Antonio González Artime
Evaluación R
Feb-2014

Ejercicio 1: Newcomb

Exploración inicial:

Para poder trabajar con la *muestra* que nos han facilitado, el primer paso será realizar una pequeña exploración inicial de ésta. Nos haremos así una pequeña idea acerca de la estructura del fichero de entrada, el tipo (o tipos) de datos que tendremos que manejar e, incluso en este caso, observaremos ya ciertos valores que, a priori, podrían parecer *atípicos* (cuando menos, resulta llamativo observar tan solo dos valores negativos dentro de un conjunto de unas 50-60 observaciones que no parece acostumbrar a tomar valores inferiores a 20).

Podemos, por tanto, concluir en este apartado, que nos enfrentamos a una muestra de *variable contínua*; organizado en una sola columna, que toma mayoritariamente valores enteros positivos de no más de dos dígitos.

Carga de datos:

Para que R nos ayude a realizar los cálculos que nos ayudarán a analizar la muestra, lo primero que debemos hacer es importarla. Para ello, y puesto que el volumen de datos es pequeño y consta de una sola variable; un simple *vector* resulta una estructura de datos adecuada y suficiente para almacenar los datos. Puesto que ya conocemos que nuestra muestra consta de enteros, le proporcionaremos este hint a la función scan.

```
> t_ray <- scan("./data/newcomb.txt", what=integer())
```

Con la función summary obtendremos rápidamente el valor de ciertas medidas que nos ayudan a conocer mejor nuestra muestra.

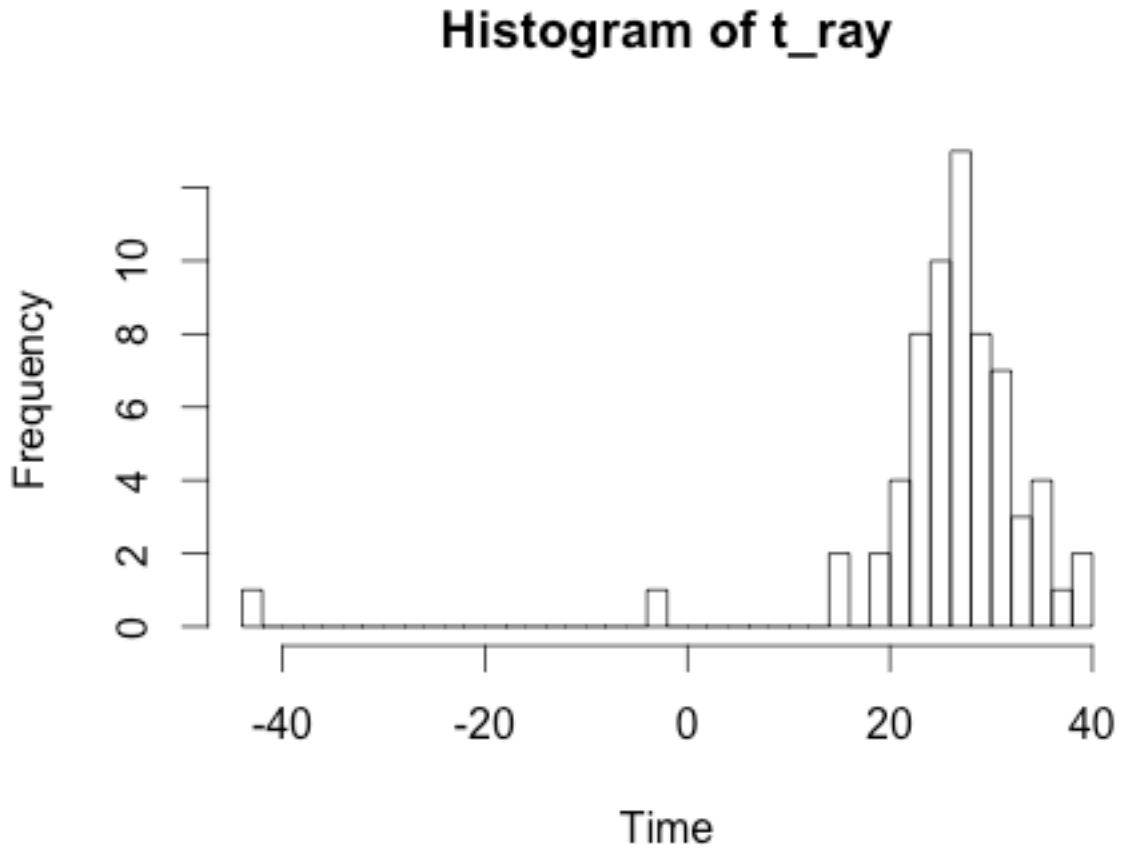
```
> summary(t_ray)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -44.00   24.00   27.00   26.21   30.75   40.00
```

No obstante, ninguno de los valores anteriores nos da una información muy interesante, puesto que al ser el conjunto tan pequeño, ya nos habíamos hecho una idea de estos valores en nuestra exploración inicial.

Análisis exploratorio:

Probablemente un histograma pudiese darnos más información sobre la muestra:

```
> hist(t_ray,40,
     main="Histogram of t_ray",
     xlab="Time",
     ylab="Frequency")
```



El histograma anterior se ve fuertemente *asimétrico* debido a los puntos negativos. Una manera de mejorarlo podrías ser la eliminación de éstos aunque, antes de nada, comprobaremos que estos datos negativos son en realidad valores atípicos basándonos en una generalización de la desigualdad de *Chebycheff* (comprobaremos si éstos se encuentran a más de tres desviaciones típicas de la media). En primer lugar, obtenemos los límites inferior y superior a partir de los cuales cualquier punto situado más haya de éstos serían considerados valores atípicos.

```
low_atyp_limit <- mean(t_ray)-3*sd(t_ray)
high_atyp_limit <- mean(t_ray)+3*sd(t_ray)
```

