Construccion de un Modelo Estadistico Base

Ariadna Jocelyn Guzmán Jiménez - A01749373

2022-09-18

```
knitr::opts_chunk$set(echo = FALSE)
```

Módulo 1: Estadística para ciencia de datos

Inteligencia artificial avanzada para la ciencia de datos I

Grupo 101

Resumen

Mediante este trabajo, se presenta el análisis e implementación del modelo de regresión lineal y ANOVA para la base de contaminación por mercurio en lagos. Esta problemática, llega a ser muy importante ya que además de afectar a los seres vivos del lago, puede afectar a los seres humanos si llegan a comer alguno de ellos. Por eso, fue necesario poder realizar un entendimiento de todos los datos para poder enfatizar y lograr a interpretar cuales son las variables candidatas para poder realizar los modelos.

Introducción

La contaminación por mercurio de peces en el agua dulce comestibles es una amenaza directa contra nuestra salud. Se llevó a cabo un estudio reciente en 53 lagos de Florida con el fin de examinar los factores que influían en el nivel de contaminación por mercurio.

En nuestra base de datos, encontramos los siguiente atributos:

- **X1** = número de indentificación
- **X2** = nombre del lago
- **X3** = alcalinidad (mg/l de carbonato de calcio)
- **X4** = PH
- **X5** = calcio (mg/l)
- **X6** = clorofila (mg/l)
- **X7** = concentración media de mercurio (parte por millón) en el tejido muscualar del grupo de peces estudiados en cada lago
- **X8** = número de peces estudiados en el lago
- **X9** = mínimo de la concentración de mercurio en cada grupo de peces
- **X10** = máximo de la concentración de mercurio en cada grupo de peces
- **X11** = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible)
- **X12** = indicador de la edad de los peces (0: jóvenes; 1: maduros)

Dadas las descripciones de cada variable, nos surgen interesantes las siguientes preguntas para poder resolver y poder ir sobre ellas para poder hacer predicciones futuras, las cuales son:

- *¿Habrá direfencia significativa entre la concentración de mercurio por la edad de los peces?*
- *¿Hay evidencia para suponer que la concentración promedio de mercurio en los lagos es dañino para la salud humana?*
- *¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida?*

En las siguientes líneas, se verá la implementación de modelos para la resolución de las preguntas anteriores y llegar a una conclusión concreta de esta problemática.

Análisis de los resultados

Lectura de datos

En la parte de la lectura de datos, importamos nuestra base y por otra parte, hacemos una nueva variable que solo cuente con los datos númericos, ya que son los que nos servirán para nuestros modelos. Se omiten solamente la columna 1 y 2 ya que la primera sólo es un id para cada fila de datos y la segunda es el nombre del lago.

Entendimiento de los datos

Dado lo anterior visualizamos medidas generales de nuestro dataset:

Dimensión del dataset

```
## Número de filas: 53
```

```
##
## Número de columnas: 12
```

```
##
## Cantidad de datos: 636
```

Cálculo de medidas estadísticas

Análisis estadístico

```
##        X1          X2                  X3               X4
##  Min.   : 1   Length:53          Min.   :  1.20   Min.   :3.600
##  1st Qu.:14   Class :character   1st Qu.:  6.60   1st Qu.:5.800
##  Median :27   Mode  :character   Median : 19.60   Median :6.800
##  Mean   :27                      Mean   : 37.53   Mean   :6.591
##  3rd Qu.:40                      3rd Qu.: 66.50   3rd Qu.:7.400
##  Max.   :53                      Max.   :128.00   Max.   :9.100
##        X5             X6               X7               X8
##  Min.   : 1.1   Min.   :  0.70   Min.   :0.0400   Min.   : 4.00
##  1st Qu.: 3.3   1st Qu.:  4.60   1st Qu.:0.2700   1st Qu.:10.00
##  Median :12.6   Median : 12.80   Median :0.4800   Median :12.00
```

```
##  Mean   :22.2   Mean   : 23.12   Mean   :0.5272   Mean   :13.06
##  3rd Qu.:35.6   3rd Qu.: 24.70   3rd Qu.:0.7700   3rd Qu.:12.00
##  Max.   :90.7   Max.   :152.40   Max.   :1.3300   Max.   :44.00
##        X9              X10              X11              X12
##  Min.   :0.0400   Min.   :0.0600   Min.   :0.0400   Min.   :0.0000
##  1st Qu.:0.0900   1st Qu.:0.4800   1st Qu.:0.2500   1st Qu.:1.0000
##  Median :0.2500   Median :0.8400   Median :0.4500   Median :1.0000
##  Mean   :0.2798   Mean   :0.8745   Mean   :0.5132   Mean   :0.8113
##  3rd Qu.:0.3300   3rd Qu.:1.3300   3rd Qu.:0.7000   3rd Qu.:1.0000
##  Max.   :0.9200   Max.   :2.0400   Max.   :1.5300   Max.   :1.0000
```

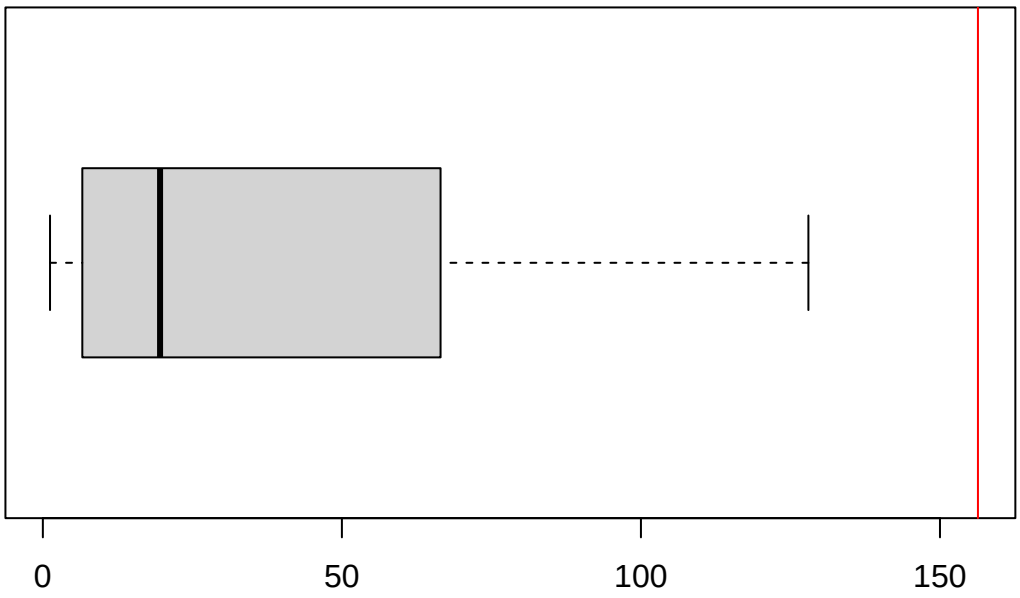
	Minimo	Q1	Mediana	Media	Q3	Maximo	Desviacion Estandar	Moda
X3	1.20	6.60	19.60	37.5301887	66.50	128.00	38.2035267	17.30
X4	3.60	5.80	6.80	6.5905660	7.40	9.10	1.2884493	5.80
X5	1.10	3.30	12.60	22.2018868	35.60	90.70	24.9325744	3.00
X6	0.70	4.60	12.80	23.1169811	24.70	152.40	30.8163214	1.60
X7	0.04	0.27	0.48	0.5271698	0.77	1.33	0.3410356	0.34
X8	4.00	10.00	12.00	13.0566038	12.00	44.00	8.5606773	12.00
X9	0.04	0.09	0.25	0.2798113	0.33	0.92	0.2264058	0.04
X10	0.06	0.48	0.84	0.8745283	1.33	2.04	0.5220469	0.06
X11	0.04	0.25	0.45	0.5132075	0.70	1.53	0.3387294	0.16
X12	0.00	1.00	1.00	0.8113208	1.00	1.00	0.3949977	1.00

La función summary aplicada para la base de datos, nos permitió visualizar un resumen general en medidas estadísticas de cada uno de los datos almacenados en la base “mercurio”, sin embargo, se realizó posteriormente un dataframe que almacen 2 medidas más que no están consideradas en summary, desplegando de una forma más visual e interactiva el cálculo de cada uno de los datos.

