

Estudio de datos biométricos sobre el ciervo volante.

Olivia Marcos Diaz de Otazu Irene Canales Giménez Samuel Melián Benito

today

```
library(tidyverse)
library(summarytools)
library(GGally)
library(gt)
library(flextable)
library(knitr)
library(corrplot)
library(ggplot2)
library(kableExtra)
```

Objetivo del estudio

El estudio de Estadística Descriptiva que vamos a realizar tiene como foco el análisis de datos biométricos del **ciervo volante** (*Lucanus cervus*), el mayor escarabajo de Europa y una especie cuya supervivencia está amenazada por la destrucción de su hábitat, bosques con abundancia de madera muerta.

El interés biométrico de esta especie nos lleva a dos preguntas principales:

Dimorfismo Sexual: analizaremos si existen diferencias significativas en tamaño, específicamente en la anchura de la cabeza (KB), entre machos y hembras.

Influencia del Hábitat: indagaremos si el tamaño adulto de los ejemplares está correlacionado con la calidad del hábitat de procedencia (Asturias, Cantabria u otra).

Se han tomado medidas en machos y hembras de ciervo volante de tres procedencias, midiendo las variables indicadas en la figura siguiente:

```
knitr:::include_graphics("imagenes/imagen.png")
```

10 J. T. Clark

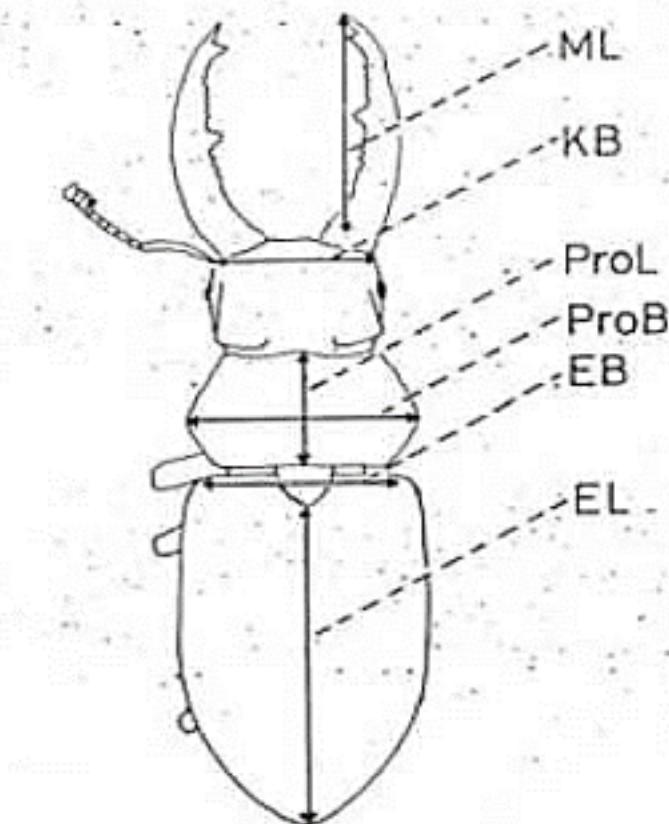


FIG. 1. Kuyten system measurements as applied to *Lucanus cervus*.



El fichero CIERVO.txt recoge, para un total de 265 ejemplares de ciervo volante, parte de estas variables. Concretamente:

- Anchura de la cabeza (KB) expresada en milímetros
- Longitud de los élitros (EL) también en milímetros
- Sexo del individuo (hembra o macho)
- Provincia de procedencia del individuo (Asturias, Cantabria u otra)

Los objetivos de este estudio son:

- Realizar un análisis descriptivo de la anchura de la cabeza de una muestra de 150 ejemplares que refleje sus principales características.
- Comparar la anchura de la cabeza de las hembras con la de los machos.
- Comparar la anchura de la cabeza de los ejemplares de las tres procedencias.
- Analizar la correlación entre la anchura de la cabeza y la longitud del élitro para la muestra.
- Analizar, para cada sexo por separado, la correlación entre la anchura de la cabeza y la longitud del élitro.
- En caso de que para alguno de los sexos ambas variables estén relacionadas, ajustar una recta de regresión que explique la anchura de la cabeza en función de la longitud del élitro e interpretarla.

Metodología

En este análisis cuantitativo de tipo descriptivo analizaremos los datos biométricos del ciervo volante. Para ello, emplearemos el software R.

Selección de la muestra

Se toma una muestra de 150 ejemplares sobre la población total, manteniendo la proporción original entre machos y hembras. Estos se tomarán de forma arbitraria empleando una semilla para asegurar la aleatoriedad y así evitar posibles sesgos.

Análisis de los datos

Con esta muestra crearemos una tabla para resumir algunos estadísticos: Media, Mediana, Primer cuartil, Tercer cuartil, Desviación estándar, el rango intercuartílico, y finalmente los valores máximo y mínimo. A partir de esta tabla crearemos el histograma.

Comparación de la anchura de la cabeza por sexo de los ejemplares

A continuación se crea una nueva tabla donde se incluya el tamaño de la muestra, la media, la mediana, desviación estándar y el IQR, pero diferenciando entre machos y hembras. A partir de ella creamos un gráfico de tipo boxplot para visualizarlo.

Comparación de la anchura de la cabeza por procedencia de los ejemplares

De igual forma, creamos una tabla diferenciando por provincia: Asturias, Cantabria, Otras y Desconocida y se inserta el boxplot correspondiente.