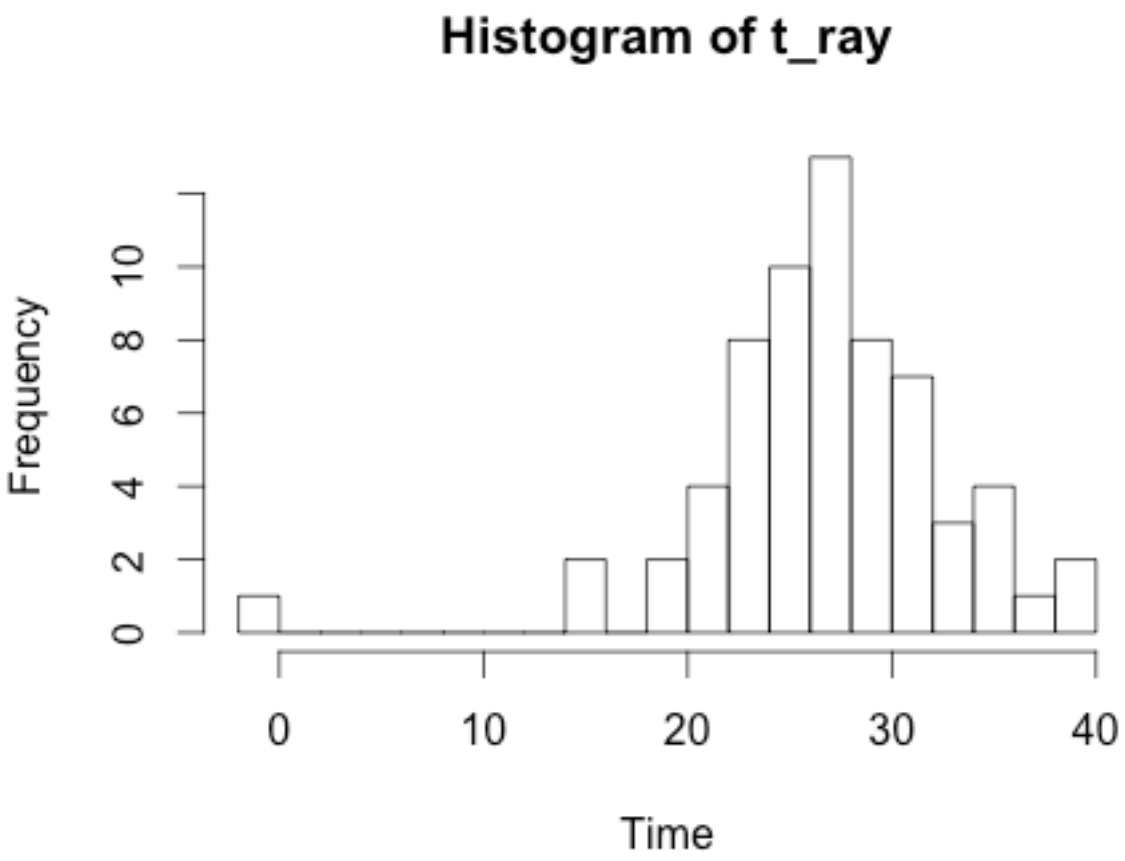
Ahora, crearemos un nuevo vector en el que indexaremos por aquellos puntos que se encuentren dentro del rango que hemos establecido.

```
t_ray_pos <- t_ray[t_ray > low_atyp_limit & t_ray < high_atyp_limit]
```

Si vemos los valores que se quedan finalmente en t_ray_pos; veremos cómo la desigualdad de Chebycheff excluye únicamente el punto negativo más bajo (el -44); mientras que mantiene el -2. Tomando este nuevo conjunto.

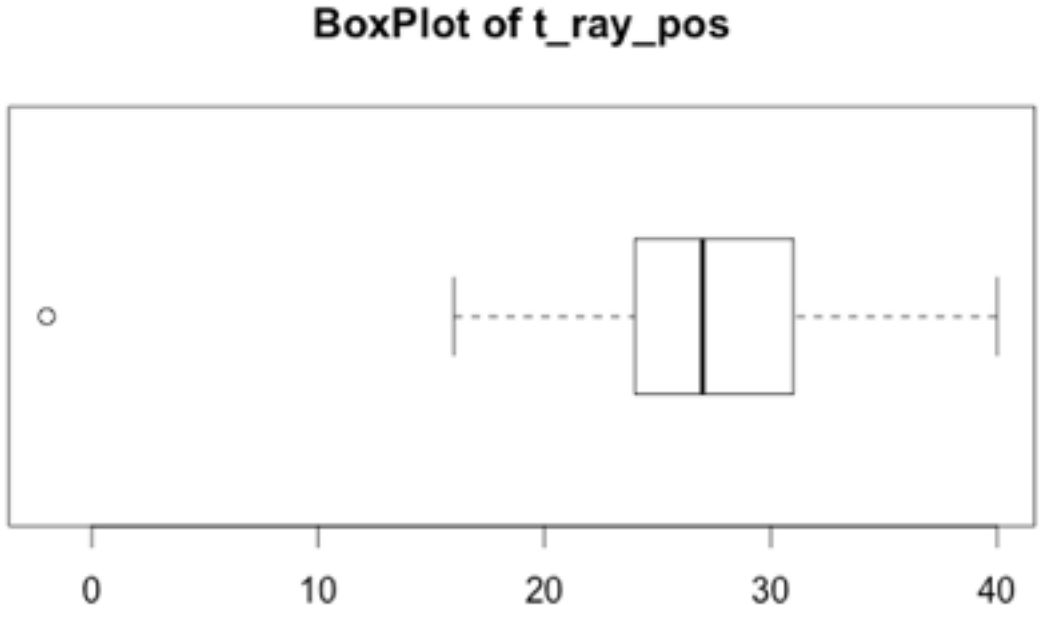
Podemos ahora volver a pintar el *histograma*:

```
hist(t_ray_pos,20,
     main="Histogram of t_ray",
     xlab="Time",
     ylab="Frequency")
```



Si aprovechamos también para dibujar un diagrama de *cajas*, veremos que éste considera como punto atípico el dato negativo que habíamos conservado:

```
boxplot(t_ray_pos, horizontal=TRUE, main="BoxPlot of t_ray_pos")
```