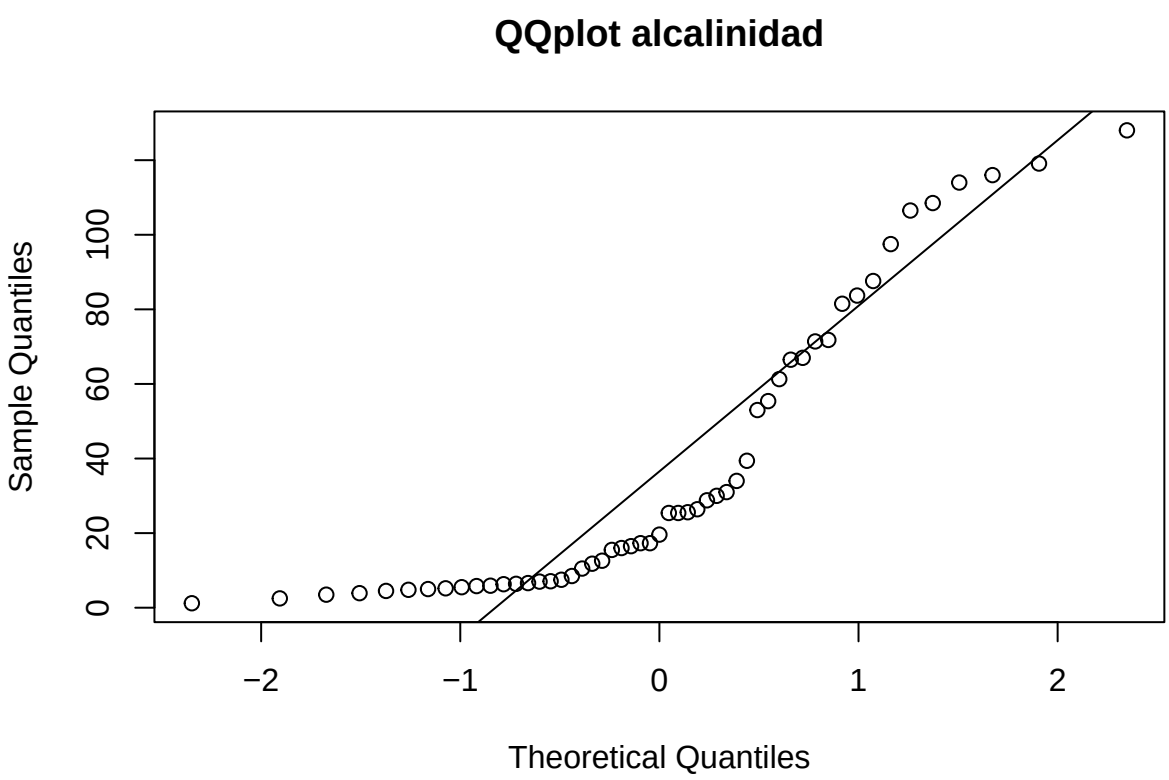
Visualización de datos

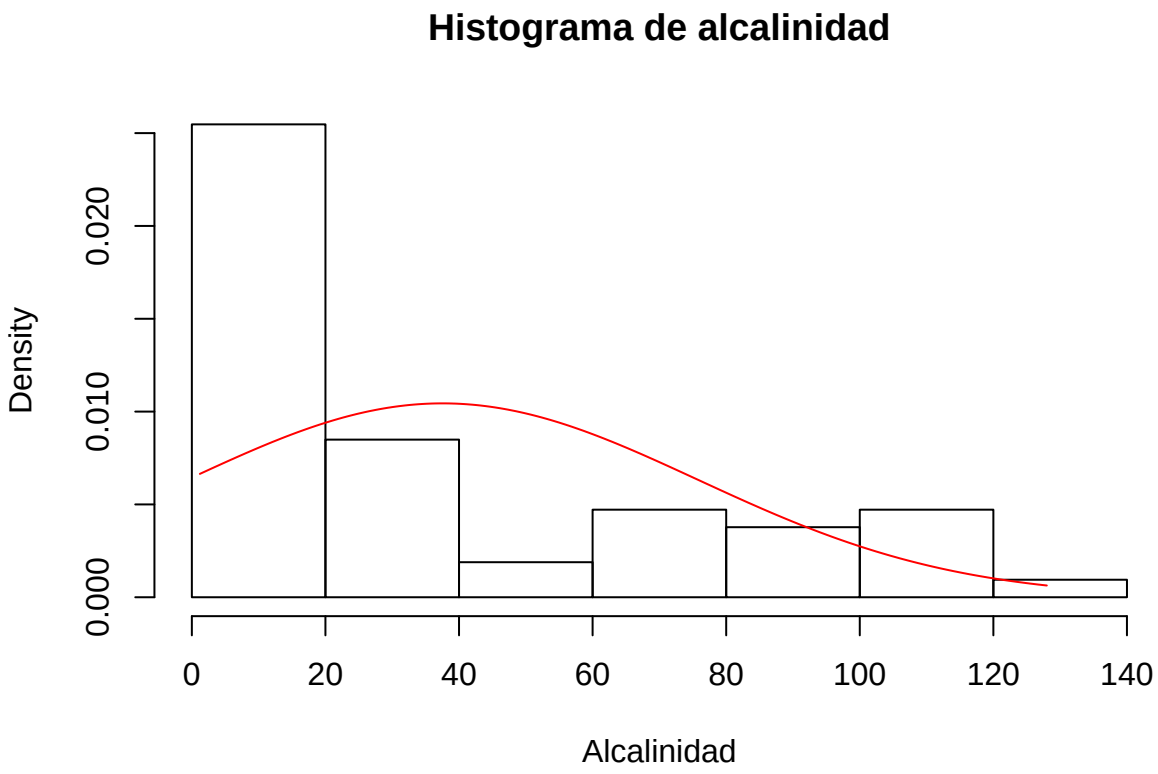
Variables cuantitativas*

- **X3**



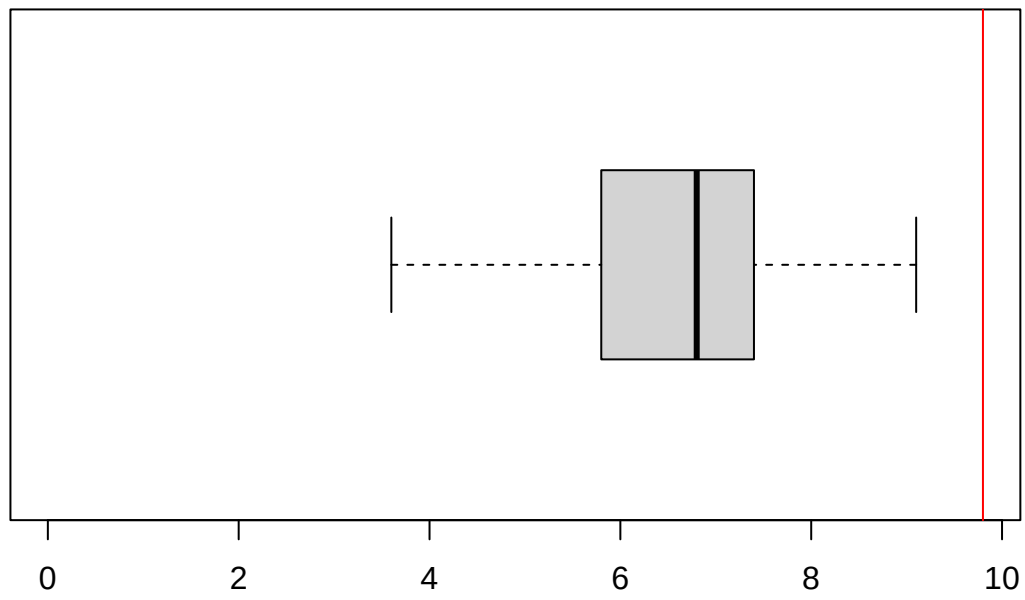

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.20    6.60   19.60   37.53   66.50  128.00
```


QQplot alcalinidad
Sample Quantiles
0
20
40
60
80
100
−2
−1
0
1
2
Theoretical Quantiles

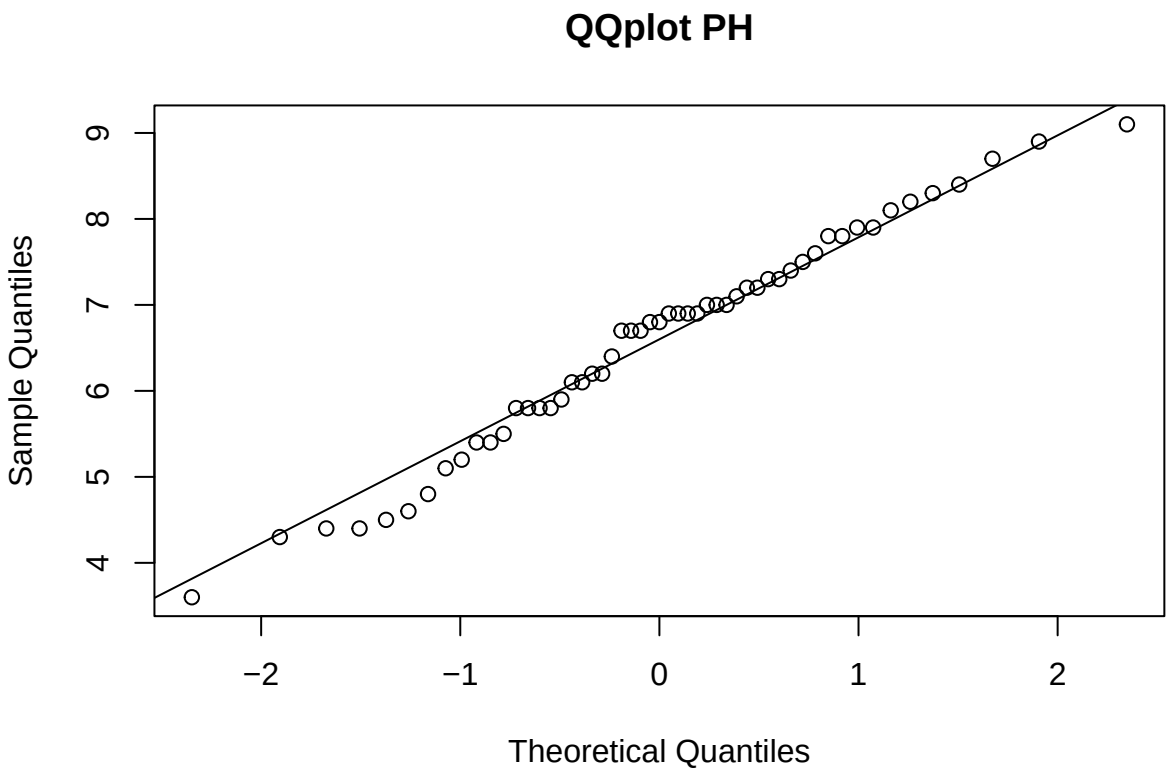


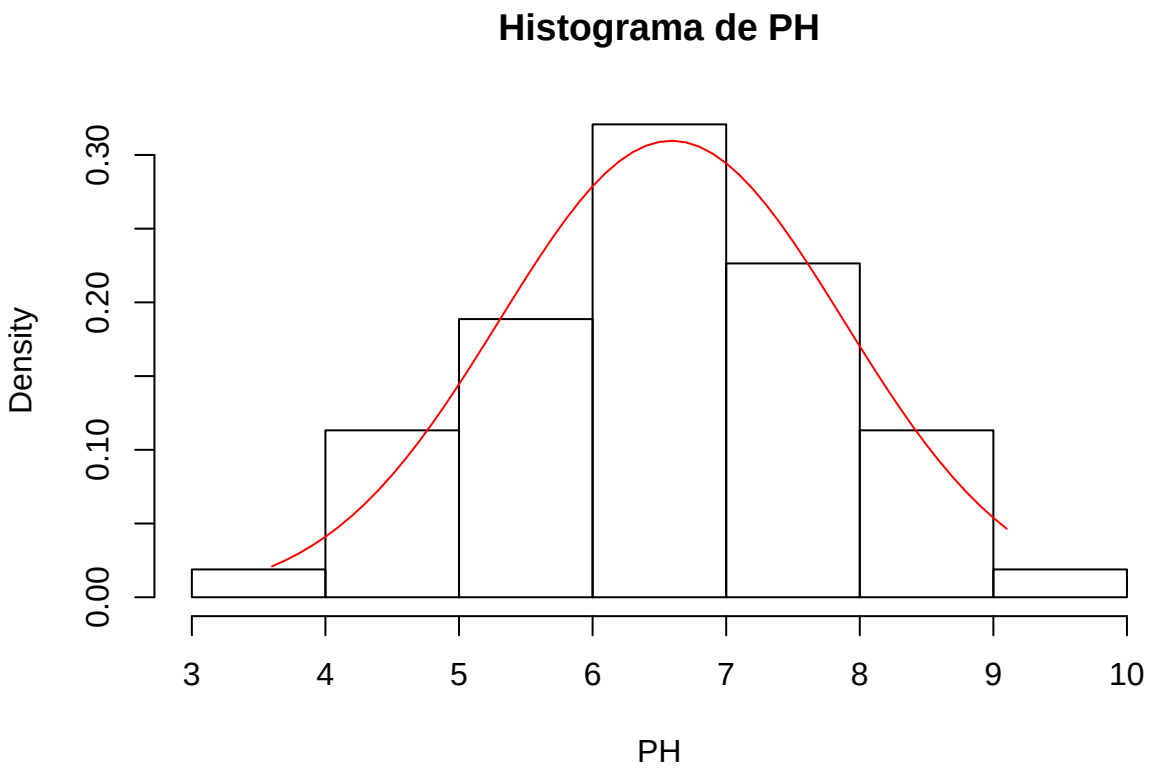
Tanto el histograma como el boxplot de alcalinidad, nos muestran una concentración de los datos hacia la izquierda, logrando visualizar un sesgo a la derecha indicando asimetría de la distribución con respecto a la media y mediana. Finalmente, de acuerdo al QQplot, se muestra una curtosis alta con una distribución leptocúrtica.

- **X4**



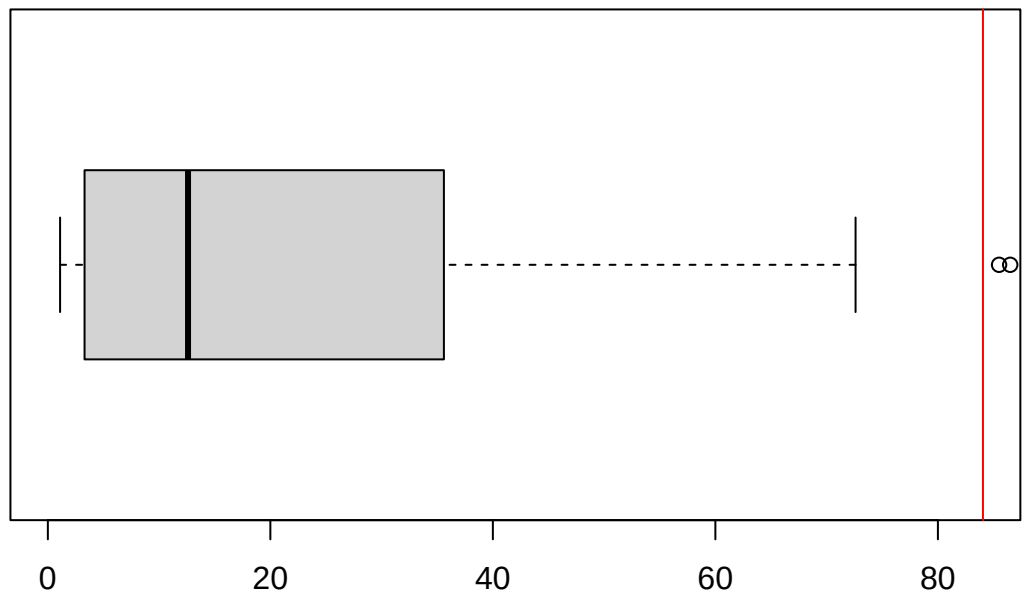

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.600   5.800   6.800   6.591   7.400   9.100
```


QQplot PH
Sample Quantiles
Theoretical Quantiles
4
5
6
7
8
9
−2
−1
0
1
2

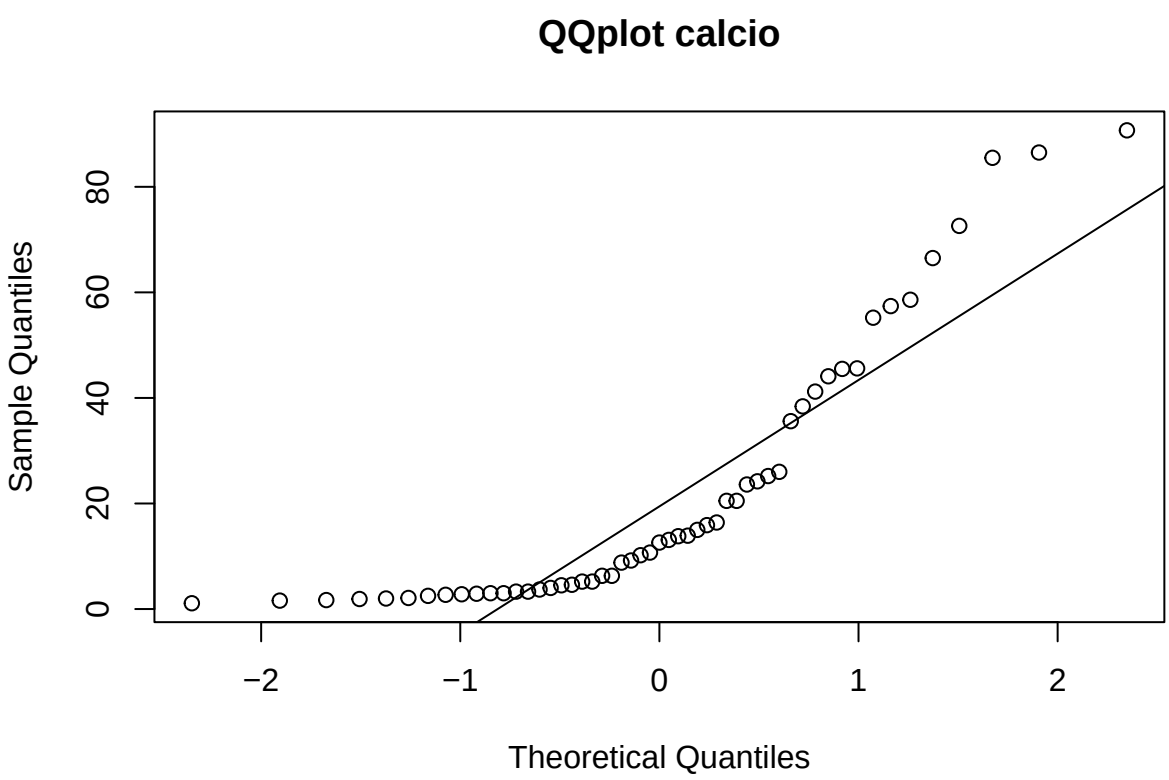


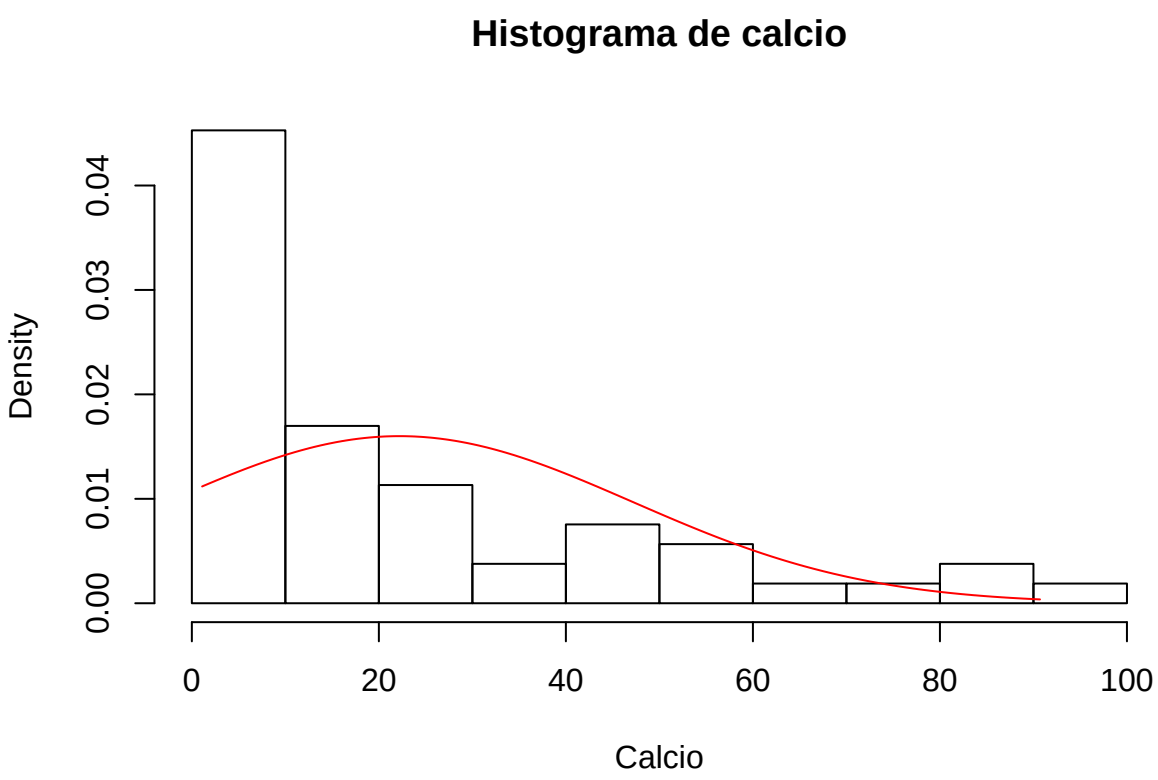
El histograma y el boxplot nos muestran una simetría en la distribución respecto a la media y mediana, logrando ver en el QQplot que la probabilidad normal es la ideal para la variable PH.

