

Regresión lineal simple

Un primer ejemplo de estadística reproducible con R markdown

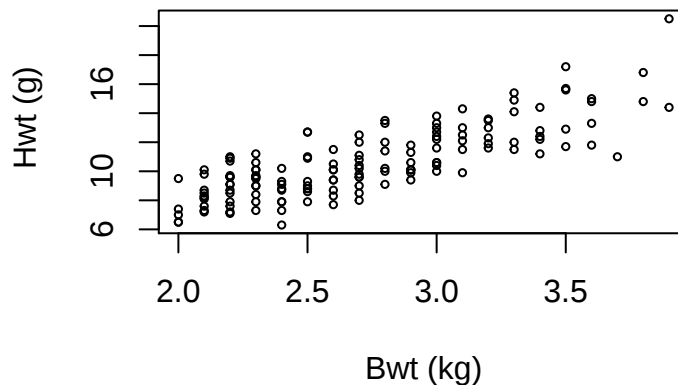
Vamos a estudiar la relación entre el peso del corazón y el peso del cuerpo del gato doméstico. Disponemos de los datos de la librería **MASS** de R, de modo que lo primero es cargar dicha librería (si no la has instalado, tendrás que hacerlo antes de cargarla en memoria) y averiguar en qué formato se presentan los datos

es decir, se trata de un **data.frame** llamado **cats**. Veamos la cabecera de los datos

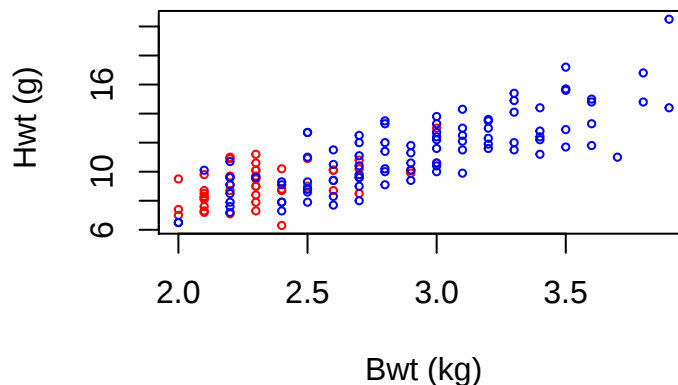
```
##   Sex Bwt Hwt
## 1  F 2.0 7.0
## 2  F 2.0 7.4
## 3  F 2.0 9.5
## 4  F 2.1 7.2
## 5  F 2.1 7.3
## 6  F 2.1 7.6
```

Hay tres variables, **sex** (female, male), **Bwt** (iniciales de body weight, peso del cuerpo, en kilogramos) y **Hwt** (heart weight, peso del corazón, en gramos), cada una de ellas en una columna

Empezamos representando las variables **Bwt** (explicativa) frente a **Hwt** (respuesta)



A la vista del diagrama, parece que cuanto mayor es el peso del gato, mayor es el peso su corazón (aunque no es imprescindible, puedes usar colores para distinguir hembras (F) de machos (M))



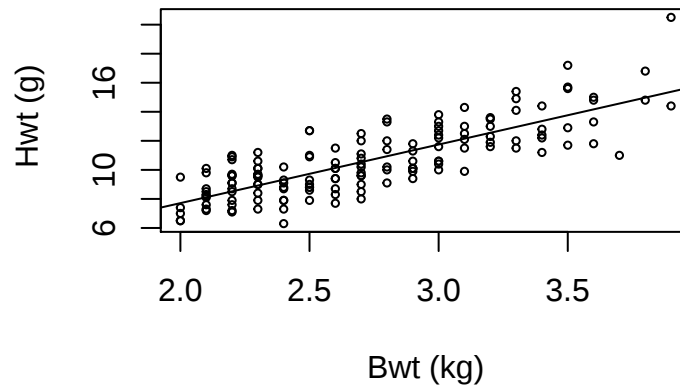
A partir del resumen del modelo de regresión lineal, obtenemos el coeficiente de correlación de Pearson

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

que no es muy bueno. Aún así, vamos a seguir adelante para analizar el resto de información que podemos obtener y los procedimientos para obtenerla.