Análisis de la correlación entre la anchura de la cabeza y la longitud de élitro

Análisis por sexo de la correlación anterior

Recta de regresión, por sexo, de la anchura de la cabeza en función de la longitud de los élitros

Resultados

Análisis de los datos

```
set.seed(78) #Semilla para asegurar la aleatoriedad a la hora de elegir la muestra de individuos y así

indices <- sample(1:nrow(datos), size = 150, replace = FALSE) #Tomamos la muestra aleatoria de 150 individuos

muestra <- datos[indices, ] #Creamos el dataframe de esta nueva muestra de 150 individuos

tabla_kb <- muestra |>
  summarise(
    N = n(),
    Media = mean(KB, na.rm = TRUE),
    Q1 = quantile(KB, 0.25, na.rm = TRUE),
    Mediana = median(KB, na.rm = TRUE),
    Q3 = quantile(KB, 0.75, na.rm = TRUE),
    DE = sd(KB, na.rm = TRUE),
    IQR = IQR(KB, na.rm = TRUE),
    Min = min(KB, na.rm = TRUE),
    Max = max(KB, na.rm = TRUE)
  )

tabla_kb |>
  kable(
    col.names = c(
      "N", "Media", "Q1", "Mediana", "Q3",
      "Desv. típica", "IQR", "Mín", "Máx"
    ),
    caption = "Estadísticos descriptivos de la anchura de la cabeza (KB)",
    digits = 2,
    booktabs = TRUE,
    align = "c"
  ) |>
  kable_styling(latex_options = "hold_position")
```

Table 1: Estadísticos descriptivos de la anchura de la cabeza (KB)

| N | Media | Q1 | Mediana | Q3 | Desv. típica | IQR | Mín | Máx |
|-----|-------|------|---------|------|--------------|------|-----|------|
| 150 | 12.15 | 8.96 | 10.8 | 15.1 | 3.97 | 6.14 | 7.1 | 24.1 |

```
muestra |> ggplot(aes(KB)) +
  geom_histogram(
    aes(y = after_stat(density)),
    fill = "#A6C8EO",
    color = "grey30",
    bins = 30,
    alpha = 0.7
  ) +
  geom_density(
    color = "#08306B",
    linewidth = 0.8,
    alpha = 0.5
```

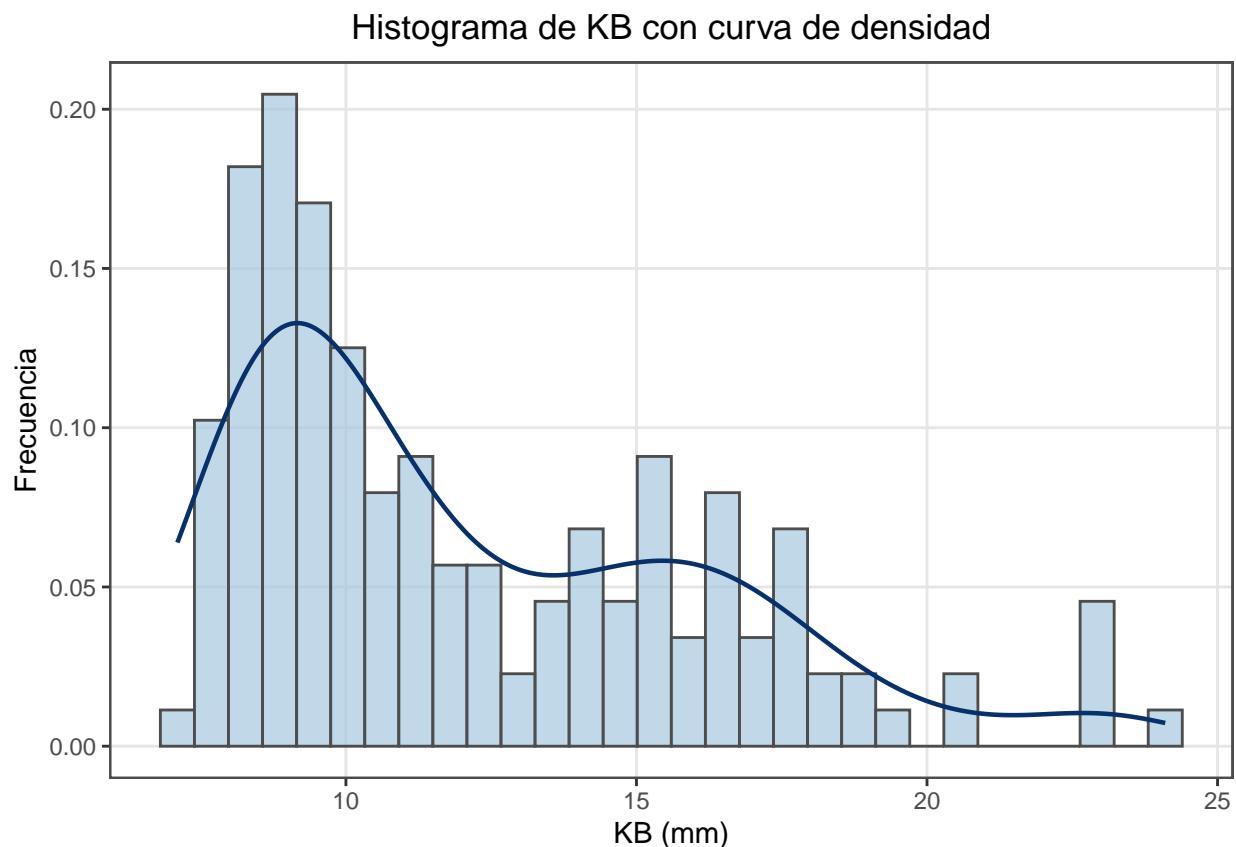
```

) + 

labs(
  title = "Histograma de KB con curva de densidad",
  x = "KB (mm)",
  y = "Frecuencia"
) + 

theme_bw() +
theme(
  plot.title = element_text(hjust = 0.5),
  axis.title = element_text(),
  axis.text = element_text(),
  panel.grid.major = element_line(color = "grey90"),
  panel.grid.minor = element_blank(),
  panel.border = element_rect(color = "grey30", fill = NA, size = 0.8)
)