- **X5**



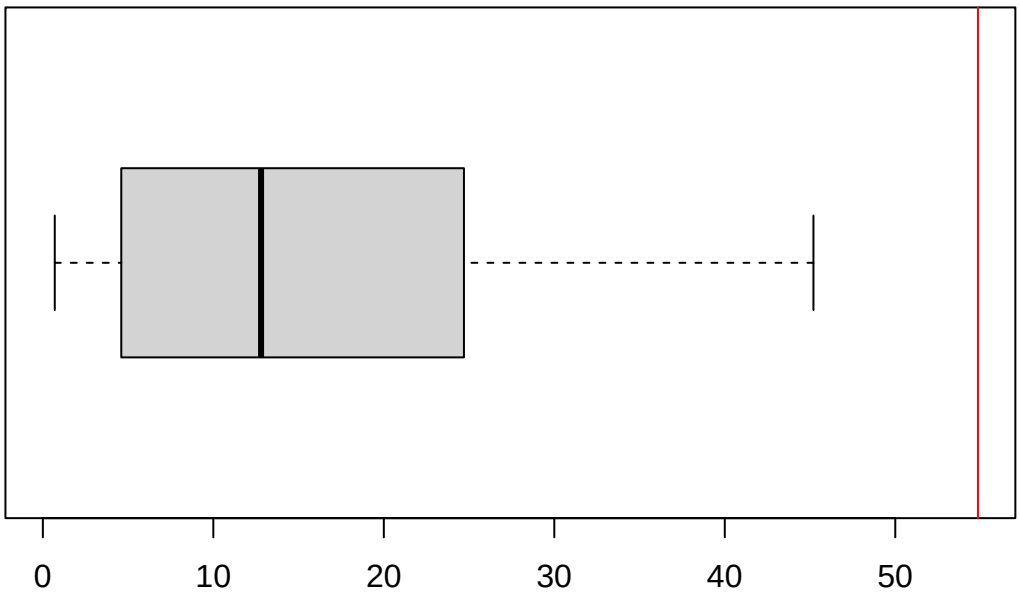

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.1     3.3    12.6    22.2    35.6    90.7
```


QQplot calcio
Sample Quantiles
0
20
40
60
80
−2
−1
0
1
2
Theoretical Quantiles

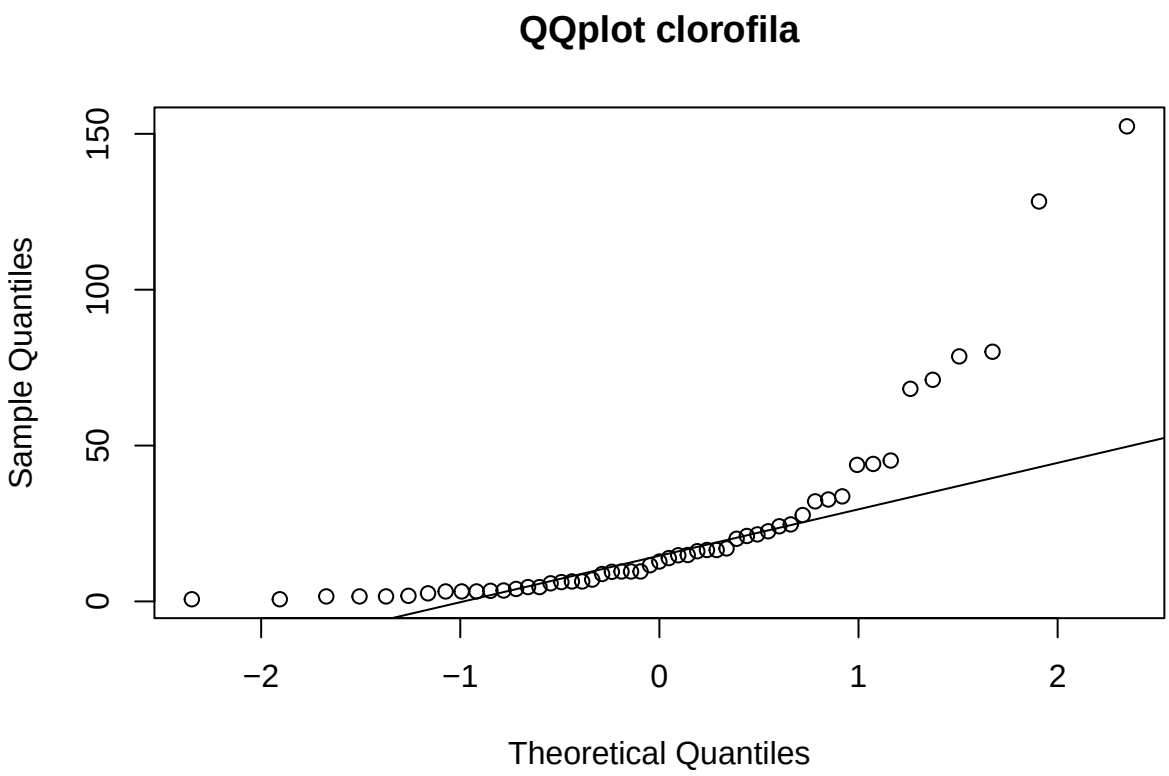


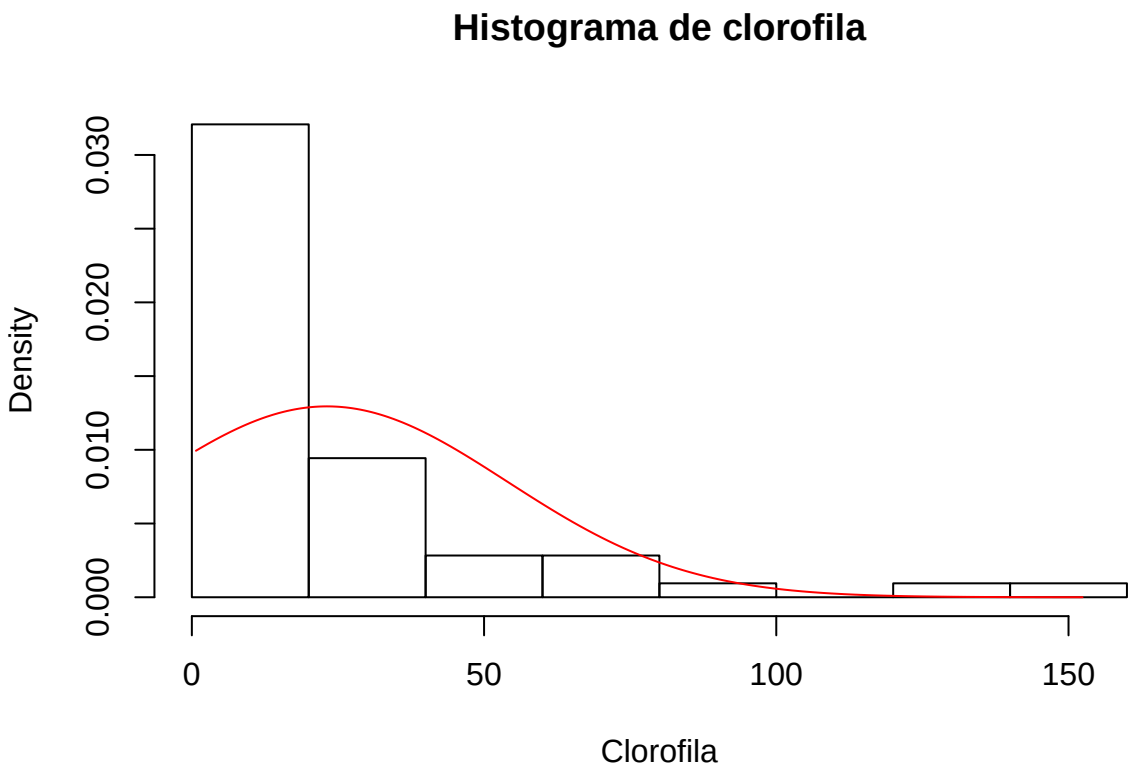
El histograma y el boxplot para el calcio, nos muestran una distribución cargada hacia la izquierda con respecto a la media y a la moda, viendo que se cuenta con un sesgo hacia la derecha y la presencia de datos atípicos y extremos, los cuales son diferentes a las demás observaciones deñ grupo de calcio. Por otra parte, en QQplot, nos lo confirma mostrando una asimetría positiva.

- **X6**



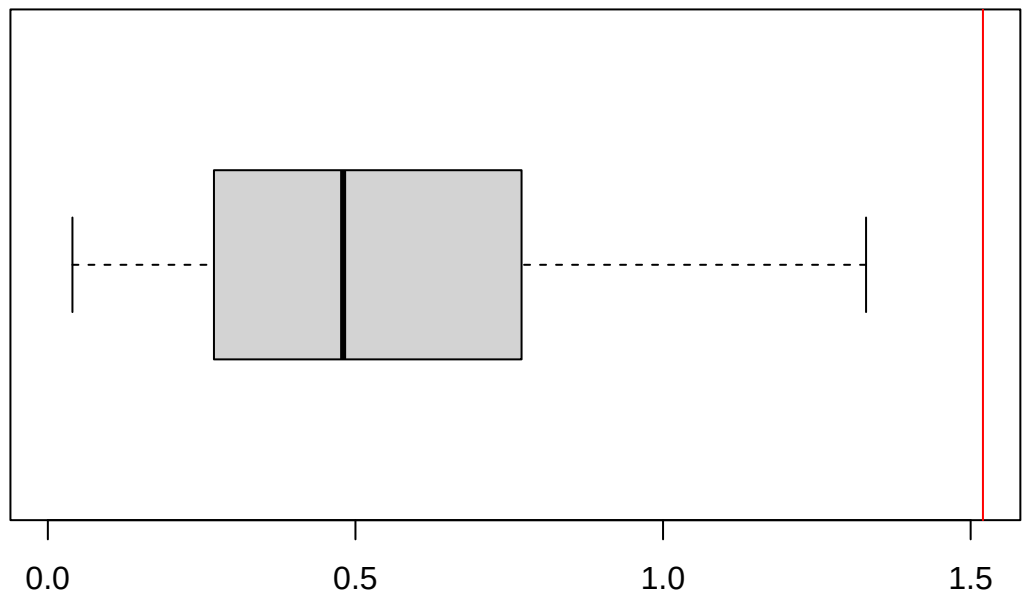

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.70    3.75    9.60   13.76   20.55   45.20
```


QQplot clorofila
Sample Quantiles
0
50
100
150
−2
−1
0
1
2
Theoretical Quantiles

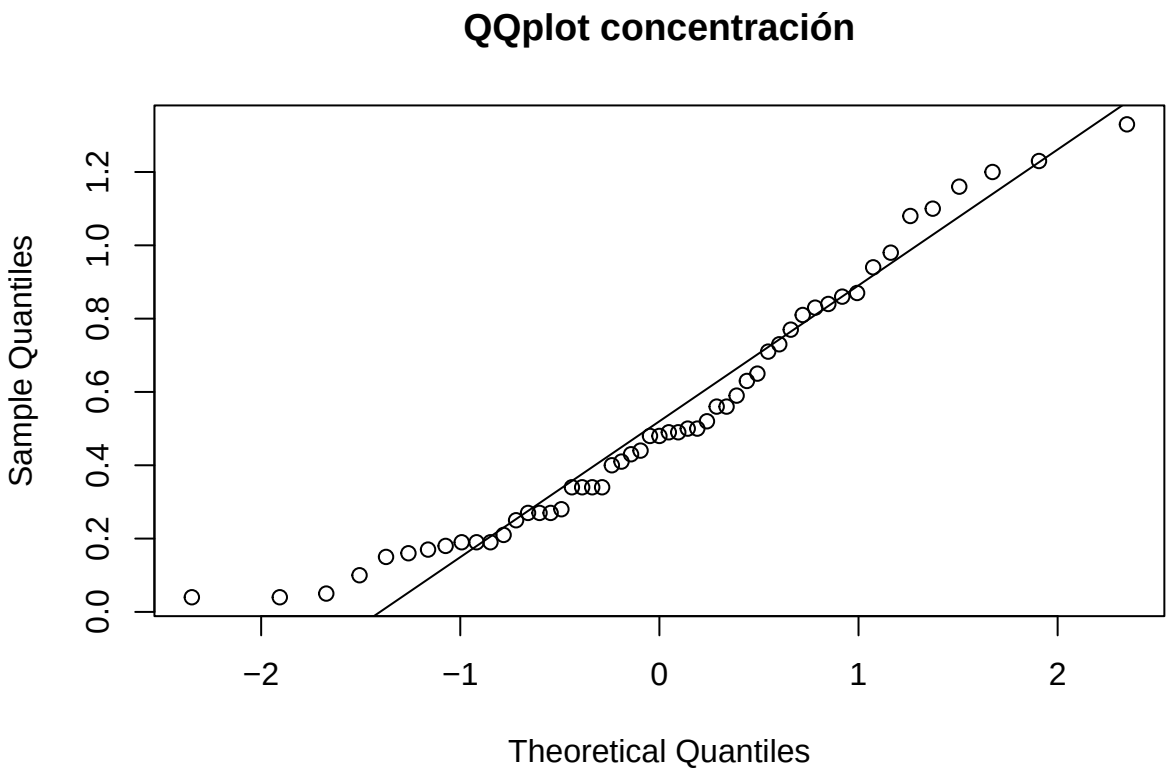


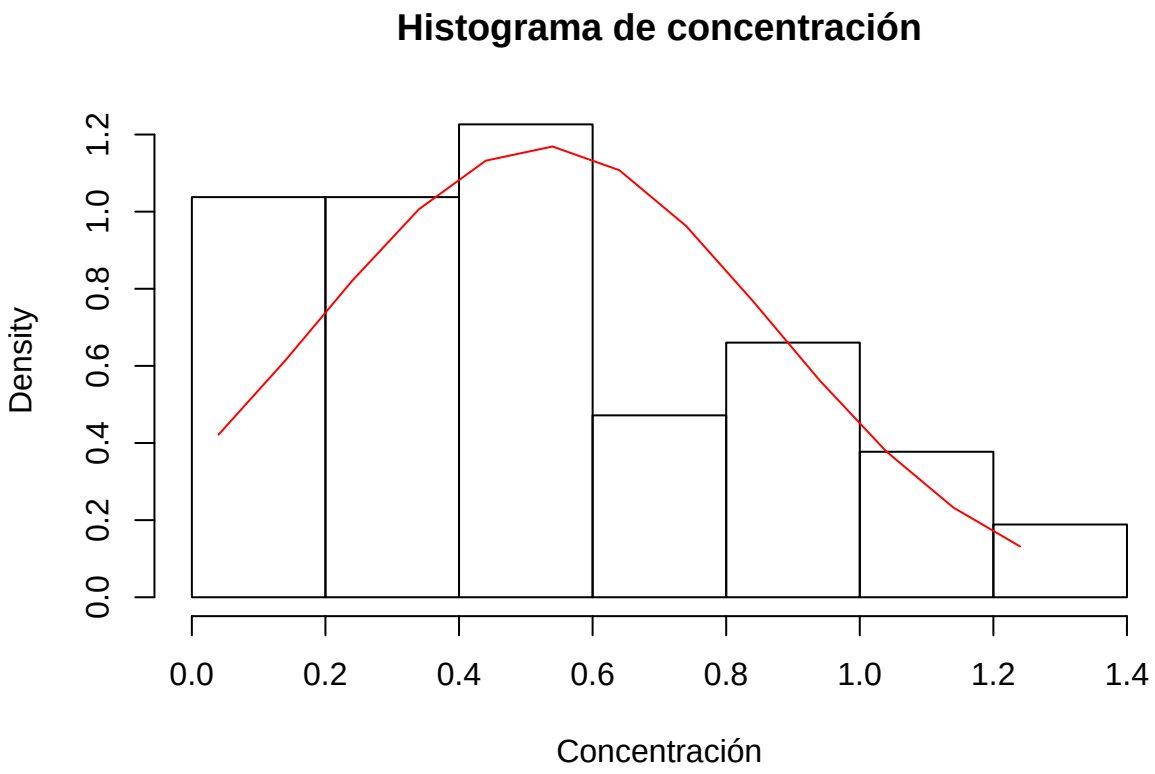
Para la clorofila, el caso es similar. Se nos muestra una distribución cargada hacia la izquierda con respecto a la media y a la moda, viendo que se cuenta con un sesgo hacia la derecha pero no cuenta con la presencia de datos atípicos. Por otra parte, en QQplot, nos lo confirma mostrando una asimetría positiva.