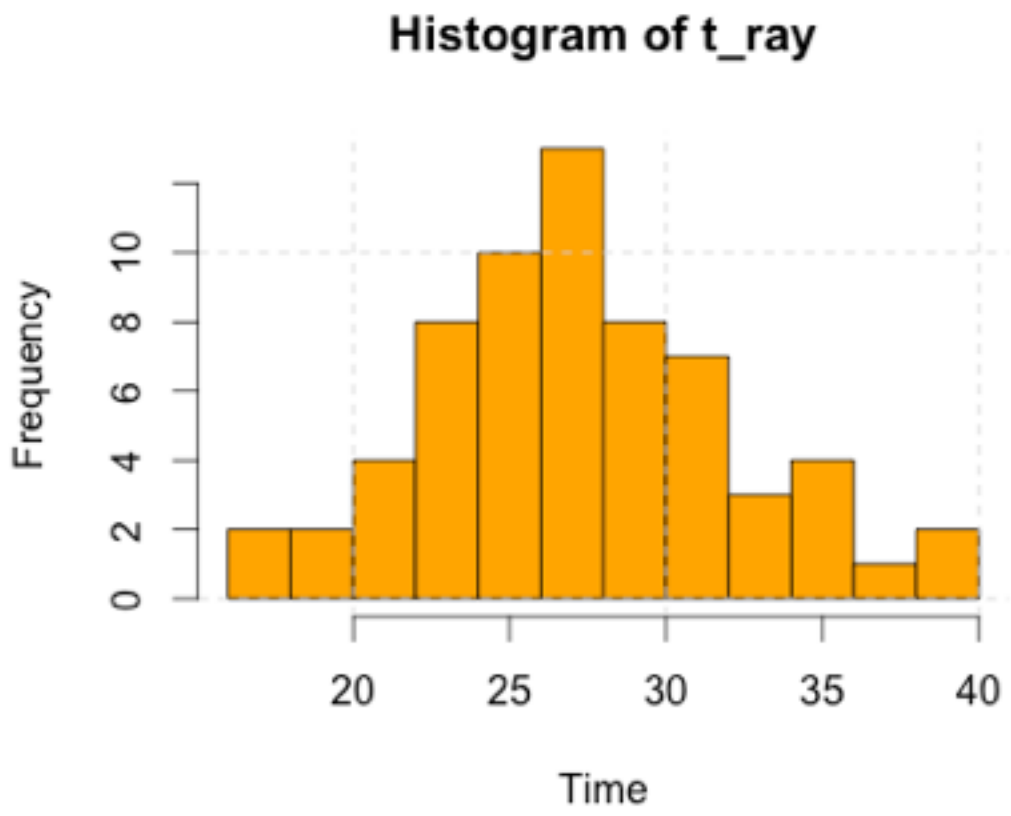
Además del punto atípico, podemos observar *el rango intercuartílico*, la mediana, etc. En lugar de utilizar la generalización sobre la desigualdad de Chebycheff, comprobaremos ahora porqué el valor negativo que nos queda en la muestra es considerado atípico. Para ello, nos basaremos en el criterio de que será conservado atípico si es menor que el primer cuartil y su distancia al cuartil de referencia es mayor que una vez y media el rango intercuartílico.

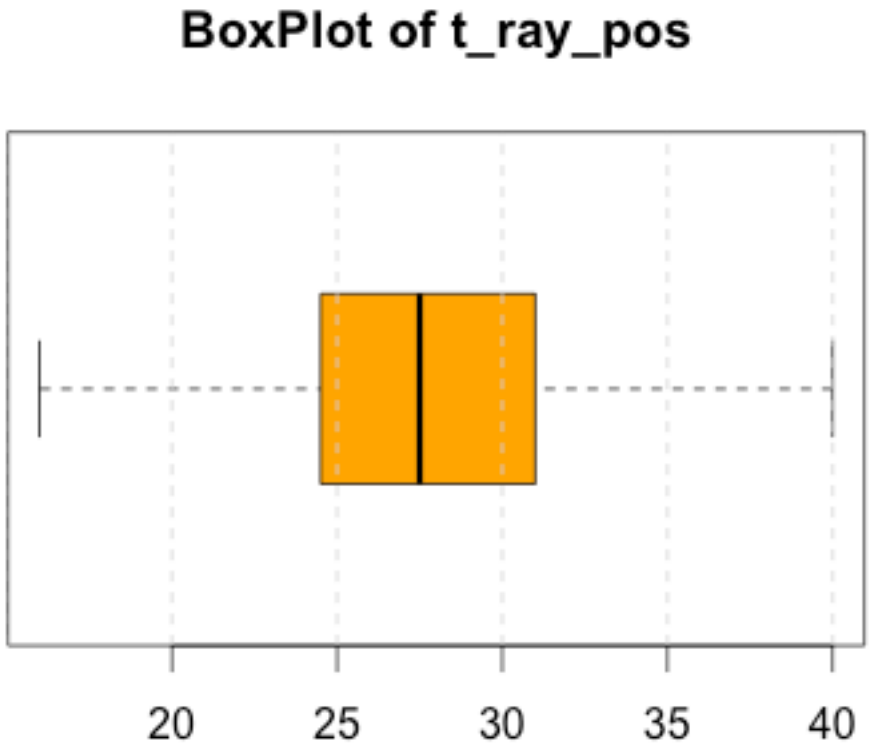
```
low_atyp_q=quantile(t_ray)[2] - 1.5*(quantile(t_ray)[4]-quantile(t_ray)
[2])
high_atyp_q=quantile(t_ray)[4] + 1.5*(quantile(t_ray)[4]-quantile(t_ray)
[2])
```

Por tanto, cualquier valor inferior a 13,875 (low_atyp_q) o superior a 40.875 (high_atyp_q) sería considerado atípico.

Eliminaremos de la muestra dichos puntos atípicos y volveremos a pintar nuestros diagramas.

```
hist(t_ray_pos,15,
     main="Histogram of t_ray",
     xlab="Time",
     ylab="Frequency",
     col="orange")
abline(h=seq(0,10,10), lty=2, col="lightgrey")
abline(v=seq(0,40,10), lty=2, col="lightgrey")
```






```
boxplot(t_ray_pos, horizontal=TRUE, main="BoxPlot of t_ray_pos",
col="orange")
abline(v=seq(0,40,5), lty=2, col="lightgrey")
```

En este punto tendríamos ya un conjunto bastante más estable. Muestra de ello es que la media de los datos sería ahora bastante próxima a la mediana (27.75 vs 27.5) y tendríamos una desviación típica de 5.083; que no está tan mal (aunque podría ser mejor) teniendo en cuenta que las mediciones se realizaron en el siglo XIX.