```



Comparación de la anchura de la cabeza por sexos

```

tabla_kb_sexo <- muestra |>
  group_by(SEXO) |>
  summarise(
    N = n(),
    Media = mean(KB, na.rm = TRUE),
    DE = sd(KB, na.rm = TRUE),
    )

```

```

Mediana = median(KB, na.rm = TRUE),
IQR = IQR(KB, na.rm = TRUE),
)

tabla_kb_sexo |>
kable(
  col.names = c("Sexo", "N", "Media", "Desv. típica", "Mediana", "IQR"),
  caption = "\ Estadísticos descriptivos de la anchura de la cabeza (KB) por sexo",
  digits = 2,
  booktabs = TRUE,
  align = "c",
  escape = FALSE # permite que LaTeX interprete la caption
) |>
kable_styling(
  latex_options = "hold_position",
  font_size = 10
)

```

Table 2: Estadísticos descriptivos de la anchura de la cabeza (KB) por sexo

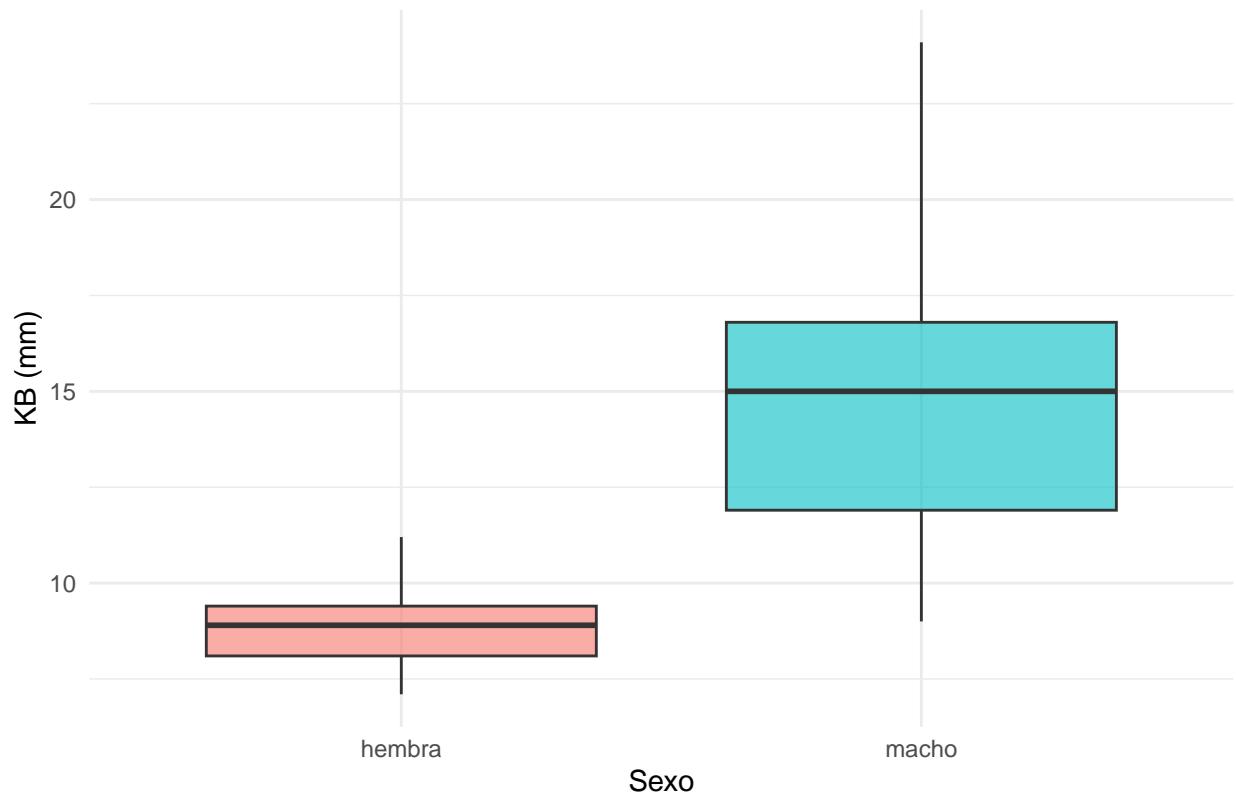
| | Sexo | N | Media | Desv. típica | Mediana | IQR |
|--|--------|----|-------|--------------|---------|-----|
| | hembra | 69 | 8.93 | 0.91 | 8.9 | 1.3 |
| | macho | 81 | 14.89 | 3.48 | 15.0 | 4.9 |

```

ggplot(muestra, aes(x = SEXO, y = KB, fill = SEXO)) +
  geom_boxplot(alpha = 0.6) +
  labs(
    title = "Anchura de la cabeza (KB) por sexo",
    x = "Sexo",
    y = "KB (mm)"
  ) +
  theme_minimal() +
  theme(legend.position = "none")