- **X7**



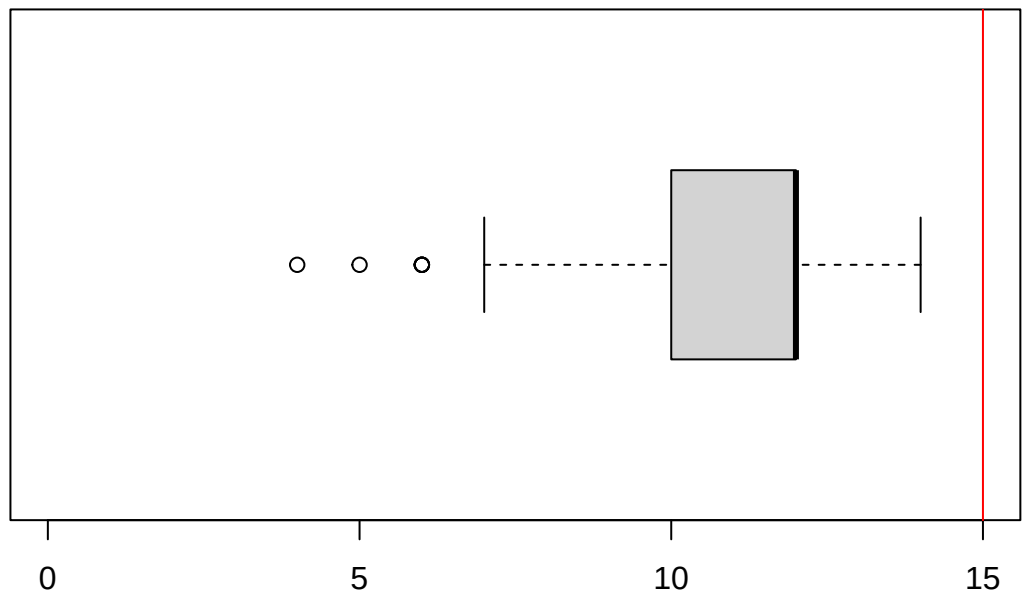

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0400  0.2700  0.4800  0.5272  0.7700  1.3300
```


QQplot concentración
Sample Quantiles
0.0
0.2
0.4
0.6
0.8
1.0
1.2
−2
−1
0
1
2
Theoretical Quantiles

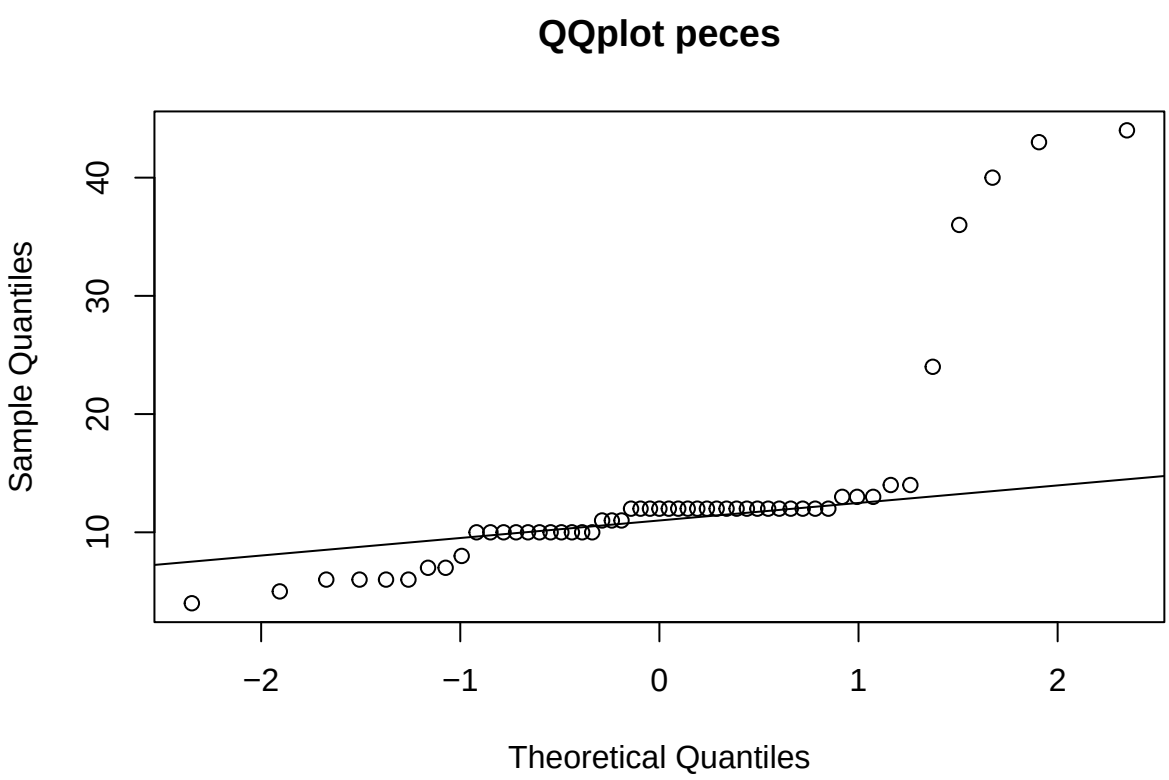


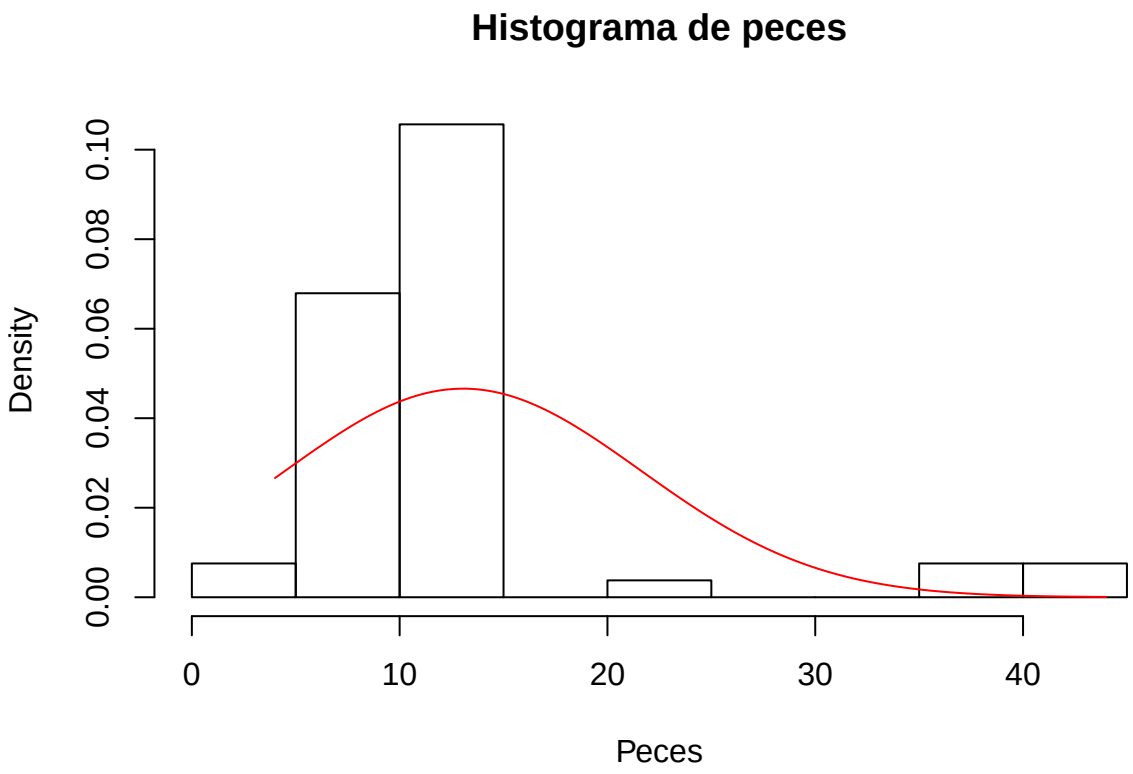
Para la concentración de la media, se muestra igualmetne una distribución cargada hacia la izquierda con respecto a la media y a la moda, sesgo hacia la derecha y sin la presencia de datos atípicos. Por otra parte, en QQplot, nos lo confirma mostrando una asimetría positiva.

- **X8**



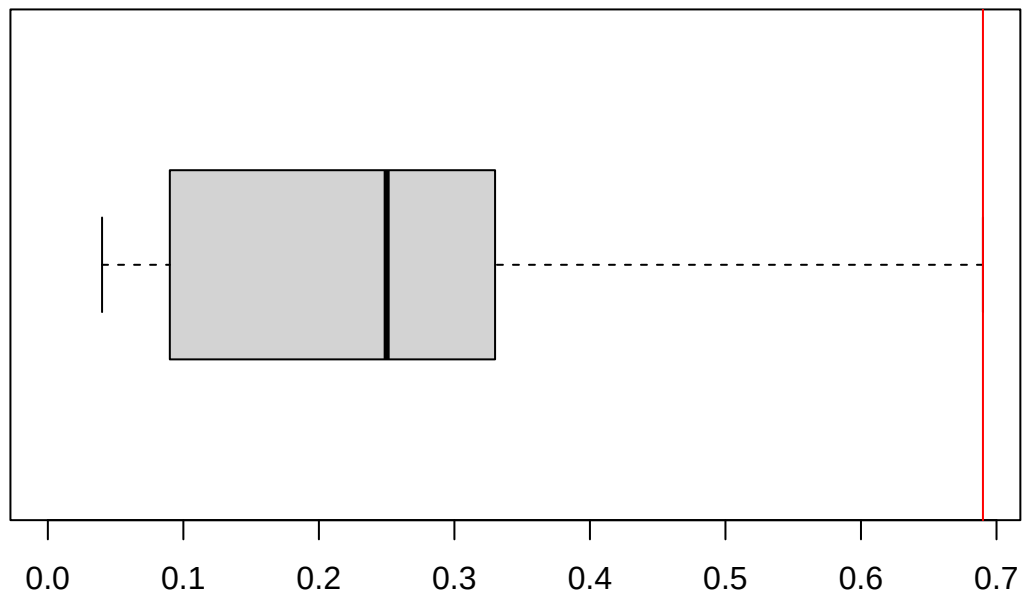

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00   10.00   12.00   10.52   12.00   14.00
```


QQplot peces
Sample Quantiles
10
20
30
40
−2
−1
0
1
2
Theoretical Quantiles

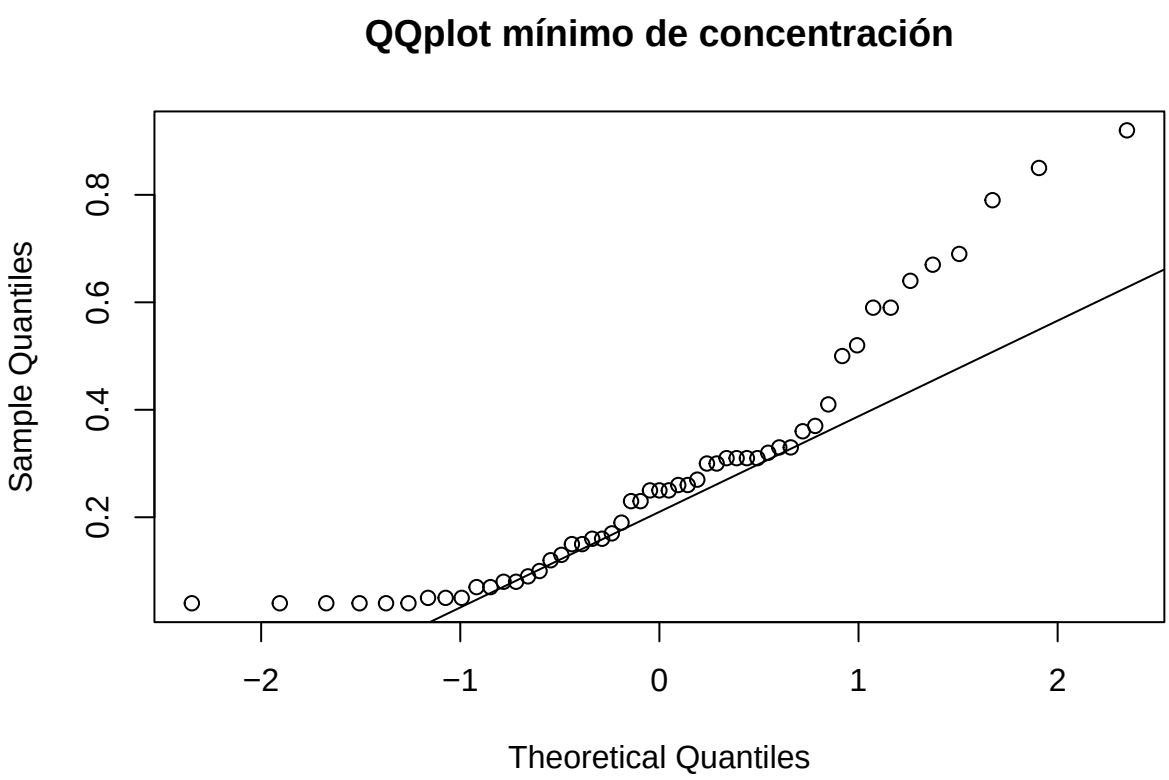


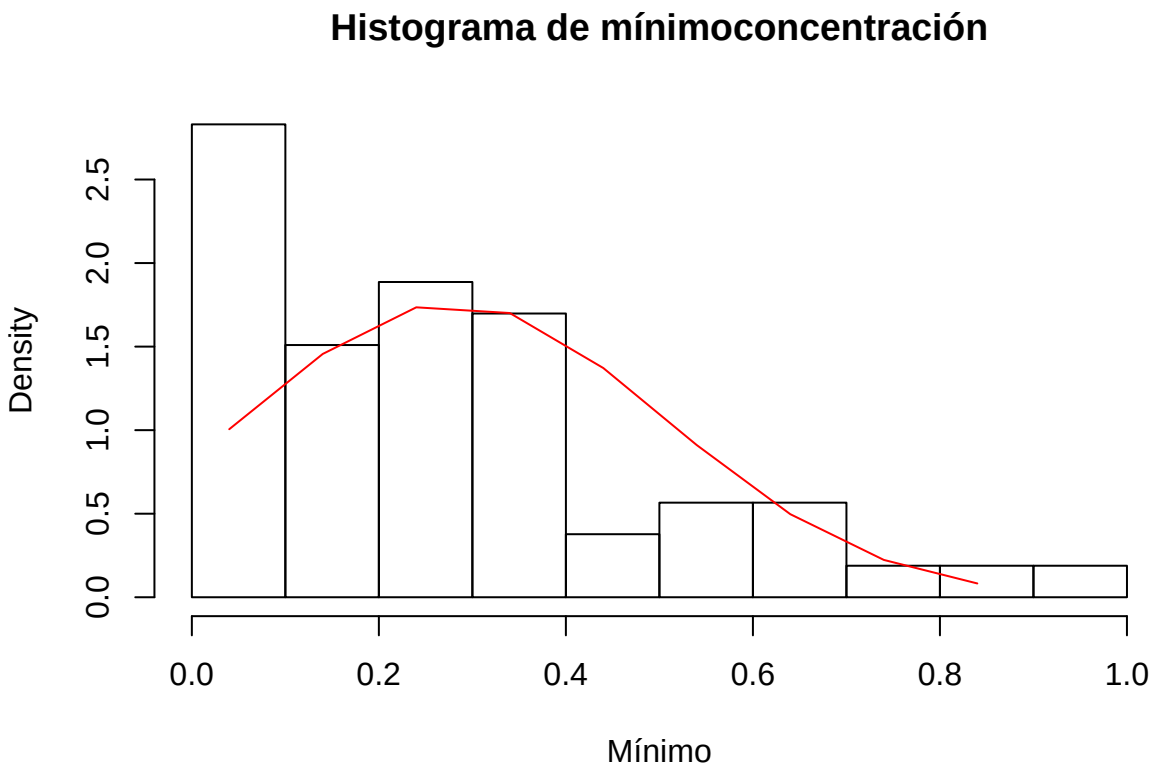
En la cantidad de peces del lago, se ve una casi simetría con respecto a la moda y la media en la distribución, pero, aquí si contamos con la presencia de 3 valores atípicos/extremos. El QQplot nos muestra una distribución con colas gruesas, es decir, baja curtosis y distribución platicúrtica.