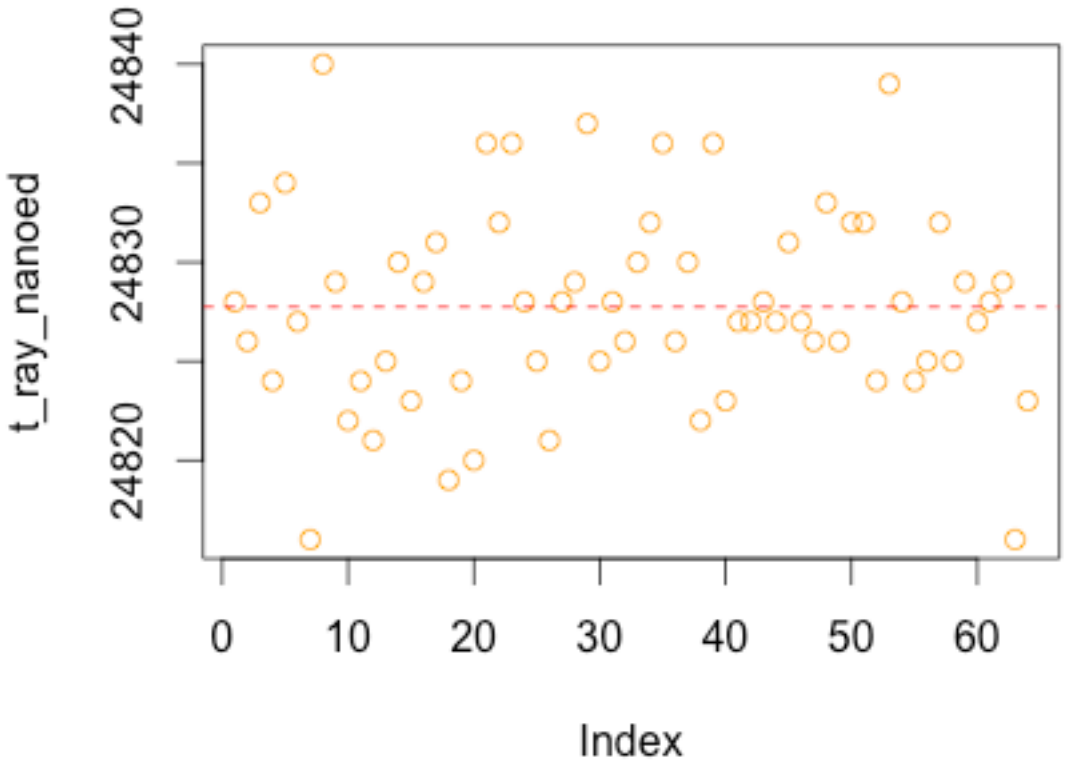
Para ser capaces de dar un valor en kilómetros/segundo o primero que tenemos que hacer es decodificar los datos. Una vez decodificados y puesto que la muestra carece ya de valores atípicos; utilizaremos la media para realizar el cálculo y pintaremos la muestra así como una línea roja mostrando la media.

```
t_ray_to_nano <- function(x) x+24800
t_ray_nanoed <- t_ray_to_nano(t_ray_pos)

plot(t_ray_nanoed, col="orange")
abline(h=mean(t_ray_nanoed), lty=2, col="red")
```

Finalmente, calculamos el valor solicitado en kilómetros/segundo:

```
dist_km = 7400/1000
speed=dist_km/(mean(t_ray_nanoed)*10^-9)
```



speed

```
[1] 298053.6
```

(kilometros/segundo)

Ejercicio 2: Mamíferos

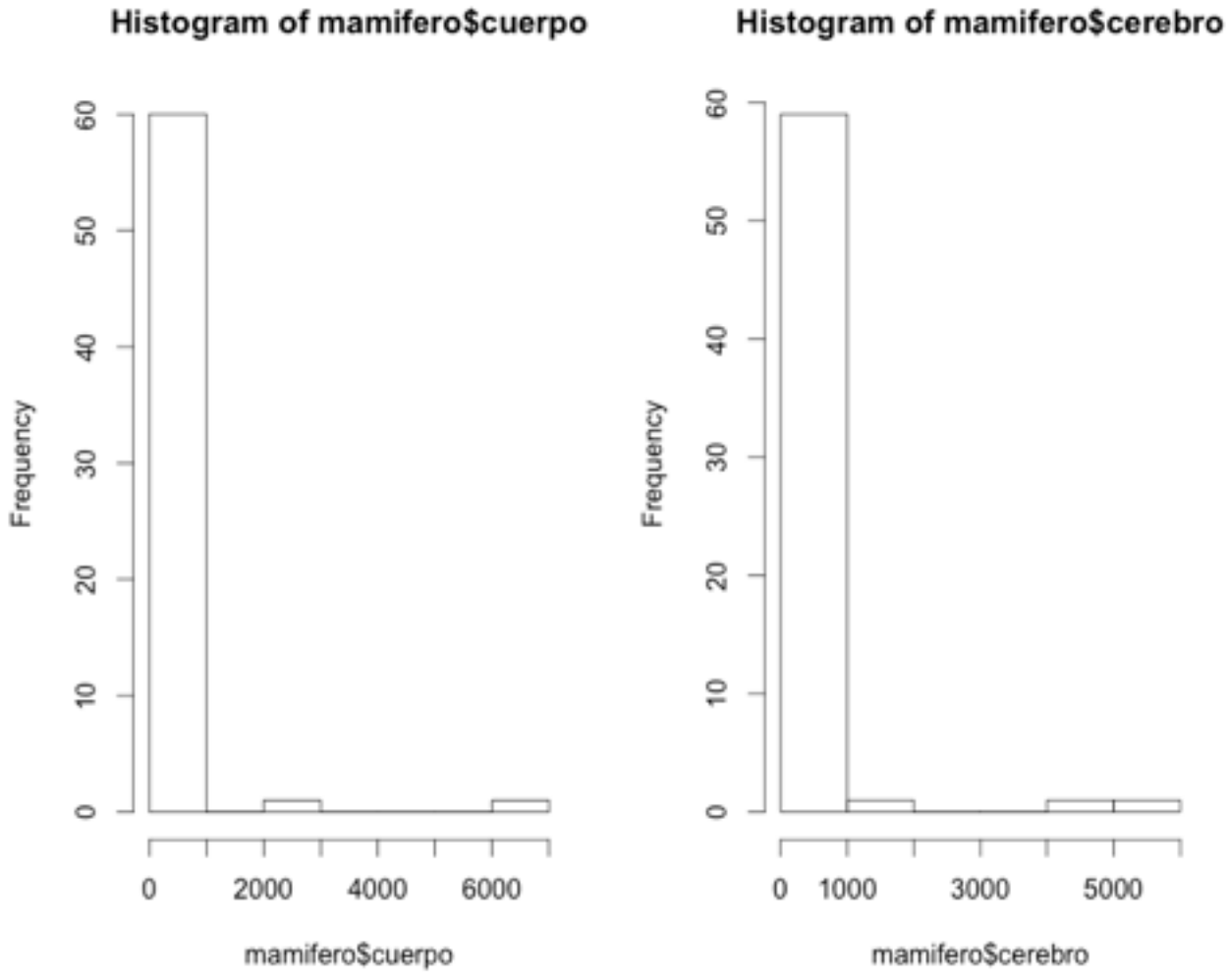
Carga de datos:

```
load("./data/mamifero")
```

