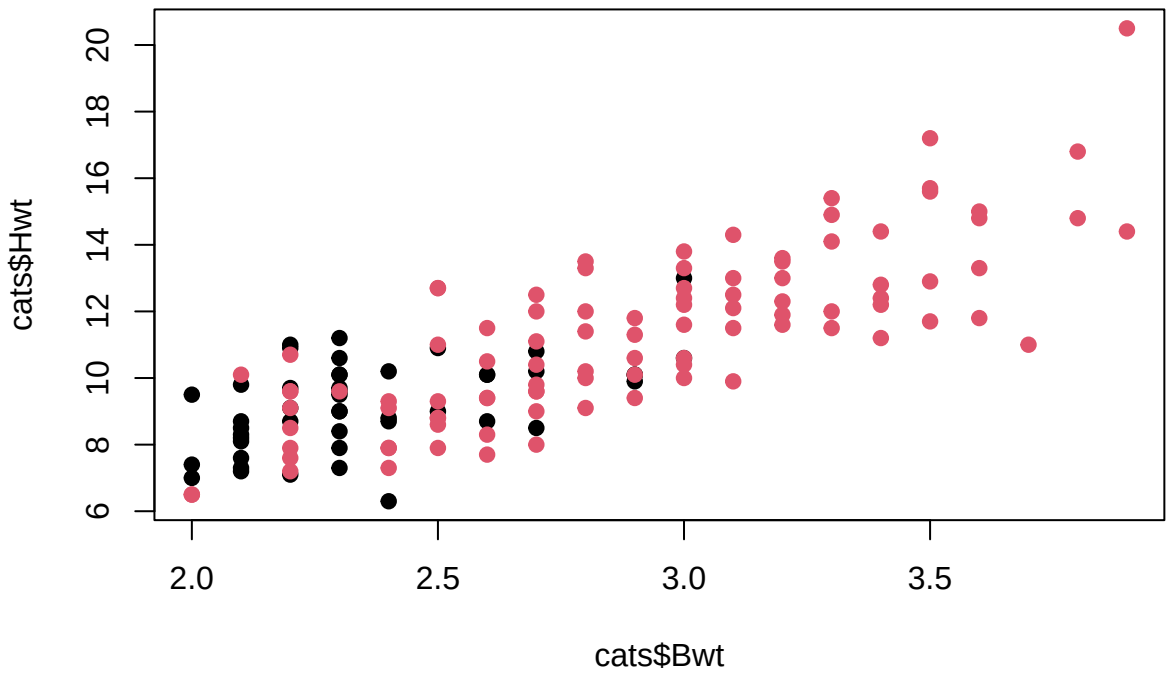
```
## [1] 0.5320497
```

```
(cor(gatas$Bwt, gatas$Hwt))^2
```

```
## [1] 0.2830768
```

los coeficientes de correlación y terminación son pequeños. Si distinguimos entre machos y hembras

```
plot(cats$Bwt, cats$Hwt, pch = 19, col = as.factor(cats$Sex))
```



se aprecia nítidamente que en los machos la tendencia lineal es más acusada que en las hembras. A la vista de esto, no tiene sentido hacer el estudio de regresión lineal sólo para las hembras.