```

Anchura de la cabeza (KB) por sexo



Comparación de la anchura de la cabeza por provincias

```
tabla_kb_provincia <- muestra |>
  group_by(PROVINCIA) |>
  summarise(
    N = n(),
    Media = mean(KB, na.rm = TRUE),
    DE = sd(KB, na.rm = TRUE),
    Mediana = median(KB, na.rm = TRUE),
    IQR = IQR(KB, na.rm = TRUE),
  )
  print(tabla_kb_provincia)

## # A tibble: 4 x 6
##   PROVINCIA      N  Media     DE  Mediana    IQR
##   <fct>     <int> <dbl> <dbl> <dbl> <dbl>
## 1 Asturias     102  11.7  3.68  10.6  5.78
## 2 Cantabria     17  12.9  3.79  11.5  6.9
## 3 Otras         24  13.5  5.10  11.0  6.99
## 4 Desconocida     7  12.0  3.67  11.2  3.55
```

Presentamos ahora un boxplot o diagrama de cajas y bigotes de la anchura de la cabeza para cada provincia

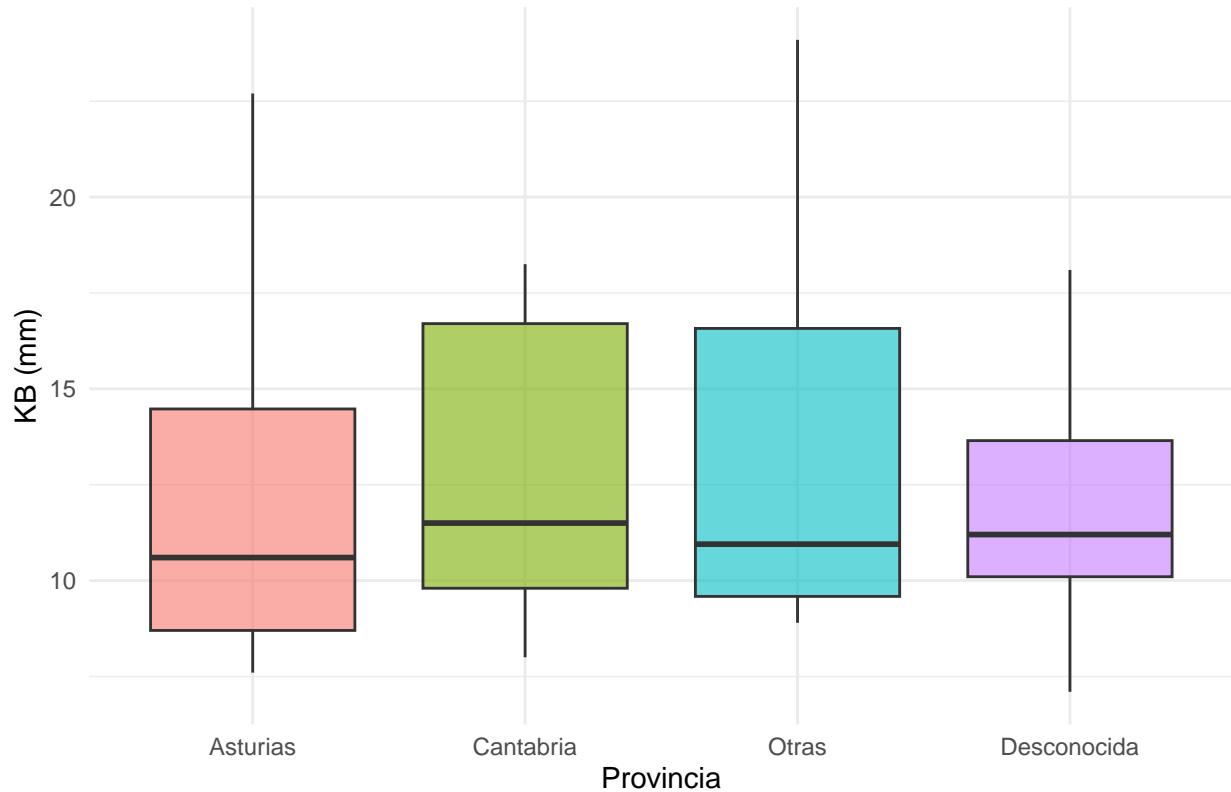
```
ggplot(muestra, aes(x = PROVINCIA, y = KB, fill = PROVINCIA)) +
  geom_boxplot(alpha = 0.6) +
  labs(
    title = "Anchura de la cabeza (KB) por provincia",
```

```

x = "Provincia",
y = "KB (mm)"
) +
theme_minimal() +
theme(legend.position = "none")