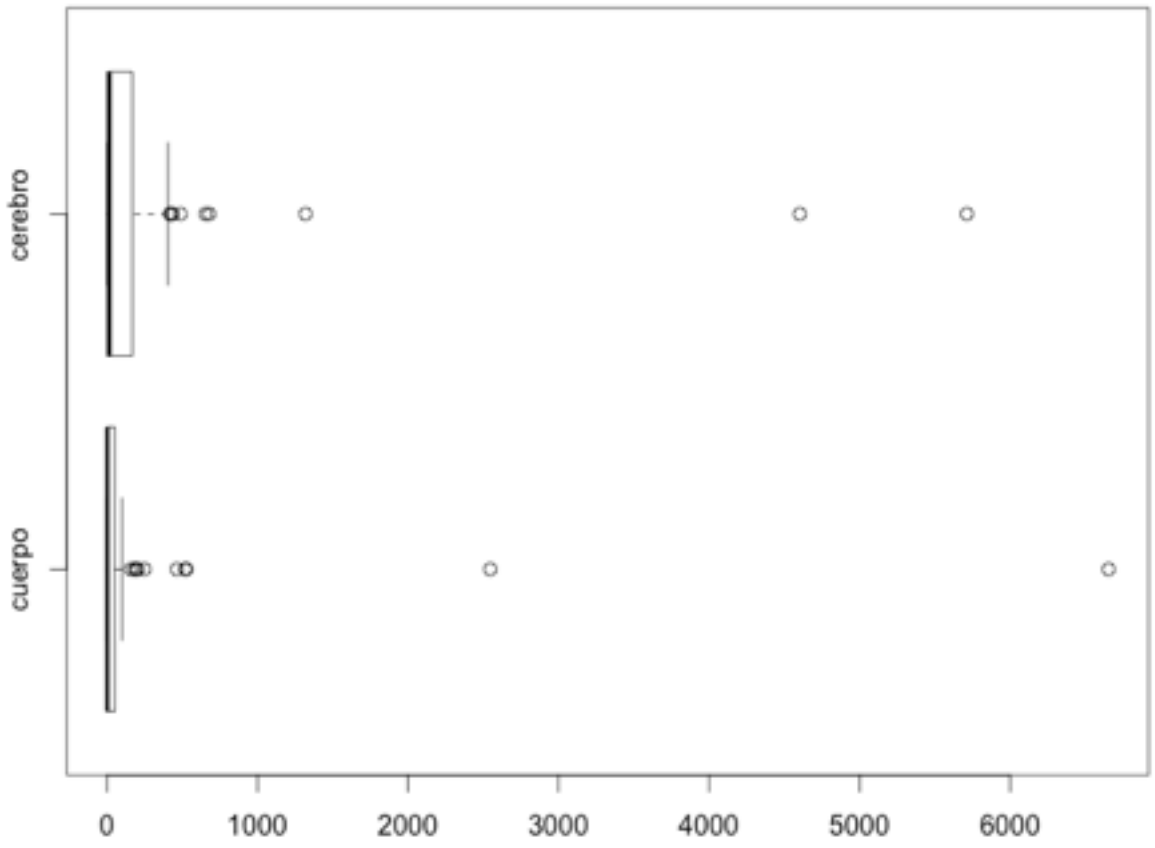
Análisis descriptivo:

Tras echarle un vistazo a los datos contenidos en el dataframe facilitado, vemos que nos encontramos ante una muestra de dos variables contínuas cuantitativas y numéricas, peso en kilogramos del cuerpo de un mamífero; y peso en gramos del cerebro de dicho mamífero. Podemos ver el histograma para determinar que la mayor parte de la muestra toma valores inferiores a 1000 (ya sean kg o gramos).

```
par(mfrow=c(1,2))
hist(mamifero$cuerpo)
hist(mamifero$cerebro)
```