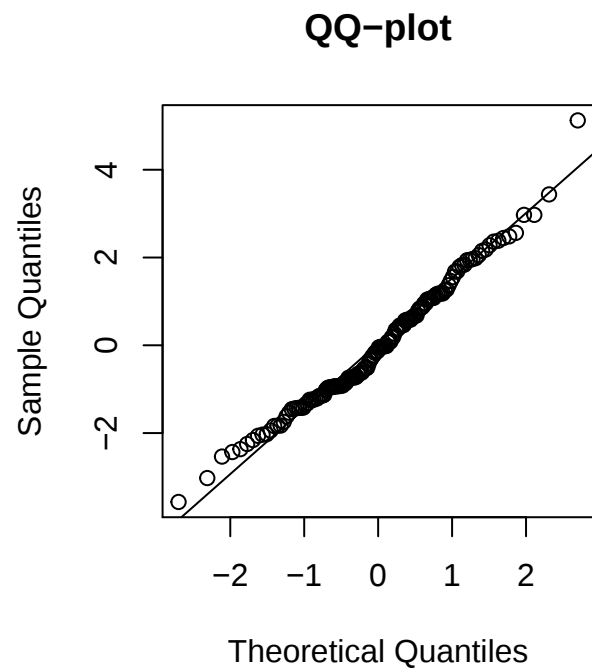
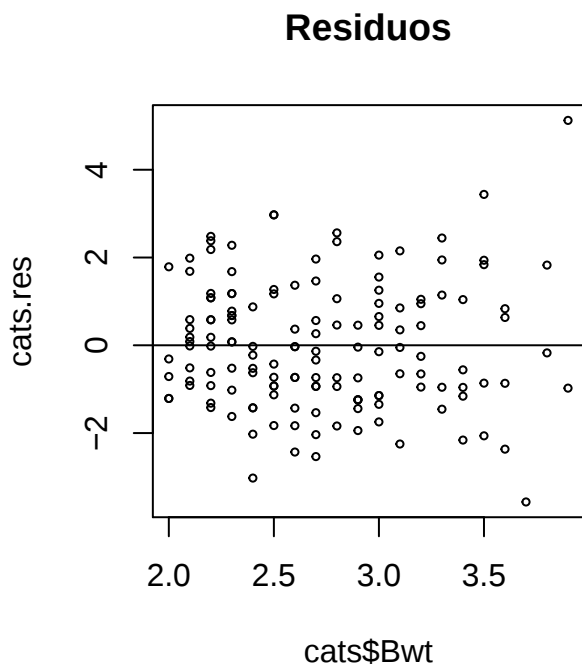
Representar la recta de regresión

junto con la nube de dispersión



Además, es fundamental comprobar que se cumplen las hipótesis necesarias para que funcione la técnica de regresión lineal:

- Normalidad en y para cada valor de x
- Homocedasticidad en los residuos



- **Normalidad:** gráfico izquierda, se observan ligeras desviaciones respecto de la normalidad (falta de simetría respecto del cero en algunas columnas), sobre todo para los individuos más grandes. Gráficode la derecha, el qq-plot muestra un ajuste bastante bueno.
- **Homocedasticidad:** gráfico izquierda, no se infringe severamente, aunque convendría eliminar algunos outliers.

Calcular el peso del corazón predicho por el modelo para el peso del cuerpo de 2, 2.25, 2.5, 2.75 y 3kg:

```
##           1           2           3           4           5
## 7.711463  8.719979  9.728494 10.737010 11.745526
```

Cálculo del intervalo de confianza para los coeficientes de la recta de regresión

Nos referimos a la recta teórica

```
confint(lmXY)
```

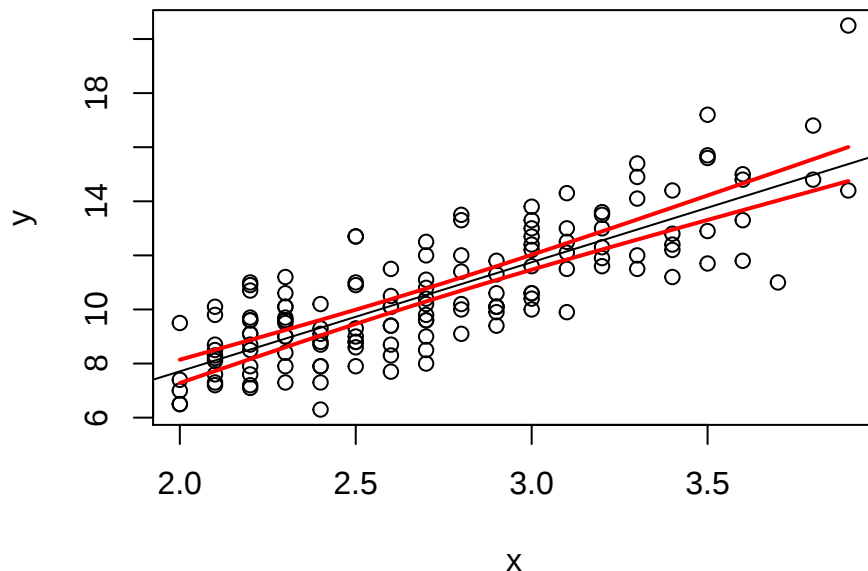
```
##           2.5 %    97.5 %
## (Intercept) -1.725163  1.011838
## x           3.539343  4.528782
```

La pendiente de la recta es significativamente distinta de cero.

Representar las bandas de confianza

La idea es

- calcular el valor estimado (esperado) de la variable y para muchos valores de la variable explicativa x
- calcular el intervalo de confianza para cada uno de esos valores predichos
- al representar conjuntamente cada uno de esos intervalos, el resultado son las bandas de confianza: una superior y otra inferior, al unir todos los extremos superiores e inferiores de los intervalos.



Muchos puntos escapan a las bandas de confianza. Esto es coherente con un coeficiente de correlación bajo (poca tendencia de los puntos a estar alineados).