- **X9**



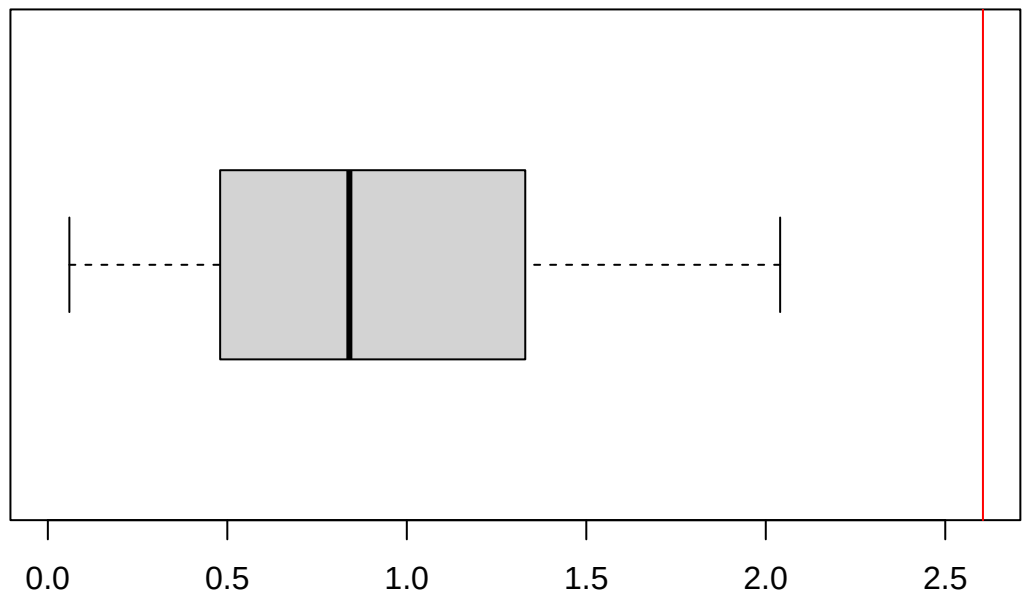

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0400  0.0825  0.2400  0.2454  0.3175  0.6900
```


QQplot mínimo de concentración
Sample Quantiles
0.2
0.4
0.6
0.8
−2
−1
0
1
2
Theoretical Quantiles

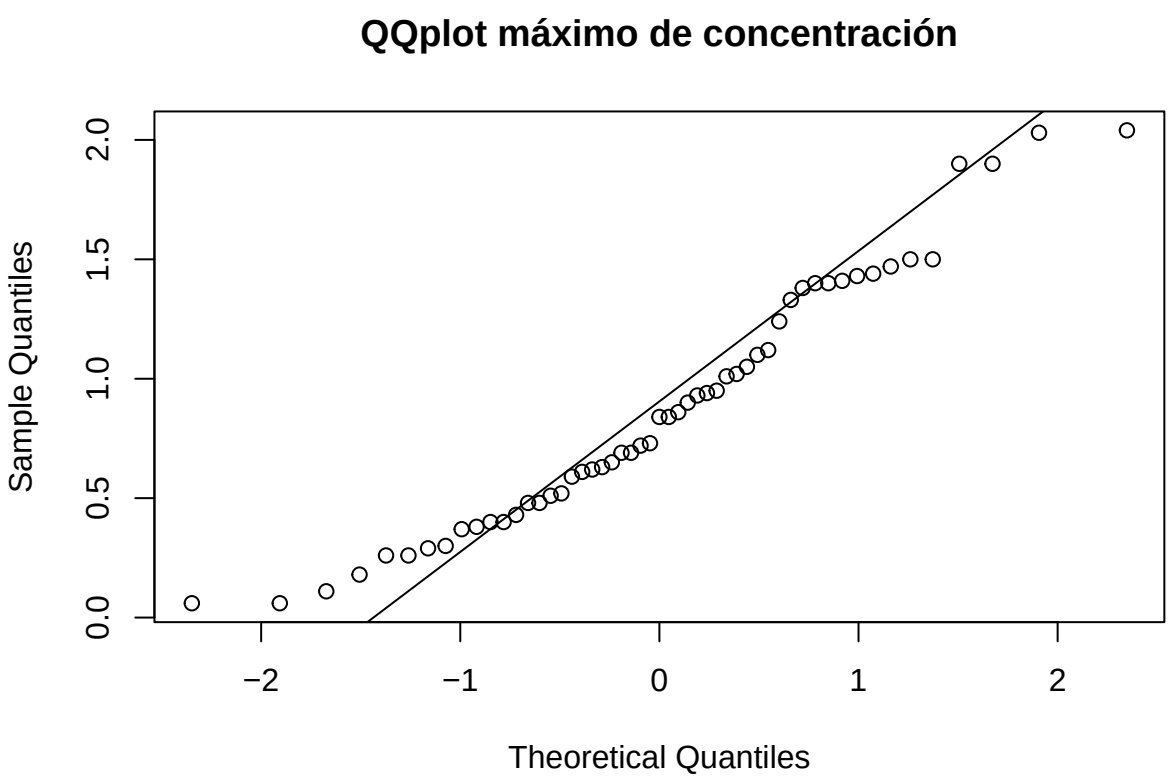


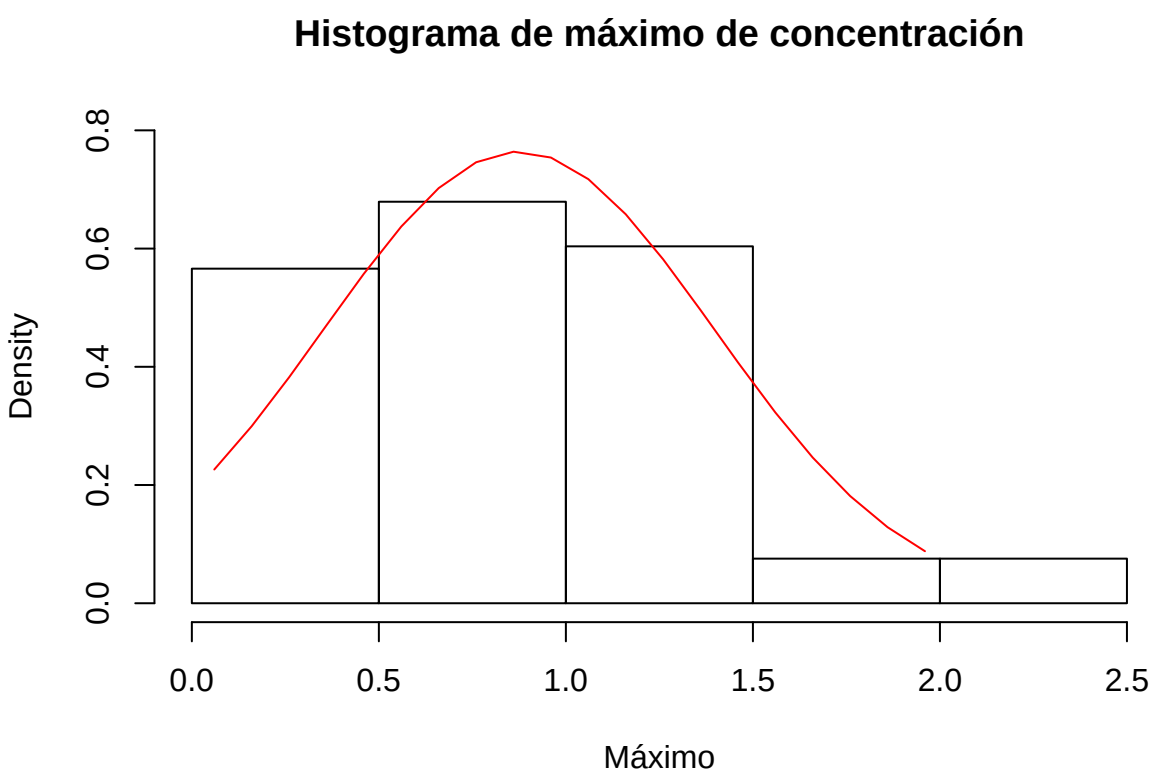
Se muestra un sesgo a la derecha con la distribución recargada hacia la izquierda de acuerdo a la media y mediana. Esta variable, no presenta datos atípicos y el QQplot, nos indica una asimetría positiva.

- **X10**



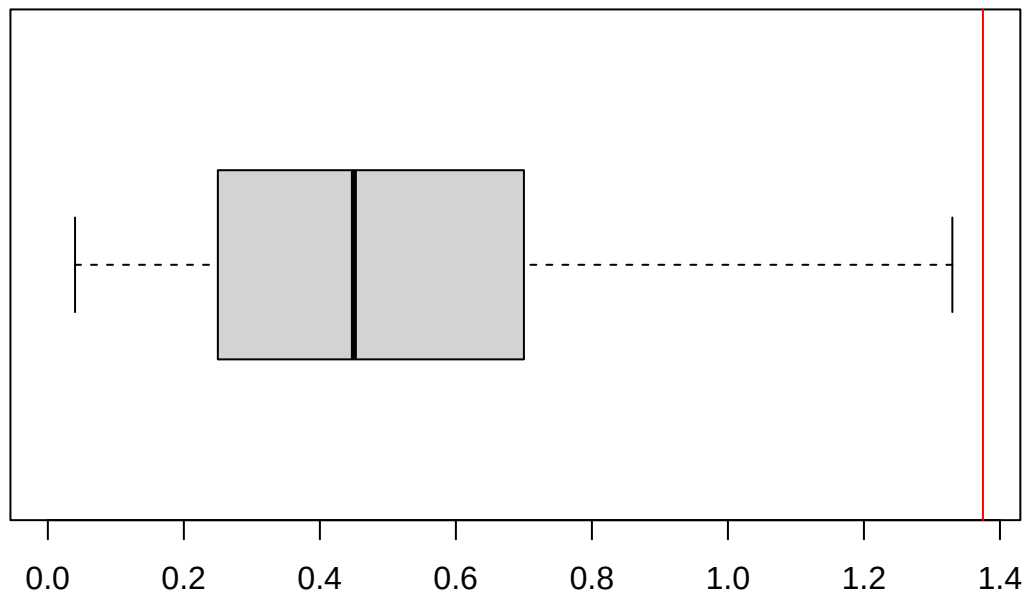

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0600  0.4800  0.8400  0.8745  1.3300  2.0400
```


QQplot máximo de concentración
Sample Quantiles
0.0
0.5
1.0
1.5
2.0
−2
−1
0
1
2
Theoretical Quantiles

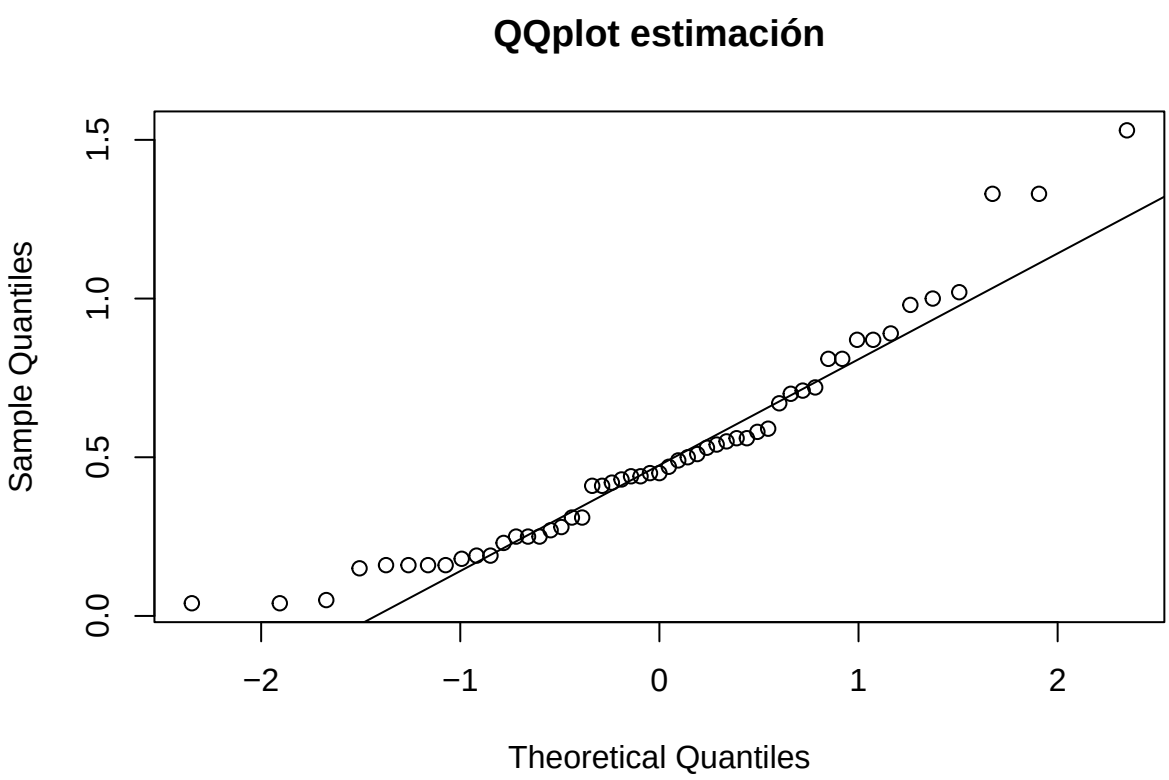


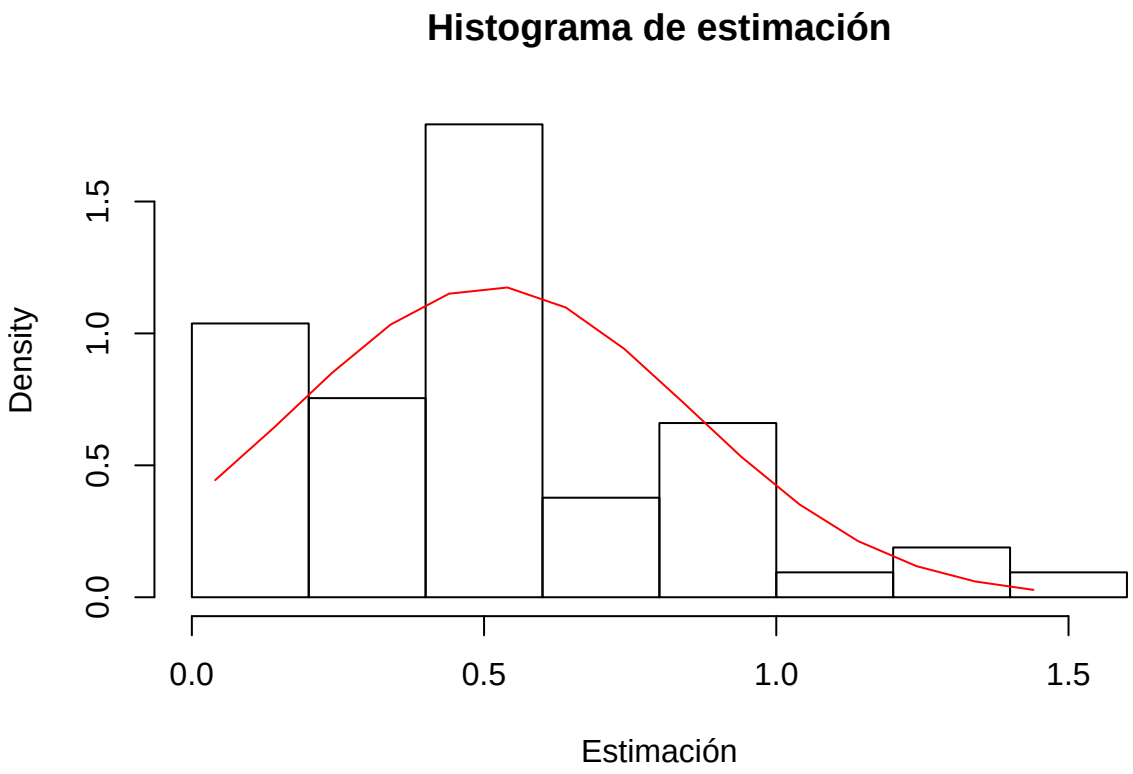
El máximo de la concentración nos muestra una simetría en los datos, ya que están de esta forma distribuidos de acuerdo a la mediana y la moda. Finalmente, el QQplot nos indica una probabilidad normal casi ideal.

- **X11**




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0400  0.2500  0.4500  0.4937  0.6775  1.3300
```

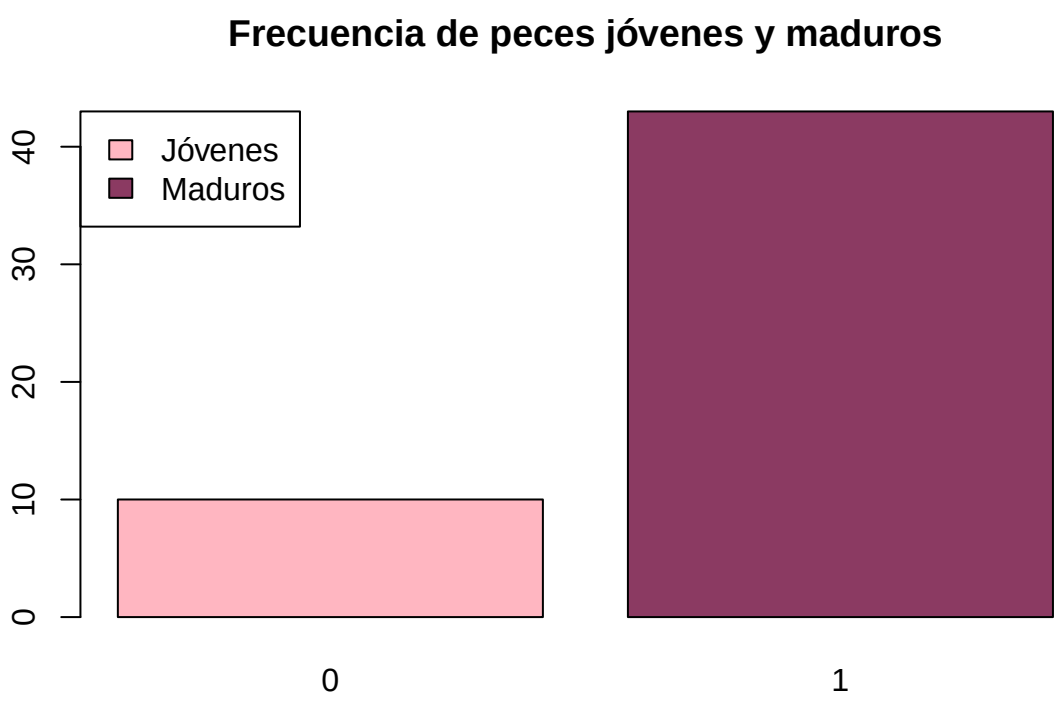

QQplot estimación
Sample Quantiles
Theoretical Quantiles
0.0
0.5
1.0
1.5
−2
−1
0
1
2



Estimación con sesgo hacia la derecha, distribución recargada hacia la izquierda con respecto a la media y mediana y sin datos atípicos. QQplot con asimetría positiva.

Medidas cualitativas

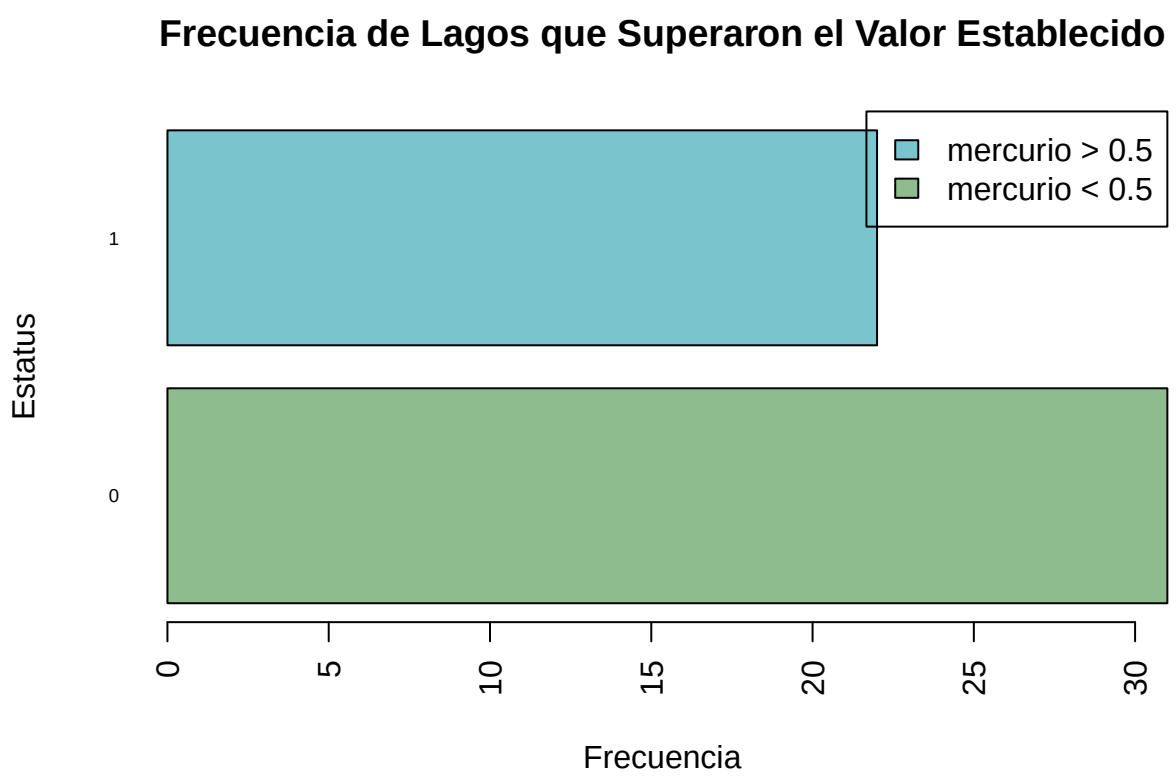
- **X12**



Aquí notamos que son más la cantidad de peces maduros que jóvenes, lo que puede indicar un problema en que el mercurio, esta afectando más a los jóvenes.

- **Medidición de lagos que superaron los 0.5 mg de hg/kg**

```
## [1] "Tabla de Distribución de Lagos que Superaron los 0.5 mg de Hg/Kg: "