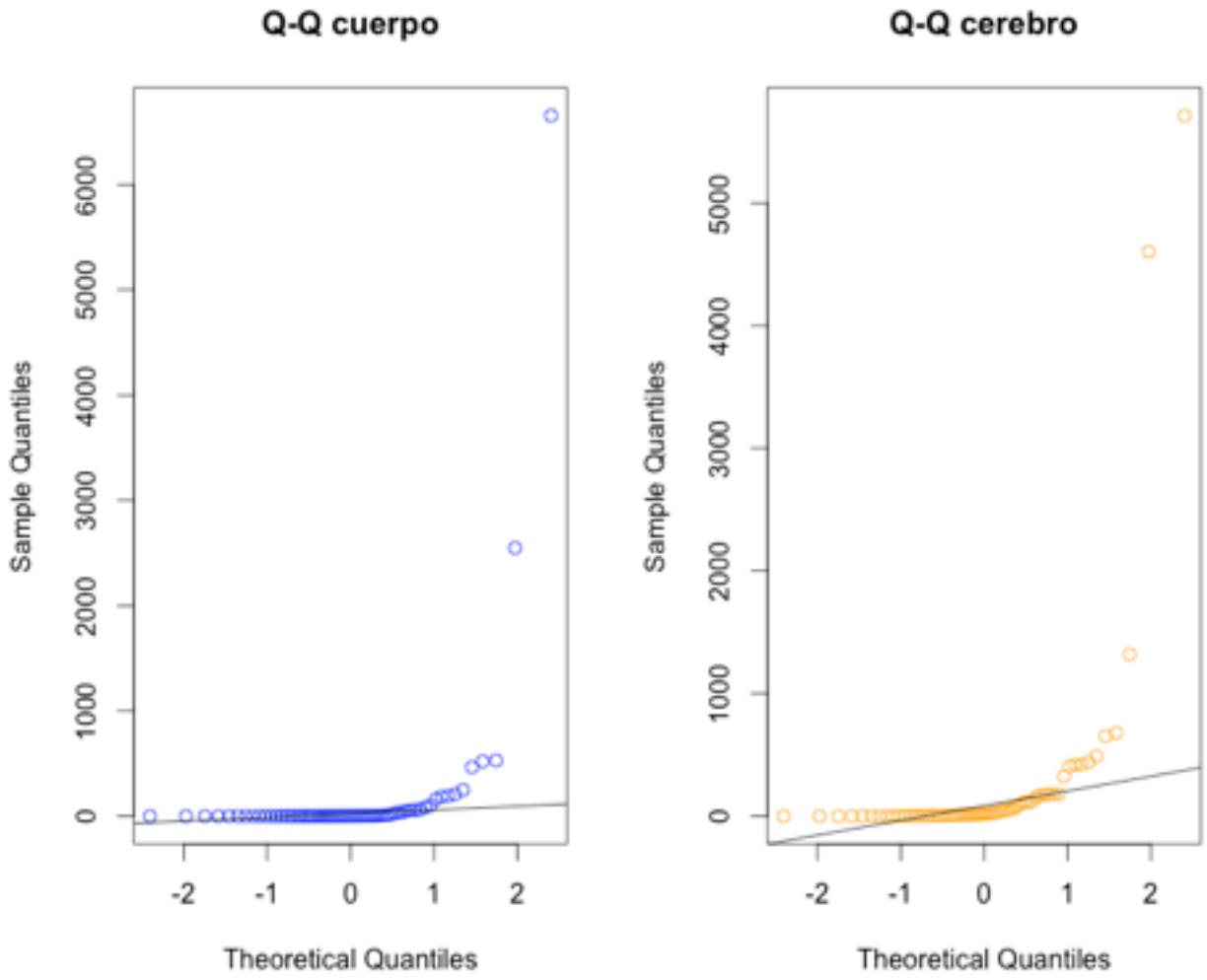


Si pintamos el diagrama de cajas, veremos que la muestra está muy concentrada, aunque aparecen puntos “muy” atípicos (posiblemente los correspondientes a mamíferos grandes, como una ballena o un elefante).



Estas desviaciones sustanciales se ven claramente en el gráfico Q-Q, con colas a la derecha y, que muestra para ambas variables una forma bastante exponencial, indicando anormalidad.

```
par(mfrow=c(1,2))
qqnorm(mamifero$cuerpo, main="Q-Q cuerpo", col="blue")
qqline(mamifero$cuerpo)
qqnorm(mamifero$cerebro, main="Q-Q cerebro", col="orange")
qqline(mamifero$cerebro)
```



Dependencia:

La gráfica Q-Q anterior, además de darnos una idea de la localidad relativa de los datos y su escala, nos mostraba también una semejanza entre ambas que nos hace pensar que las variables pudiesen ser fuertemente dependientes. Comprobémoslo.

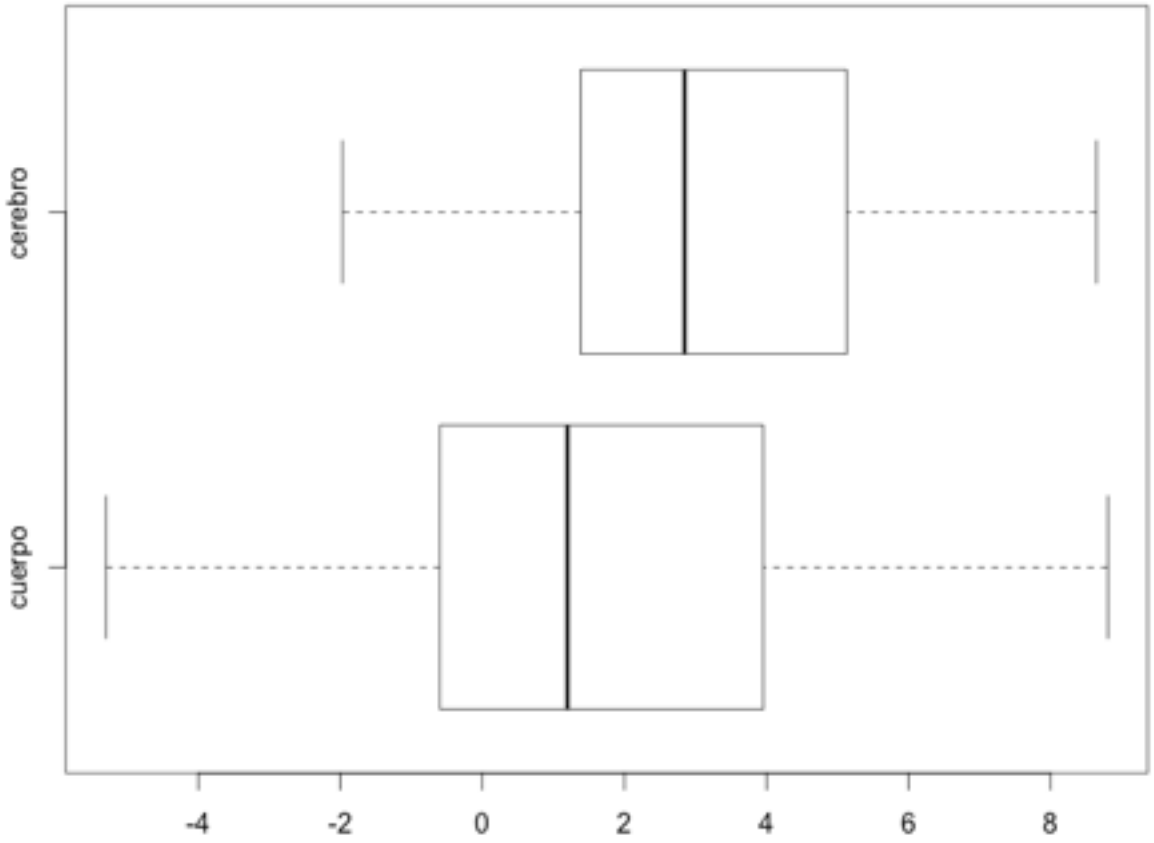
La correlación nos muestra una dependencia fuerte y positiva entre ambas variables.

```
> cor(mamifero)
            cuerpo   cerebro
cuerpo  1.0000000 0.9341639
cerebro 0.9341639 1.0000000
```