```

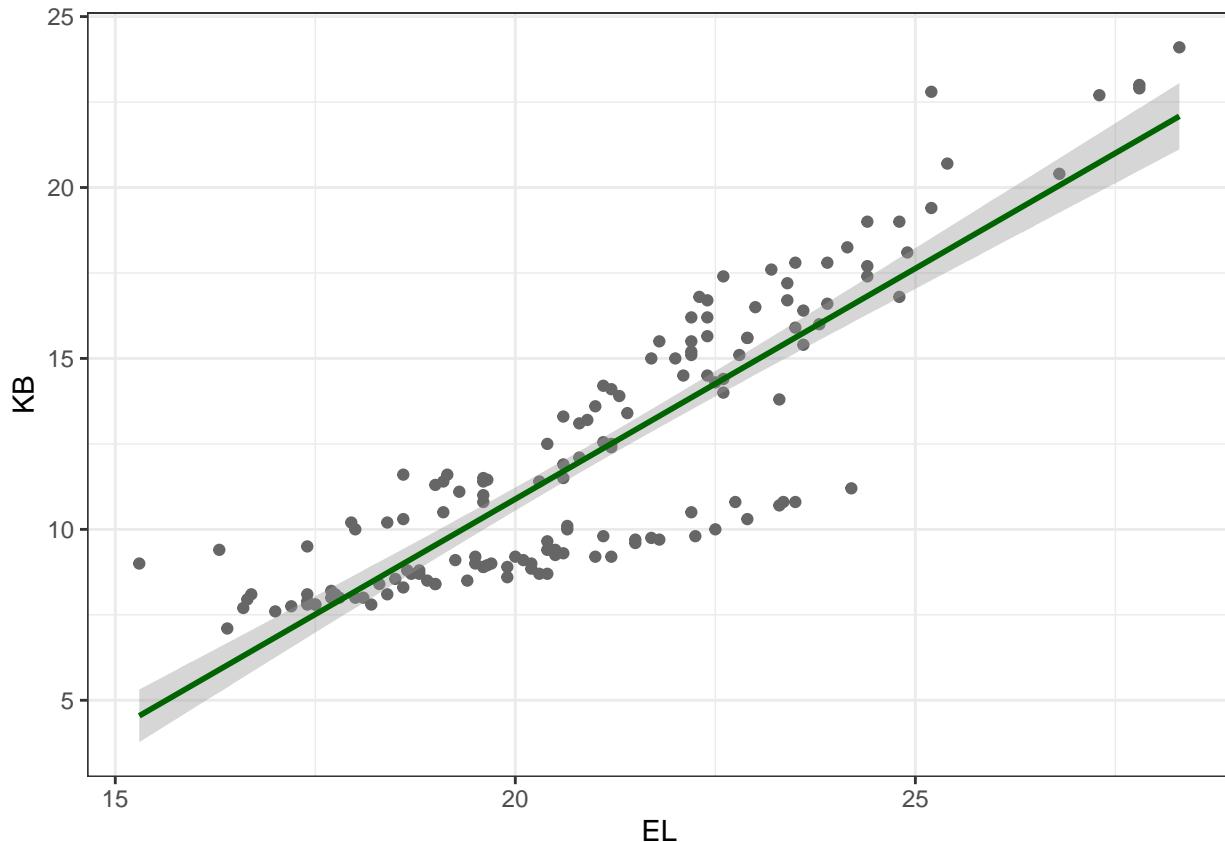
Anchura de la cabeza (KB) por provincia



```

# Recta de regresión de la variable EL en función de KB
muestra |>
  select(KB, EL) |> # Selecciona las columnas 'EL' y 'KB'
  ggplot(aes(x=EL, y = KB)) + # Define las variables para el gráfico
  geom_point(
    color = "grey40",
  ) + # Añade puntos al gráfico
  theme_bw() + # Aplica un tema en blanco y negro
  geom_smooth(
    method = lm,
    se = TRUE,
    color = "darkgreen"
  ) # Añade una línea de regresión lineal sin intervalo de confianza
## `geom_smooth()` using formula = 'y ~ x'

```



```
# Creación y resumen de un modelo lineal para 'est' en función de 'mg'
modelo <- muestra |>
  select(KB, EL) |> # Selecciona las variables 'KB' y 'EL'
  with(lm(KB ~ EL)) # Crea un modelo lineal de 'KB' en función de 'EL'

#Valor de r^2
r2 <- summary(modelo)$r.squared
r2

## [1] 0.7560048

#Valor de r
r <- sign(coef(modelo)[2]) * sqrt(r2)
r

##          EL
## 0.8694853

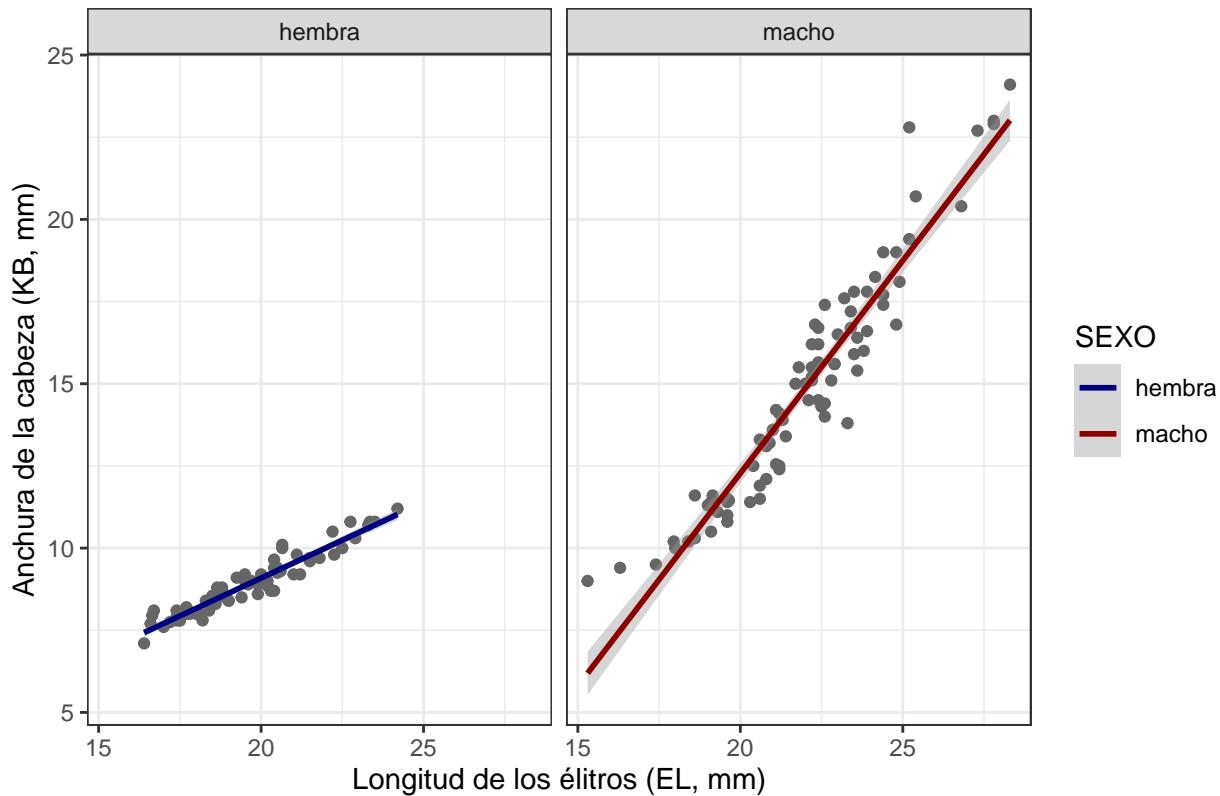
# Recta de regresión de la variable 'est' en función de 'mg' filtrando por tipo 'C'
muestra |>
  # 'filter(tipo == "C")' filtra el dataframe para incluir solo las filas donde la columna 'tipo' es igual a 'C'
  ggplot(aes(x = EL, y = KB)) +
  facet_wrap(~ SEXO) +
  geom_point(
    color = "grey40",
  ) +
  labs(
    x = "Longitud de los élitros (EL, mm)",
```

```

y = "Anchura de la cabeza (KB, mm)",
title = ""
) +
# 'geom_smooth(method = "lm")' añade una línea de ajuste lineal al gráfico. 'method = "lm"' indica qu
geom_smooth(
  method = "lm",
  aes(color = SEXO),
) +
scale_color_manual(values = c("hembra" = "navy", "macho" = "darkred")) +
theme_bw();

## `geom_smooth()` using formula = 'y ~ x'