##
##  0  1
## 31 22
```

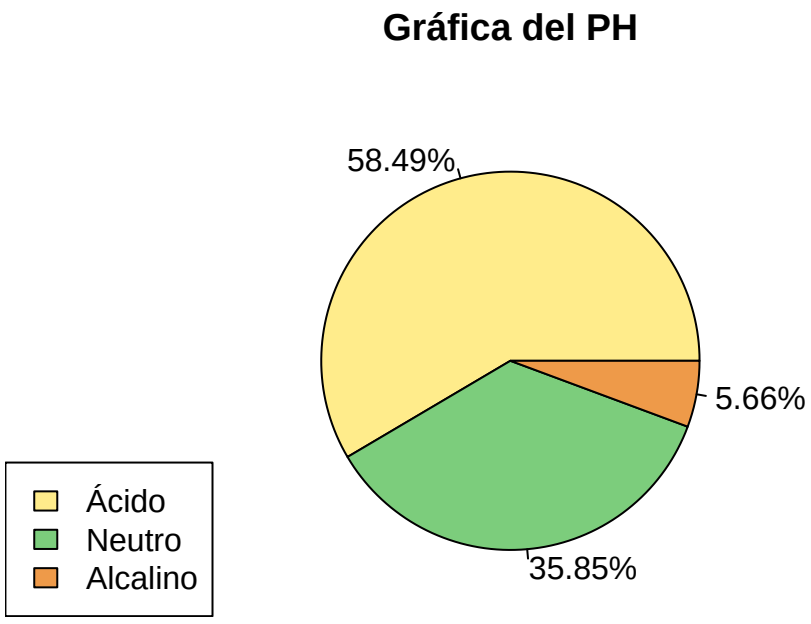


Con esta gráfica, visualizamos la cantidad de datos que superan la normativa de referencia de niveles máximos de mercurio, siendo más de la mitad los que no lo logran.

- **División y frecuencia de ph**

```
## [1] "Tabla de Distribución del PH: "
```

```
##
##    Ácido Alcalino   Neutro
##       31       19        3
```

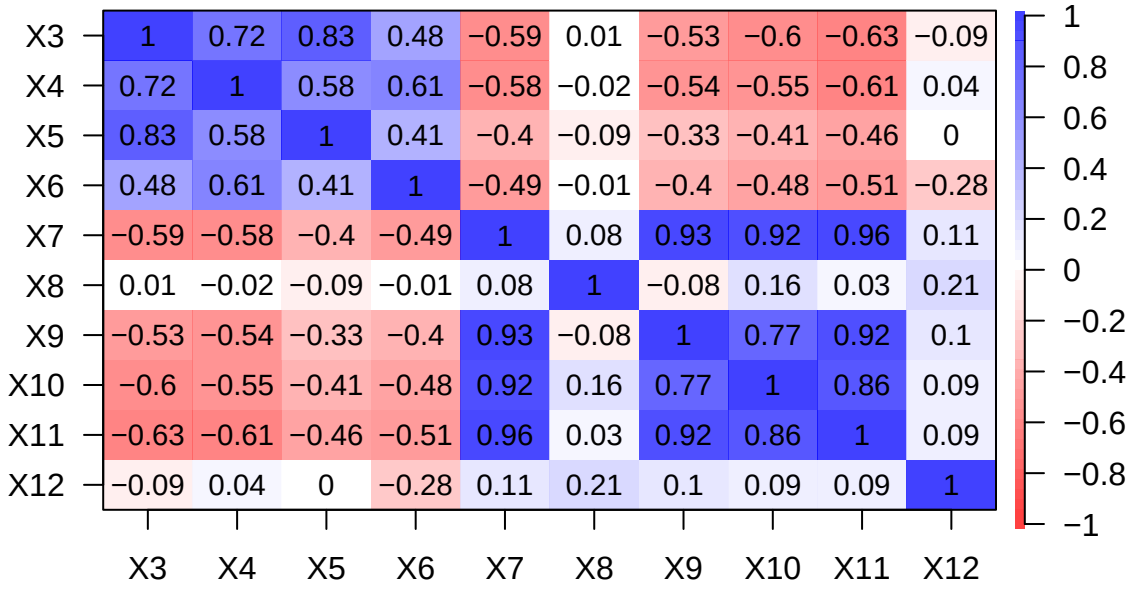


En esta gráfica de pastel, vemos que dominan más las variables con ph ácido, seguido del neutro y finalmente alcalino, lo cual muestra el gran afectamiento hacia los peces ya que les puede provocar el no funcionamiento de algunas células gracias a la producción prolongada de ácido.

Correlaciones

Matriz de correlacion

```
## Registered S3 method overwritten by 'psych':
##   method         from
##   plot.residuals rmutil
```



Debido a la matriz de correlación anteriormente desplegada, logramos ver que hay una mayor interacción y dependencia entre las variables:

- *X7 y X9* -> 0.93
- *X7 y X10* -> 0.92
- *X7 y X11* -> 0.96 Lo que da a indicar que la concentración media de mercurio en el tejido muscular del grupo de peces estudiados, tienen más relación con la estimación, el mínimo y el máximo de peces estudiados.

Sin embargo, escoger esyas variables puede ser perjudicial para la implementación de nuestro modelos, ya que se puede dar un problema de multicolinealidad y dificultar los resultados de interpretación, es por ello que escogeremos a las que indican una correlación debilmente negativa, es decir **X3, X4, X5, X6**

Implementación de modelos estadísticos

ANOVA

Mediante el método ANOVA, buscamos resolver la pregunta: **¿Habrá direfencia significativa entre la concentración de mercurio por la edad de los peces?**

Para ello, utilizaremos las variables:

- **X7** -> Concentración mdia de mercurio
- **X12** -> Indicador de la edad de los peces (jóvenes o maduros)

Es necesario para poder empezar, hacer un análisis sobre cada valor de concentración para saber si pertenece a un pez joven o maduro y de esta forma conocer cuantos datos corresponden a cada uno. Para este problema, utilizaremos un nivel de significancia de $\alpha = 0.05$

```
## [1] "Jovenes"
```

```
##  [1] 1.33 0.04 0.44 0.05 0.41 0.50 0.87 0.56 0.04 0.27
```

```
## [1] "Maduros"
```

```
##  [1] 1.23 1.20 0.27 0.48 0.19 0.83 0.81 0.71 0.50 0.49 1.16 0.15 0.19 0.77 1.08
## [16] 0.98 0.63 0.56 0.73 0.34 0.59 0.34 0.84 0.34 0.28 0.34 0.17 0.18 0.19 0.49
## [31] 1.10 0.16 0.10 0.48 0.21 0.86 0.52 0.65 0.94 0.40 0.43 0.25 0.27
```

```
## [1] "Media por edad"
```

```
##  [1] 1.33 0.04 0.44 0.05 0.41 0.50 0.87 0.56 0.04 0.27 1.23 1.20 0.27 0.48 0.19
## [16] 0.83 0.81 0.71 0.50 0.49 1.16 0.15 0.19 0.77 1.08 0.98 0.63 0.56 0.73 0.34
## [31] 0.59 0.34 0.84 0.34 0.28 0.34 0.17 0.18 0.19 0.49 1.10 0.16 0.10 0.48 0.21
## [46] 0.86 0.52 0.65 0.94 0.40 0.43 0.25 0.27
```

Por consiguiente, creamos ahora una variable llamada *edad* para poder tener nuestras clasificaciones correspondientes a ella y pasamos a la implementación del método ANOVA a través de la función *aov*

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## edad         1  0.072 0.07151    0.61  0.438
## Residuals   51  5.976 0.11718
```

Mediante los resultados de nuestro ANOVA, verificamos que la diferencia entre medias **no es estadísticamente significativa** es decir, **no hay diferencia** entre estos valores, lo que nos da por conclusión que la relación entre la edad y la concentración no es suficiente evidencia para conocer como se está comportando y afectando el problema del mercurio en los lagos.

Debido a lo anterior, nos concentramos en la siguiente pregunta: **¿Hay evidencia para suponer que la concentración promedio de mercurio en los lagos es dañino para la salud humana?**

Para esto consideraremos nuestra variable cualitativa que examinaba cuáles eran los mayores a 0.5 y los menores, esto de acuerdo a la normatividad de referencia para evaluar los niveles máximos de Hg, por lo que como en el ejemplo anterior, dividiremos las variables y examinaremos cuantas corresponden a cada una.

```
## [1] "Menor"
```

```
##  [1] 0.04 0.44 0.27 0.48 0.19 0.50 0.49 0.05 0.15 0.19 0.41 0.34 0.34 0.50 0.34
## [16] 0.28 0.34 0.17 0.18 0.19 0.04 0.49 0.16 0.10 0.48 0.21 0.27 0.40 0.43 0.25
## [31] 0.27
```

```
## [1] "Mayor"
```

```
##  [1] 1.23 1.33 1.20 0.83 0.81 0.71 1.16 0.77 1.08 0.98 0.63 0.56 0.73 0.59 0.84
## [16] 0.87 0.56 1.10 0.86 0.52 0.65 0.94
```

```
## [1] "Reglamento"
```

```
##  [1] 0.04 0.44 0.27 0.48 0.19 0.50 0.49 0.05 0.15 0.19 0.41 0.34 0.34 0.50 0.34
## [16] 0.28 0.34 0.17 0.18 0.19 0.04 0.49 0.16 0.10 0.48 0.21 0.27 0.40 0.43 0.25
## [31] 0.27 1.23 1.33 1.20 0.83 0.81 0.71 1.16 0.77 1.08 0.98 0.63 0.56 0.73 0.59
## [46] 0.84 0.87 0.56 1.10 0.86 0.52 0.65 0.94
```

Posteriormente pasamos a la implementación del ANOVA, realizando una variable nueva "reglamento". De igual manera, se utilizará un nivel de significancia $\alpha = 0.05$

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## reglamento   1  4.201   4.201     116 9.68e-15 ***
## Residuals   51  1.847   0.036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A diferencia del anterior, logramos ver que en esta ocasión $p < \alpha$, mostrandonos que las diferencias entre algunas de las medias son estadísticamente significativas, es decir, no todas las medias de la población son iguales, lo que indica que hay la evidencia para suponer que la concentración promedio de mercurio es dañina para la salud humana.

A través de este resultado, podemos visualizar que se pierde un grado de libertad, al contar con 52 en total siendo n = 53 y la variación entre tratamientos nos da un valor de 116, llegando a visualizar que como se mencionó anteriormente, ninguna de las concentraciones tiene un efecto igual a 0.

Análisis de cada reglamento por concentración

```
## [1] "Medias del reglamento: "
```

```
##     Mayor     Menor
## 0.8613636 0.2900000
```

```
## [1] "Desviación estándar del reglamento: "
```

```
##     Mayor     Menor
## 0.2397478 0.1460593
```

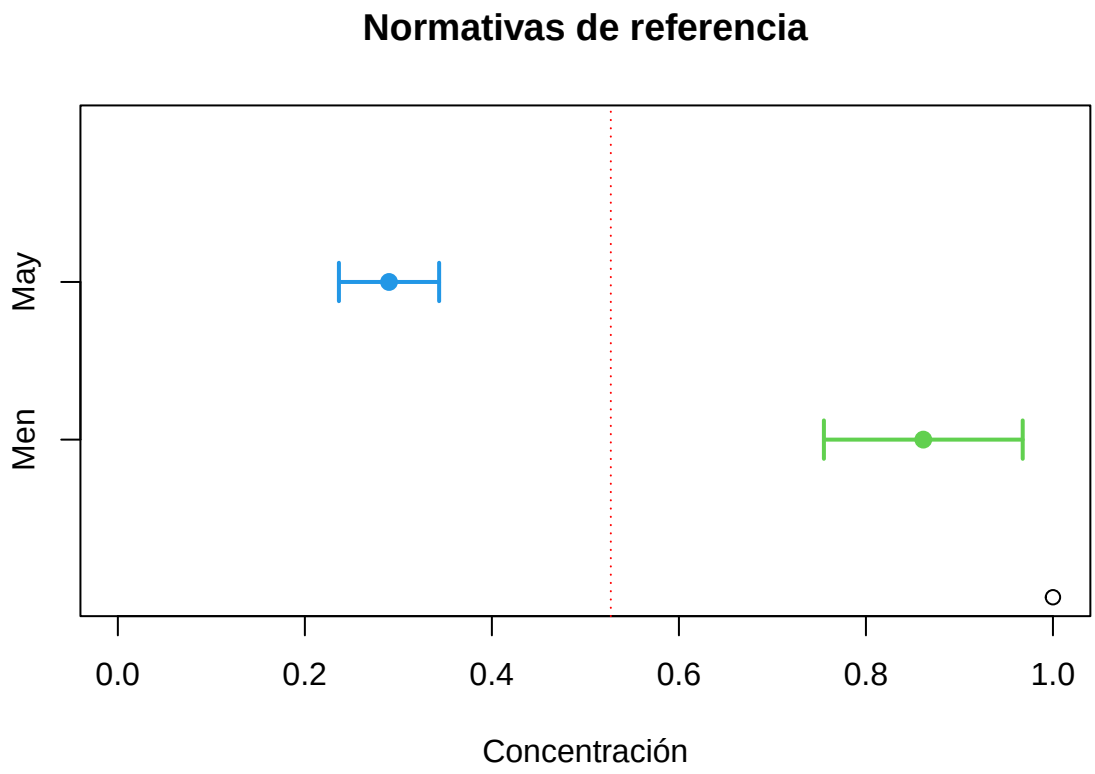
```
## [1] "Tamaño de la muestra del reglamento: "
```

```
## Mayor Menor
##    22    31
```

Invervalos de confianza

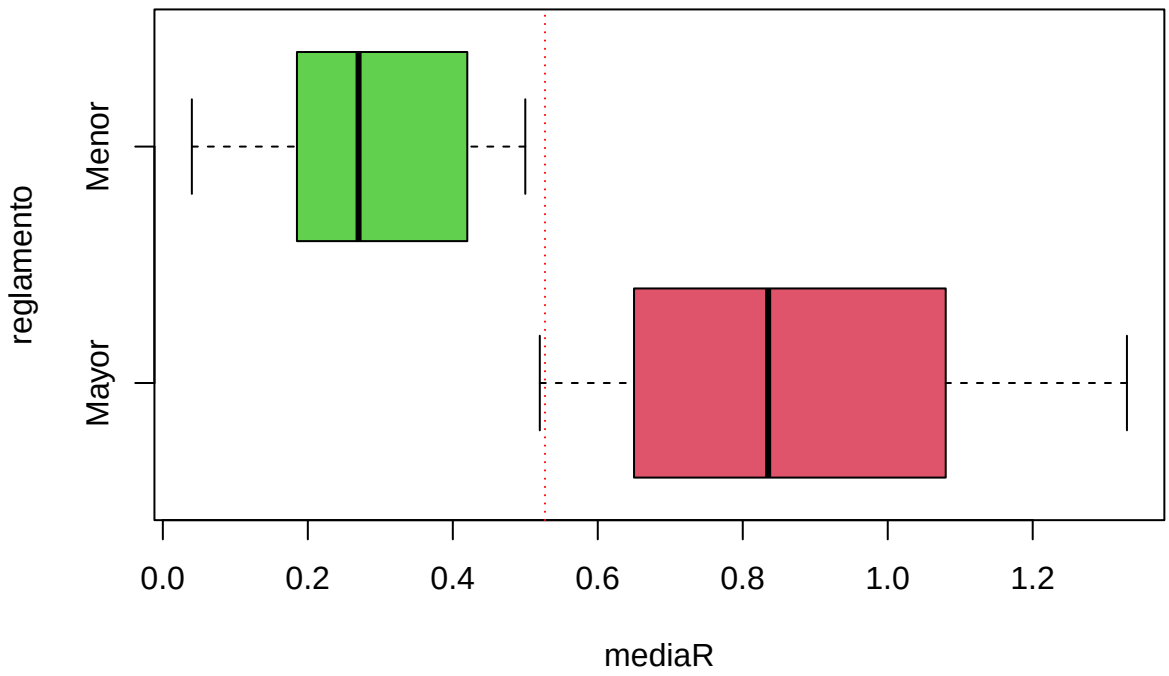
```
##     Mayor     Menor
## 0.7550654 0.2364250
```

```
##     Mayor     Menor
## 0.9676619 0.3435750
```