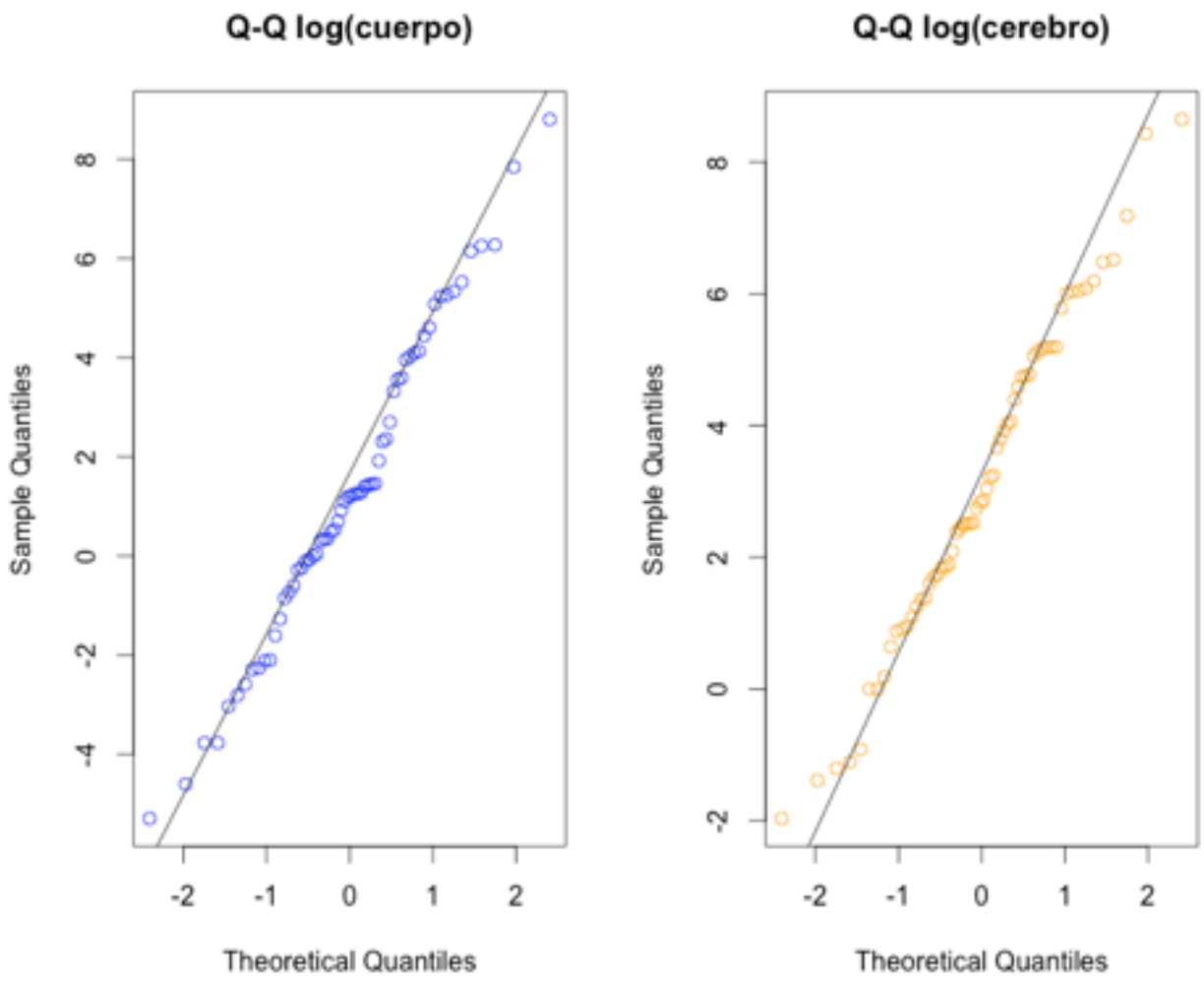
Posibles transformaciones:

La forma exponencial de la gráfica Q-Q nos lleva a pensar que tal vez una transformación logarítmica en los datos pueda normalizar la muestra bastante bien. Tras realizar la transformación, volveremos a pintar el diagrama de cajas y el Q-Q.

```
#Transformacion
lmamifero <- log(mamifero)
par(mfrow=c(1,1))
boxplot(lmamifero, horizontal=TRUE)
```




```
par(mfrow=c(1,2))
qqnorm(lmamifero$cuerpo, main="Q-Q log(cuerpo)", col="blue")
qqline(lmamifero$cuerpo)
qqnorm(lmamifero$cerebro, main="Q-Q log(cerebro)", col="orange")
qqline(lmamifero$cerebro)
```



Se aprecia una enorme mejoría en la normalidad y conseguimos además una mayor dependencia entre las variables.

```
> cor(lmamifero)
            cuerpo   cerebro
cuerpo  1.0000000 0.9595525
cerebro 0.9595525 1.0000000
```

Identificación “hombre”:

Para intentar identificar al hombre partiremos del conocimiento previo que nos ha dado la experiencia. Partimos de la base de que nuestros hábitos nos tienen más acostumbrados a realizar mediciones del peso de nuestro cuerpo que de nuestro cerebro. Por este motivo, podríamos intentar realizar un primer filtro de la muestra basándonos en lo que creemos que puede ser un peso standard de un ser humano (entre 50 y 90 kg). Desconocemos los hábitos alimenticios del ser humano medido, pero podemos suponer que la medida será de un hombre común, y no de uno con un claro sobrepeso.