```



```

modelo_hembra <- lm(
  KB ~ EL,
  data = muestra,
  subset = SEXO == "hembra"
)
modelo_macho <- lm(
  KB ~ EL,
  data = muestra,
  subset = SEXO == "macho"
)
#Valor de r^2 para la regresión lineal KB ~ EL para hembra y macho por separado
r2_hembra <- summary(modelo_hembra)$r.squared
r2_macho <- summary(modelo_macho)$r.squared

```

```
#Valor de r para la regresion lineal KB ~ EL para hembra y macho por separado
r_hembra <- sign(coef(modelo_hembra)[2]) * sqrt(r2_hembra)
r_macho <- sign(coef(modelo_macho)[2]) * sqrt(r2_macho)
```

#Crearemos una tabla que tenga el valor de r y r^2 para la regresión de hembra y lo mismo para la de macho (tabla 2x2)

#APARTADO F

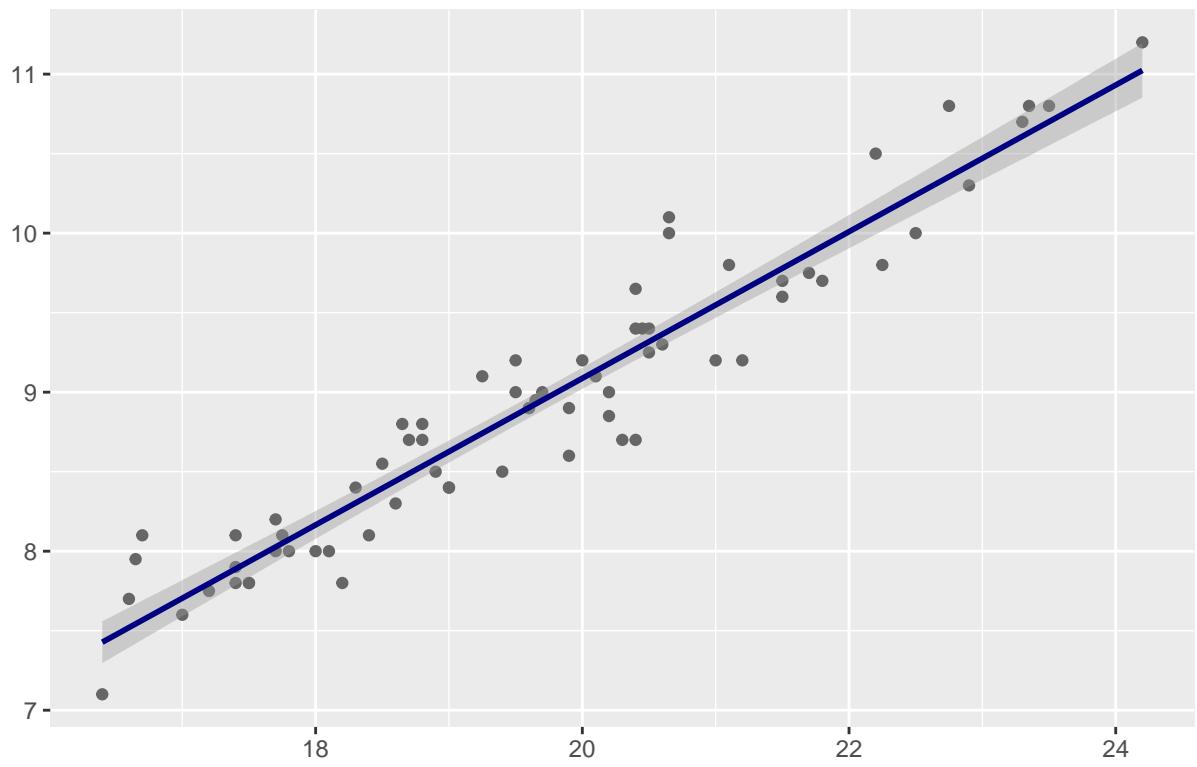
```
summary(modelo_hembra)$coefficients
```

| | Estimate | Std. Error | t value | Pr(> t) |
|----------------|------------|------------|------------|--------------|
| ## (Intercept) | -0.1332862 | 0.34590530 | -0.3853256 | 7.012168e-01 |
| ## EL | 0.4610258 | 0.01751241 | 26.3256655 | 4.700964e-37 |