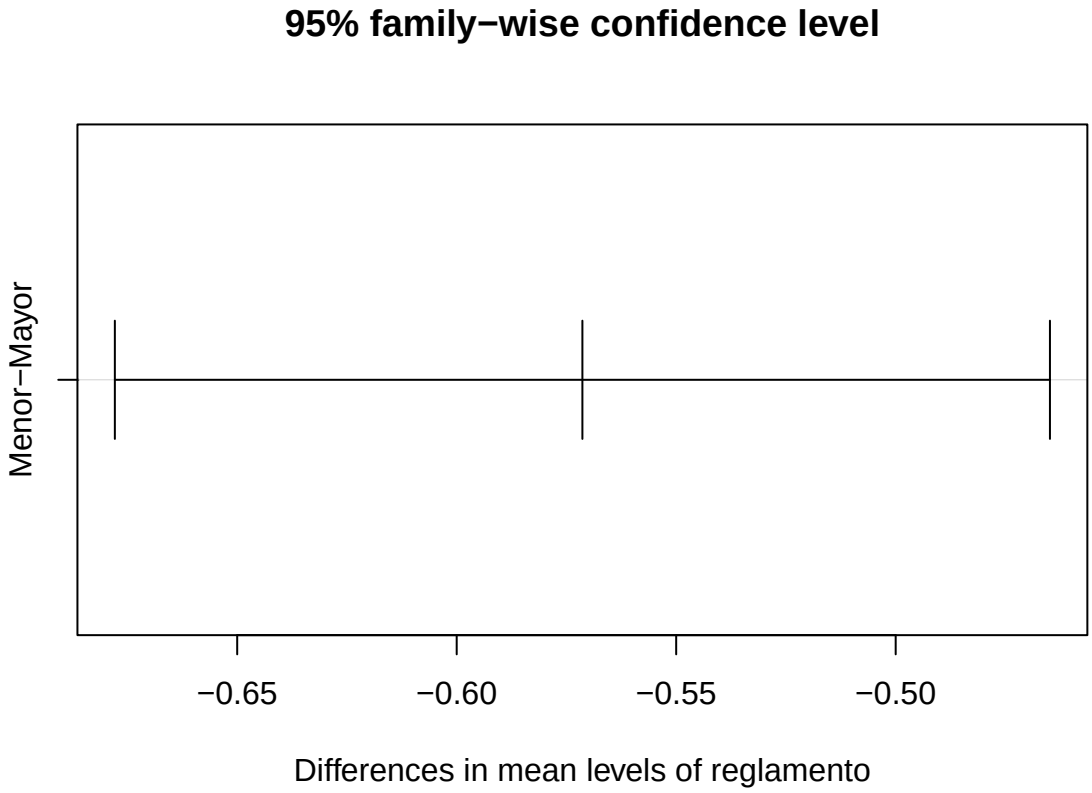
Esta gráfica nos muestra visualmente los intervalos de confianza siendo:

- **Menor** de 0.23 a 0.34
- **Mayor** de 0.75 a 0.96



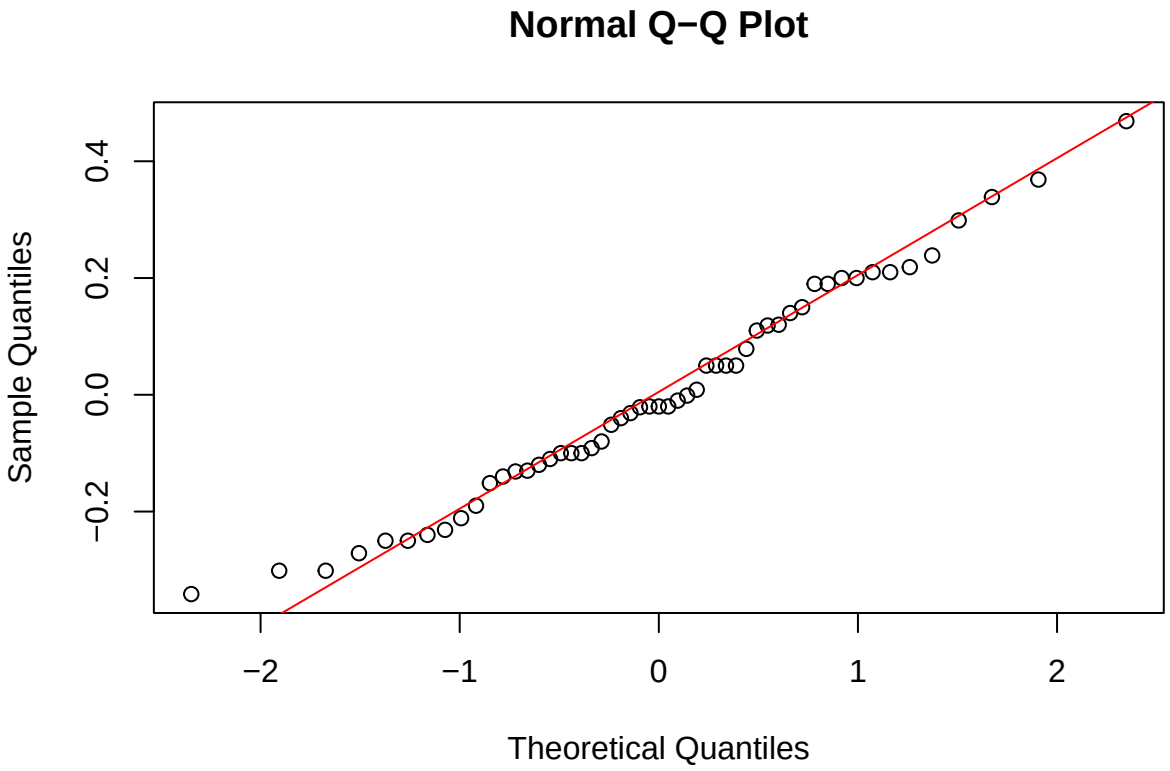
Los datos menores tienen una distribución concentrada hacia la derecha con respecto a su media, mientras que la mayor la tiene hacia la izquierda.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = mediaR ~ reglamento)
##
## $reglamento
##                    diff        lwr        upr p adj
## Menor-Mayor -0.5713636 -0.6778698 -0.4648575     0
```

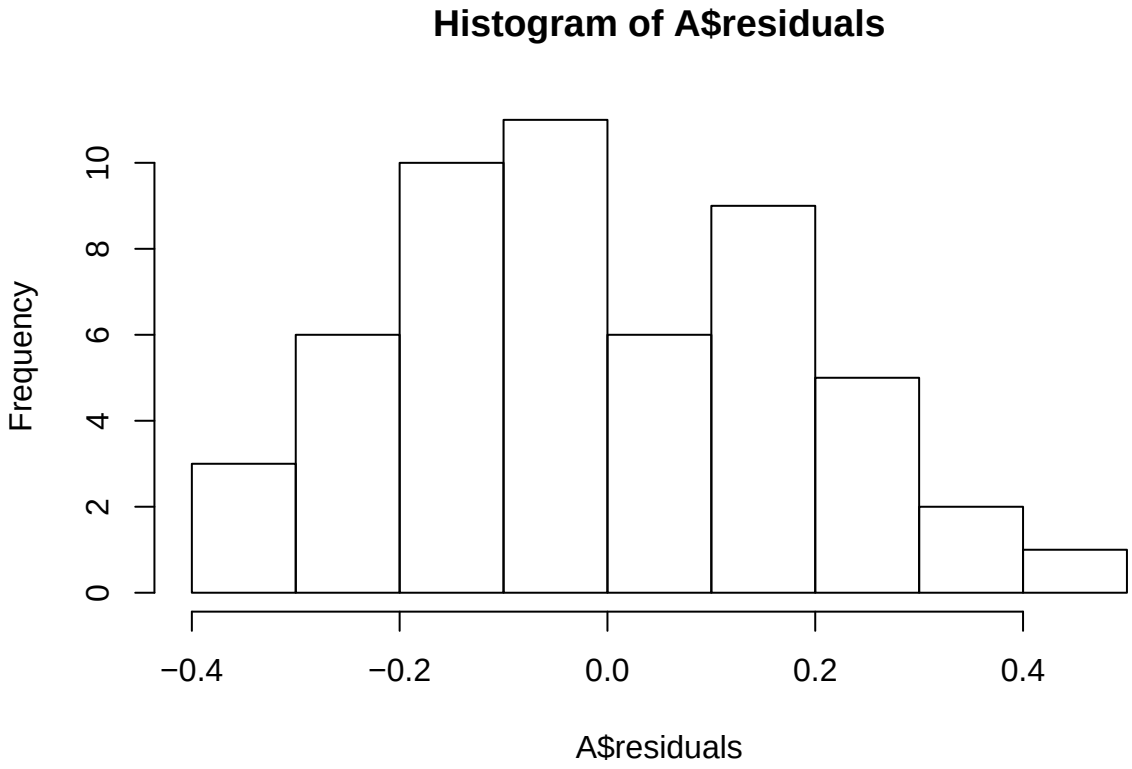


Como en esta ocasion solo nos basamos un reglamento de “Mayor” y “Menor”, no hay comparación significativa con otros pares de variables.

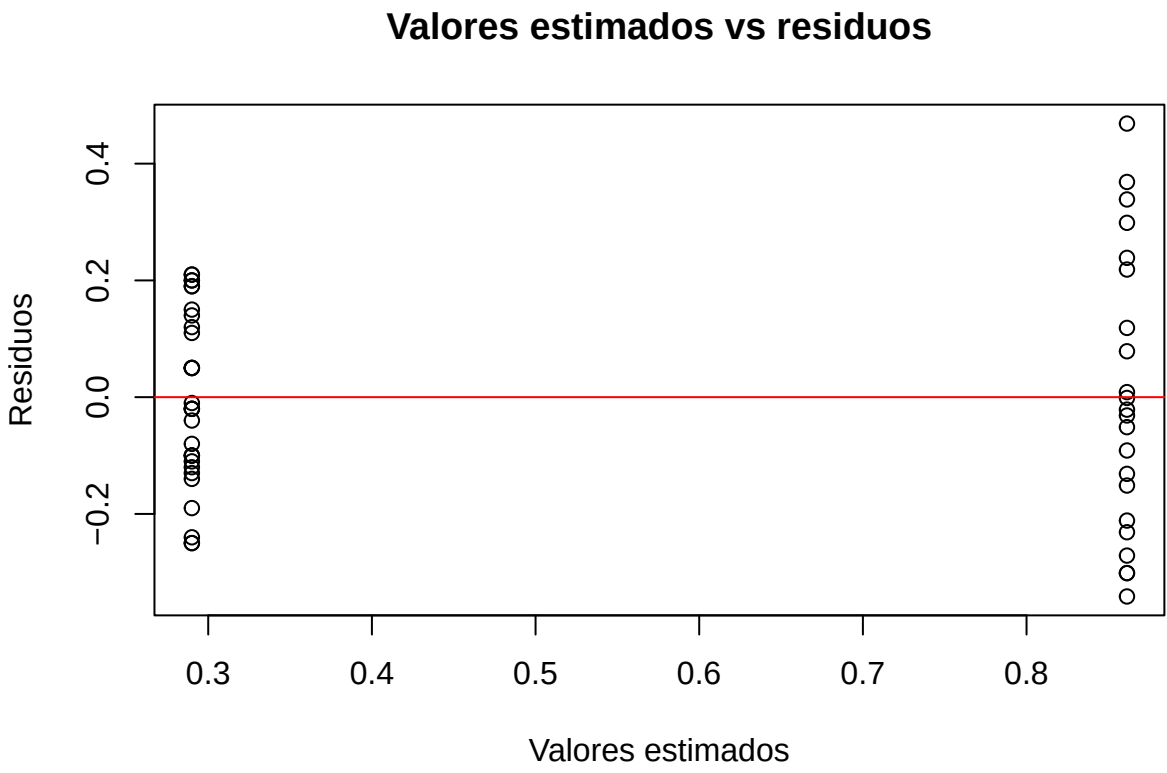
Verificación de supuestos



Normalidad

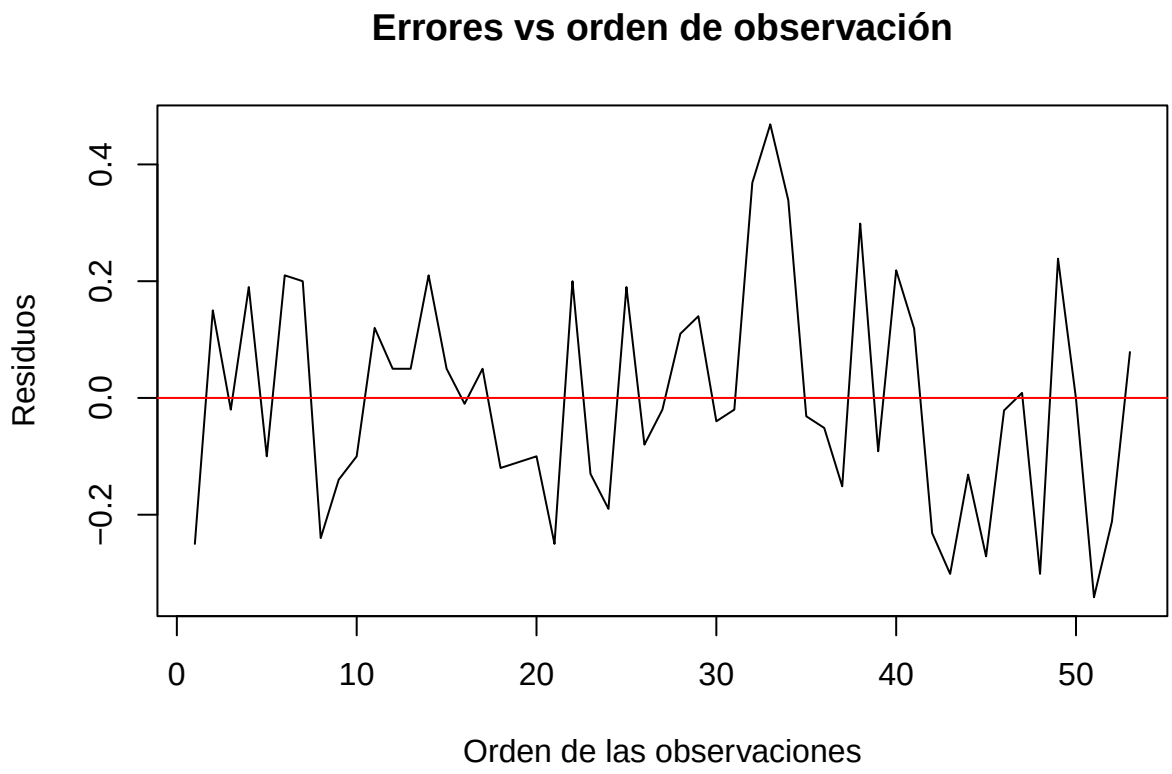


Mediante la gráfica podemos visualizar que la probabilidad Normal de nuestro análisis es ideal en su simetría.



Homocedasticidad
Gracias a esta gráfica, vemos que los residuos son constantes y cumplen con los supuestos.

####Independencia



En la independencia. vemos que nuestra autocorrelación en los errores son negativos, es decir, se observa una alternancia muy marcada de residuos positivos y negativos.

Regresión lineal múltiple

Para la implementación de nuestra regresión lineal múltiple, usaremos las variables anteriormente descritas en el gráfico de matriz de correlación para verificar cuál puede ser el mejor modelo y llegar a una conclusión de este problema.

Como inicio, pasamos dichas variables a un dataframe para tener una mejor manipulación de los datos.

```
##       X3  X4   X5    X6   X7
## 1    5.9 6.1  3.0   0.7 1.23
## 2    3.5 5.1  1.9   3.2 1.33
## 3  116.0 9.1 44.1 128.3 0.04
## 4   39.4 6.9 16.4   3.5 0.44
## 5    2.5 4.6  2.9   1.8 1.20
## 6   19.6 7.3  4.5  44.1 0.27
## 7    5.2 5.4  2.8   3.4 0.48
## 8   71.4 8.1 55.2  33.7 0.19
## 9   26.4 5.8  9.2   1.6 0.83
## 10   4.8 6.4  4.6  22.5 0.81
## 11   6.6 5.4  2.7  14.9 0.71
## 12  16.5 7.2 13.8   4.0 0.50
## 13  25.4 7.2 25.2  11.6 0.49
## 14   7.1 5.8  5.2   5.8 1.16
```