Dicho esto, filtraremos la media en base a esta suposición:

```
> possible_humans <- mamifero[mamifero$cuerpo > 50 & mamifero$cuerpo
<90,]
> possible_humans
   cuerpo cerebro
30  85.00     325
32  62.00    1320
44  55.50     175
46  52.16     440
49  60.00      81
```

Tras reducir drásticamente la población de nuestra muestra, resulta mucho más sencillo intentar seleccionar cuál creemos que será el individuo que representa al ser humano. Me costaría bastar pensar que un cerebro pesa menos de medio kilo, así que me atrevería a decir que probablemente la medida correspondiente al hombre sea la medida 32 (62kg de cuerpo y 1320g de cerebro).

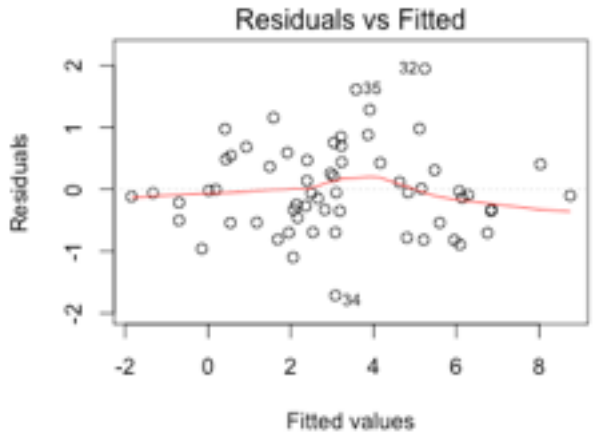
Mejora de dependencia:

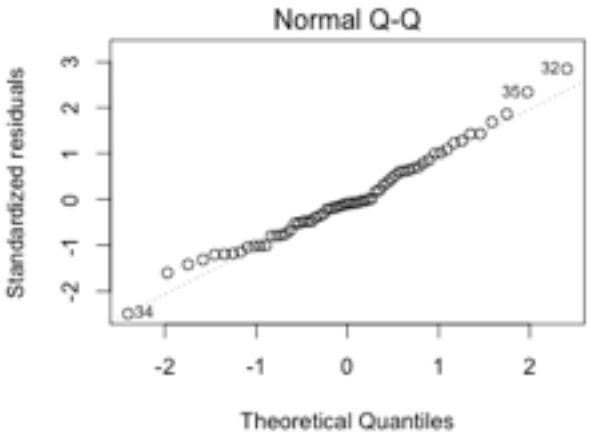
Para mejorar la dependencia, podríamos eliminar los valores de la muestra que presentan menor normalidad, es decir, aquellos puntos que nuestro modelo lineal explique peor. El modelo lineal siguiente tiene un R-Squared de 0.92 indicando que el 92% de la variabilidad de los datos quedaría explicada por el modelo (a diferencia de un 8% que no podría explicar).

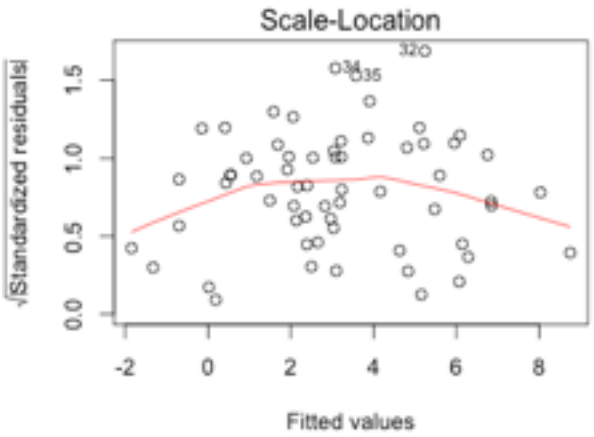
```
par(mfrow=c(2,2))
lm_lmamifero <- lm(lmamifero$cerebro ~ lmamifero$cuerpo)
summary(lm_lmamifero)
plot(lm_lmamifero)
```

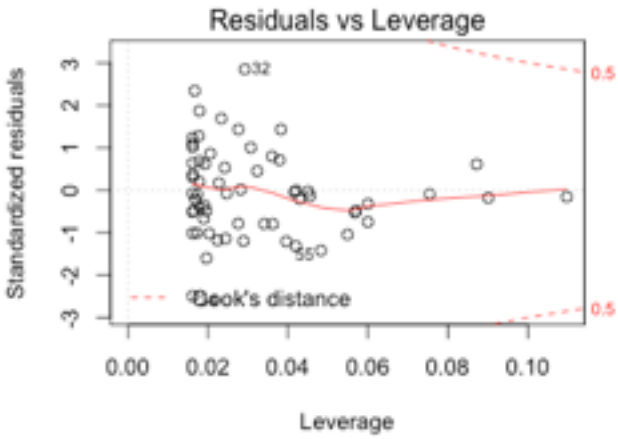
En la siguientes gráficas veremos que efectivamente el modelo parece que puede ser explicado bien linealmente. Además, en la gráfica que muestra la distancia de Cook, llama la atención el punto 55, que no aparece en las gráficas anteriores y que representa una gran influencia en el ajuste. No obstante, eliminarlo no mejoraría la dependencia puesto que no se trata de un punto atípico. Sí lo serían los puntos 32, 34 y 35 y su eliminación, vemos que hace que aumente la dependencia entre ambas variables.

```
trunc_lmamifero <- lmamifero[c(-32,-34,-35),]
cor(trunc_lmamifero)
            cuerpo   cerebro
cuerpo  1.0000000 0.9710383
cerebro 0.9710383 1.0000000
```


Residuals vs Fitted
Residuals
Fitted values
320
35
34


Normal Q-Q
Standardized residuals
Theoretical Quantiles
320
350
34


Scale-Location
√|Standardized residuals|
Fitted values
320
34
35


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
32
55
0.5
0.5