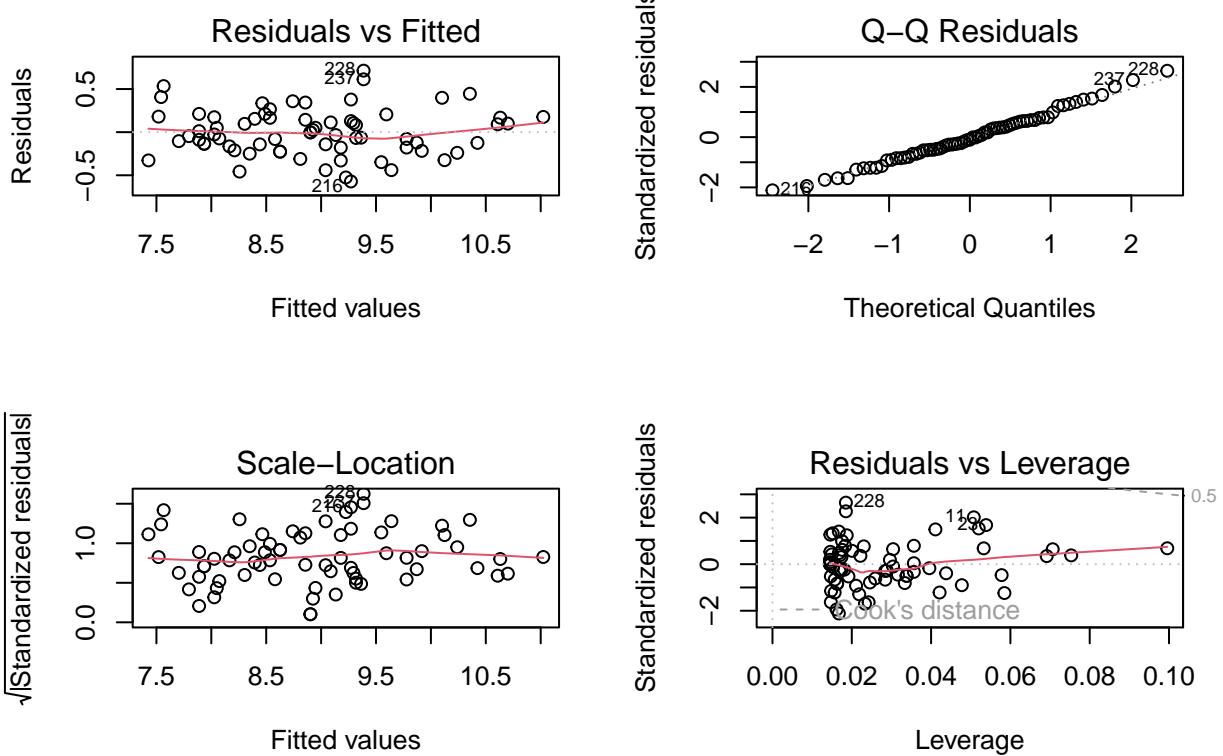
#TENDREMOS QUE DARLES FORMATO A ESTA Y TODAS LAS TABLAS (EN UNA DE LAS FOTOS DEL GRUPO ESTÁ CÓMO SE PUEDE HACER)

```
muestra |>
  filter(SEXO == "hembra") |>
  ggplot(
    aes(x = EL, y = KB),
  ) +
  geom_point(
    color = "grey40"
  ) +
  geom_smooth(
    method = "lm",
    color = "navy"
  ) +
  labs(
    x = "",
    y = "",
    title = ""
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
par(mfrow = c(2, 2))
plot(modelo_hembra)
```



```
#Nos falta también modificar si queremos los colores de las gráficas, poner bien títulos, ejes...
```

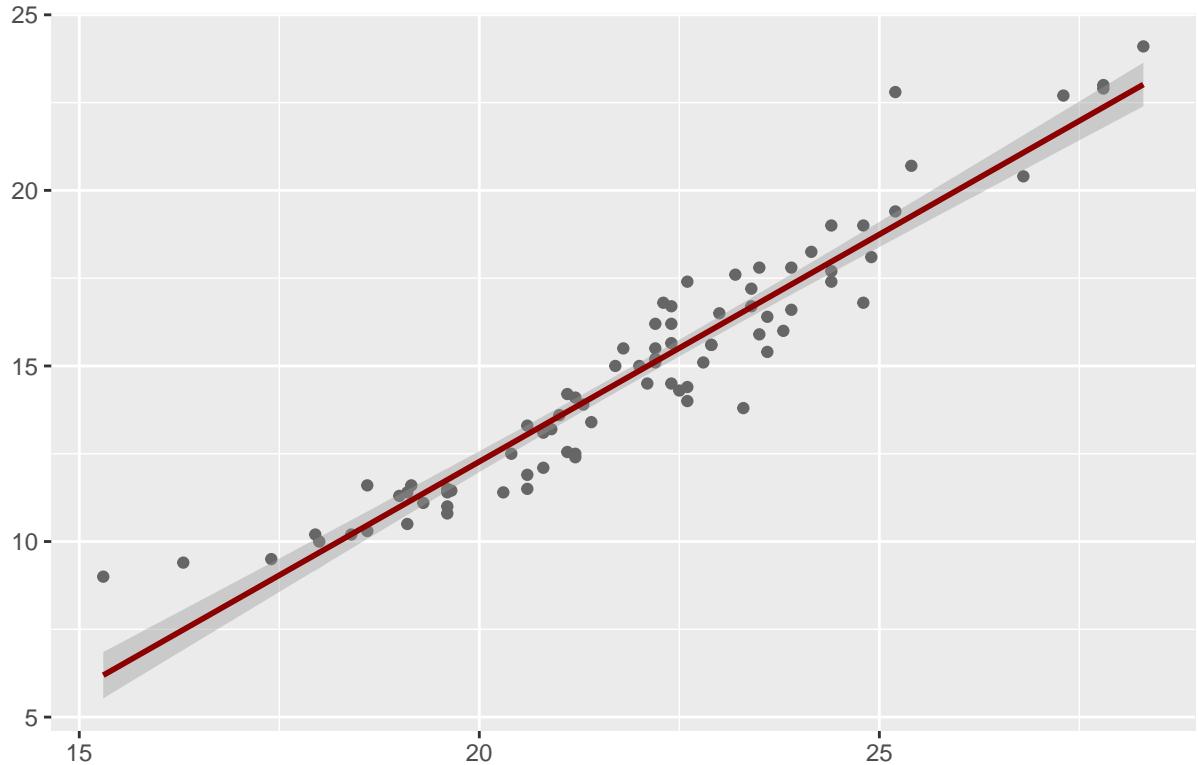
```
summary(modelo_macho)$coefficients
```

```
##             Estimate Std. Error   t value   Pr(>|t|)    
## (Intercept) -13.60437 1.02192107 -13.31254 7.002395e-22
## EL           1.29397 0.04609877  28.06950 7.559942e-43
```

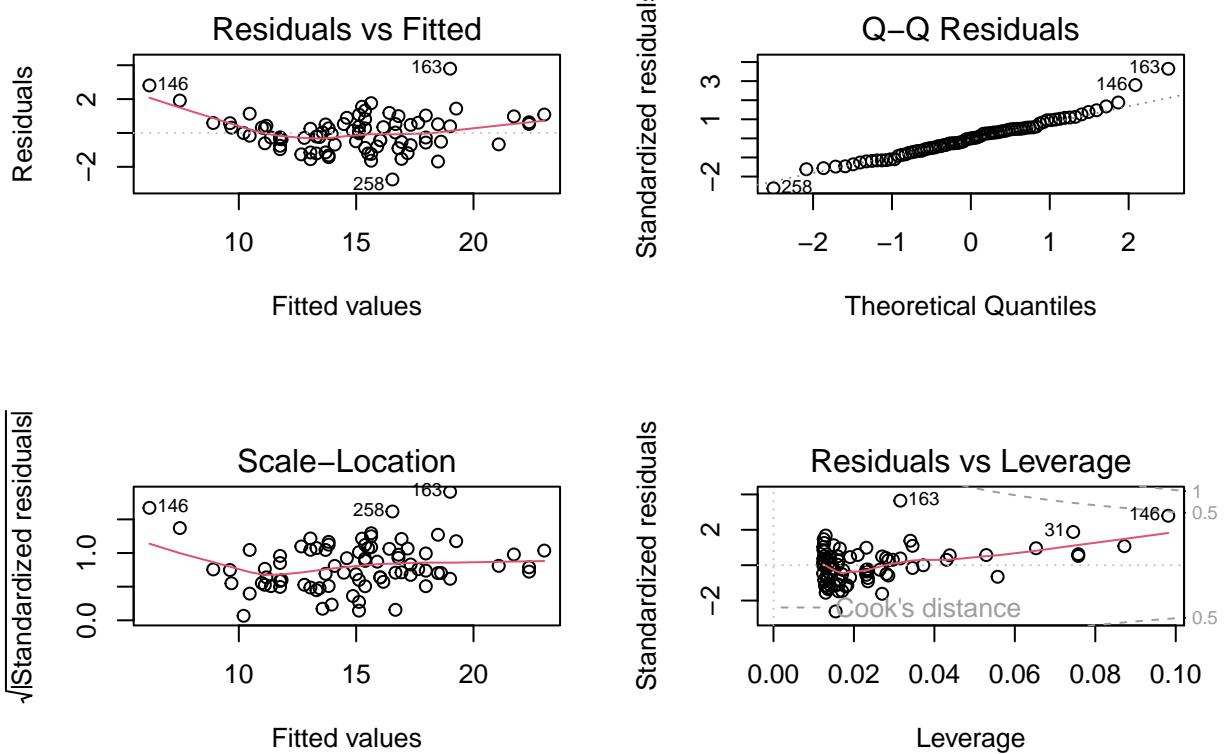
```
#Lo mismo para esta tabla
```

```
muestra |>
  filter(SEXO == "macho") |>
  ggplot(
    aes(x = EL, y = KB),
  ) +
  geom_point(
    color = "grey40"
  ) +
  geom_smooth(
    method = "lm",
    color = "darkred"
  ) +
  labs(
    x = "",
    y = "",
    title = ""
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
par(mfrow = c(2, 2))
plot(modelo_macho)
```



Conclusiones

Anexo:Código R