```
## 15 128.0 7.6 86.5  71.1 0.05
## 16  83.7 8.2 66.5  78.6 0.15
## 17 108.5 8.7 35.6  80.1 0.19
## 18  61.3 7.8 57.4  13.9 0.77
## 19   6.4 5.8  4.0   4.6 1.08
## 20  31.0 6.7 15.0  17.0 0.98
## 21   7.5 4.4  2.0   9.6 0.63
## 22  17.3 6.7 10.7   9.5 0.56
## 23  12.6 6.1  3.7  21.0 0.41
## 24   7.0 6.9  6.3  32.1 0.73
## 25  10.5 5.5  6.3   1.6 0.34
## 26  30.0 6.9 13.9  21.5 0.59
## 27  55.4 7.3 15.9  24.7 0.34
## 28   3.9 4.5  3.3   7.0 0.84
## 29   5.5 4.8  1.7  14.8 0.50
## 30   6.3 5.8  3.3   0.7 0.34
## 31  67.0 7.8 58.6  43.8 0.28
## 32  28.8 7.4 10.2  32.7 0.34
## 33   5.8 3.6  1.6   3.2 0.87
## 34   4.5 4.4  1.1   3.2 0.56
## 35 119.1 7.9 38.4  16.1 0.17
## 36  25.4 7.1  8.8  45.2 0.18
## 37 106.5 6.8 90.7  16.5 0.19
## 38  53.0 8.4 45.6 152.4 0.04
## 39   8.5 7.0  2.5  12.8 0.49
## 40  87.6 7.5 85.5  20.1 1.10
## 41 114.0 7.0 72.6   6.4 0.16
## 42  97.5 6.8 45.5   6.2 0.10
## 43  11.8 5.9 24.2   1.6 0.48
## 44  66.5 8.3 26.0  68.2 0.21
## 45  16.0 6.7 41.2  24.1 0.86
## 46   5.0 6.2 23.6   9.6 0.52
## 47  25.6 6.2 12.6  27.7 0.65
## 48  81.5 8.9 20.5   9.6 0.27
## 49   1.2 4.3  2.1   6.4 0.94
## 50  34.0 7.0 13.1   4.6 0.40
## 51  15.5 6.9  5.2  16.5 0.43
## 52  17.3 5.2  3.0   2.6 0.25
## 53  71.8 7.9 20.5   8.8 0.27
```

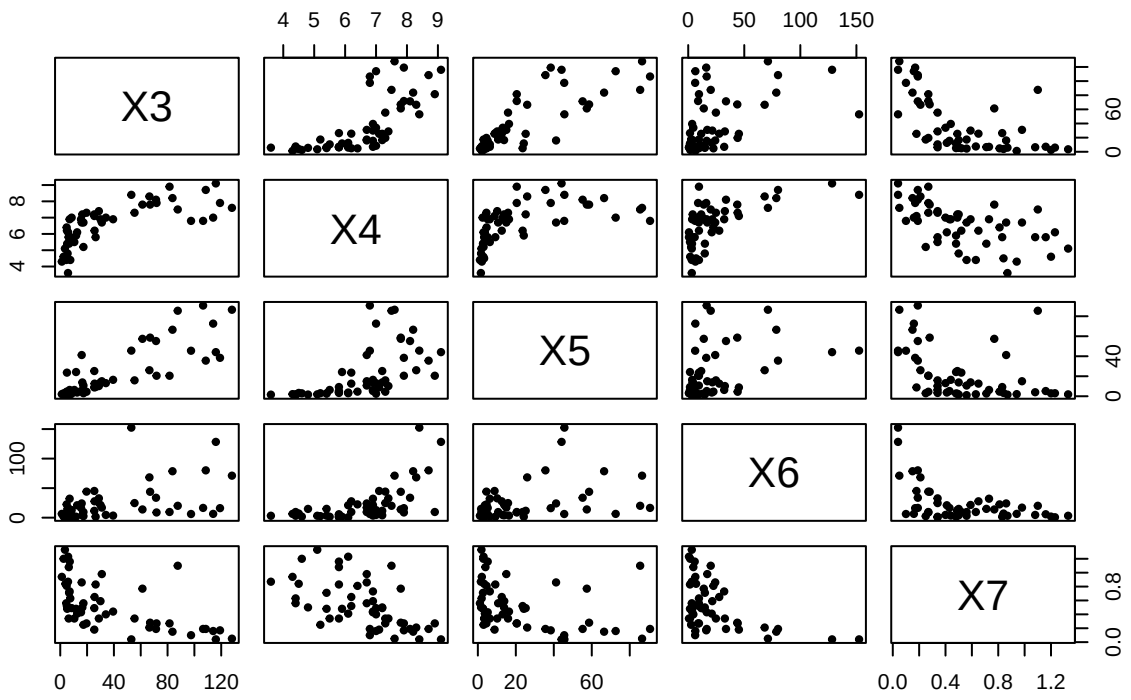
Correlacion

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha

##
## Attaching package: 'Hmisc'

## The following object is masked from 'package:psych':
##
##     describe

## The following objects are masked from 'package:base':
##
##     format.pval, units

##       X3    X4    X5    X6    X7
## X3  1.00  0.72  0.83  0.48 -0.59
## X4  0.72  1.00  0.58  0.61 -0.58
## X5  0.83  0.58  1.00  0.41 -0.40
## X6  0.48  0.61  0.41  1.00 -0.49
## X7 -0.59 -0.58 -0.40 -0.49  1.00
##
## n= 53
##
##
## P
##    X3     X4     X5     X6     X7
## X3        0.0000 0.0000 0.0003 0.0000
## X4 0.0000        0.0000 0.0000 0.0000
## X5 0.0000 0.0000        0.0023 0.0029
## X6 0.0003 0.0000 0.0023        0.0002
## X7 0.0000 0.0000 0.0029 0.0002
```

En este resultado, vemos ahora una matriz de correlación más detallada con las variables a tratar, por ello, notamos que la interacción con cada una de las variables (ejemplo: X3 con X4, X5 y X6) tiene una correlación positiva, por lo que no es recomendable hacer nuestra regresión lineal a partir de las variables mencionadas. Sin embargo, al tratarse de X7, la correlación es negativa, y vemos que pueden evitarse problemas de multicolinealidad para la implementación del modelo. Es por ello, que tomamos la vairable X7 como base para poder realizar la regresión múltiple con los valores restantes.

Por otro lado, en la segunda matriz, notamos que si establecemos un nivel de significancia $\alpha = 0.05$, nuestros valores p son menores a el, por lo que nos indica que contamos con valores significativamente estadísticos para podder implementar el modelo.

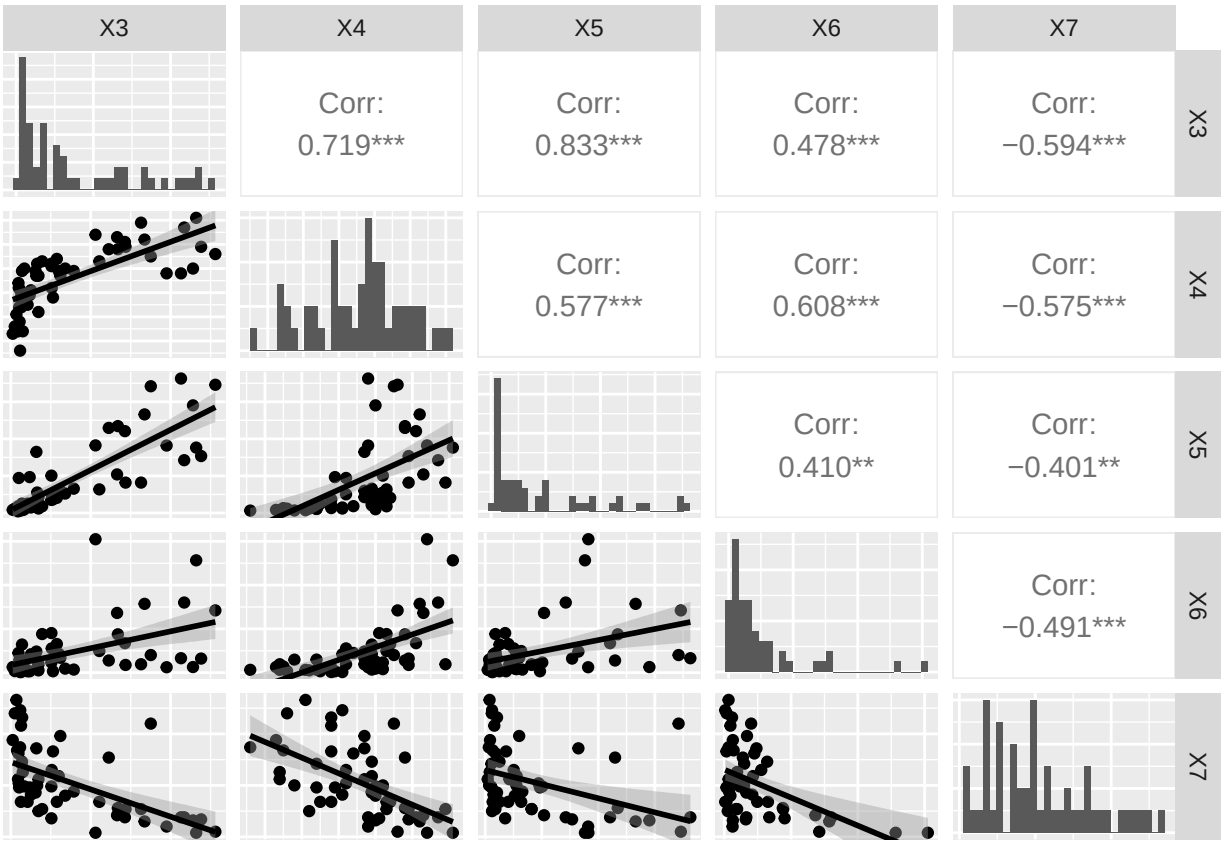


Bajo este diagrama, comprobamos correctamente que las variables tienen correlación entre cada una, visualizando altas y bajas correlaciones positivas/negativas y fuertes y débiles correlaciones positivas/negativas.

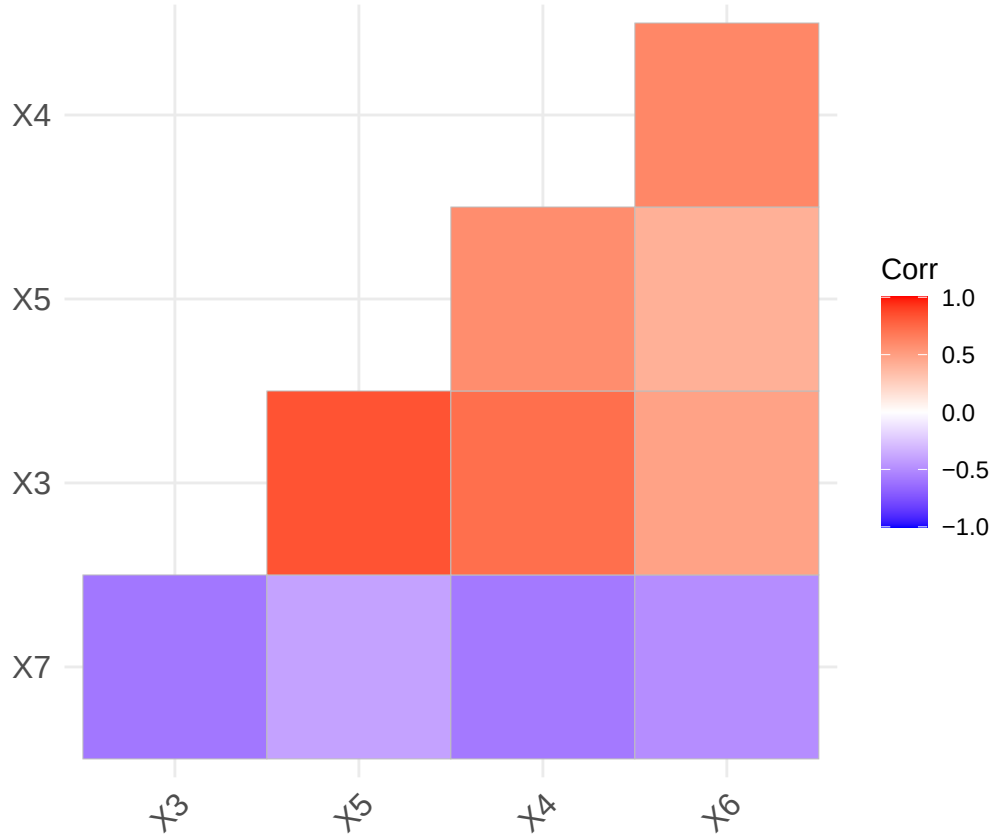
Visualización gráfica de la corrlación entre variables

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
##
## Attaching package: 'polycor'

## The following object is masked from 'package:psych':
##
##     polyserial
```



Aquí, visualizamos nuevamente que x7 contra las demás variables tienen una correlación negativa, mientras que las demás positivas, siendo X3 y X7 las más fuertes negativamente y la X5 y X3 las más fuertes positivamente.

El modelo

```
##
## Call:
## lm(formula = X7 ~ X3 + X4 + X5 + X6, data = medidas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42260 -0.19155 -0.08438  0.14334  0.62234
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.004440   0.257561   3.900 0.000299 ***
## X3          -0.005503   0.002028  -2.713 0.009224 **
## X4          -0.046709   0.045329  -1.030 0.307968
## X5           0.004129   0.002648   1.559 0.125484
## X6          -0.002361   0.001497  -1.577 0.121257
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2629 on 48 degrees of freedom
## Multiple R-squared:  0.4515, Adjusted R-squared:  0.4058
```

```
## F-statistic: 9.879 on 4 and 48 DF,  p-value: 6.499e-06
```

La función lm, nos ayuda a realizar la regresión múltiple de X7 con cada una de las variables, viendo así las intercepciones, estimación, la desviación estandar, el valor t y p.

Selección del mejor modelo

```
## Start:  AIC=-136.87
## X7 ~ X3 + X4 + X5 + X6
##
##        Df Sum of Sq    RSS     AIC
## - X4    1   0.07338 3.3904 -137.72
## <none>              3.3171 -136.87
## - X5    1   0.16803 3.4851 -136.25
## - X6    1   0.17196 3.4890 -136.19
## - X3    1   0.50874 3.8258 -131.31
##
## Step:  AIC=-137.71
## X7 ~ X3 + X5 + X6
##
##        Df Sum of Sq    RSS     AIC
## <none>              3.3904 -137.72
## - X5    1   0.18606 3.5765 -136.88
## + X4    1   0.07338 3.3171 -136.87
## - X6    1   0.35080 3.7412 -134.50
## - X3    1   0.90855 4.2990 -127.13

##
## Call:
## lm(formula = X7 ~ X3 + X5 + X6, data = medidas)
##
## Coefficients:
## (Intercept)           X3           X5           X6
##    0.744583    -0.006487     0.004333    -0.003035
```

Para la selección del mejor modelo, se toman en cuenta los resultados de la regresión realizada y la función step, hace un análisis de correlación entre las variables, analiza el modelo propuesto, detecta las variables no significativas en el modelo. Para llegar a la proposición de un nuevo modelo, se realiza el criterio de información de Akaike.

El mejor modelo

```
##
## Call:
## lm(formula = X7 ~ X3 + X5 + X6, data = medidas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38746 -0.18520 -0.07092  0.14490  0.61422
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

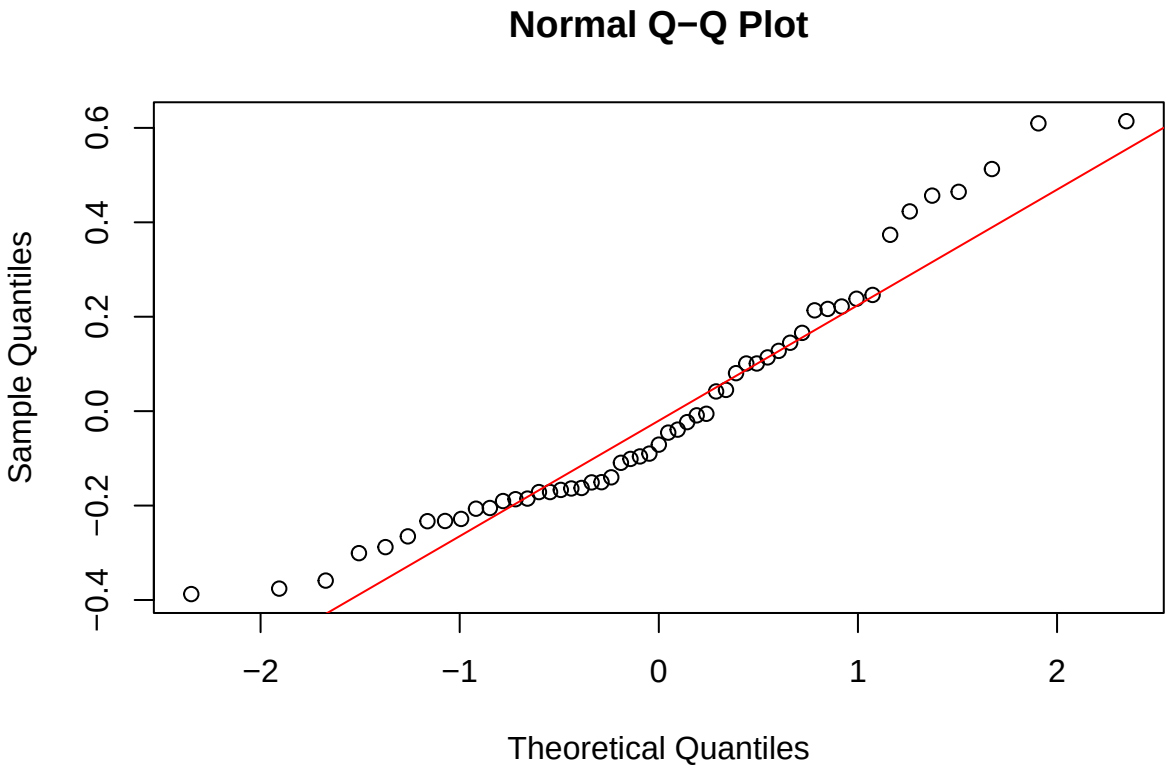
```
## (Intercept)  0.744583   0.052401  14.209  < 2e-16 ***
## X3          -0.006487   0.001790  -3.624 0.000689 ***
## X5           0.004333   0.002642   1.640 0.107445
## X6          -0.003035   0.001348  -2.252 0.028862 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.263 on 49 degrees of freedom
## Multiple R-squared:  0.4394, Adjusted R-squared:  0.4051
## F-statistic:  12.8 on 3 and 49 DF,  p-value: 2.676e-06
```

De acuerdo ahora a este nuevo modelo plantead por la función, vemos que de disminuyeron menos grados de libertad y que los errores medios son menores. De igual forma

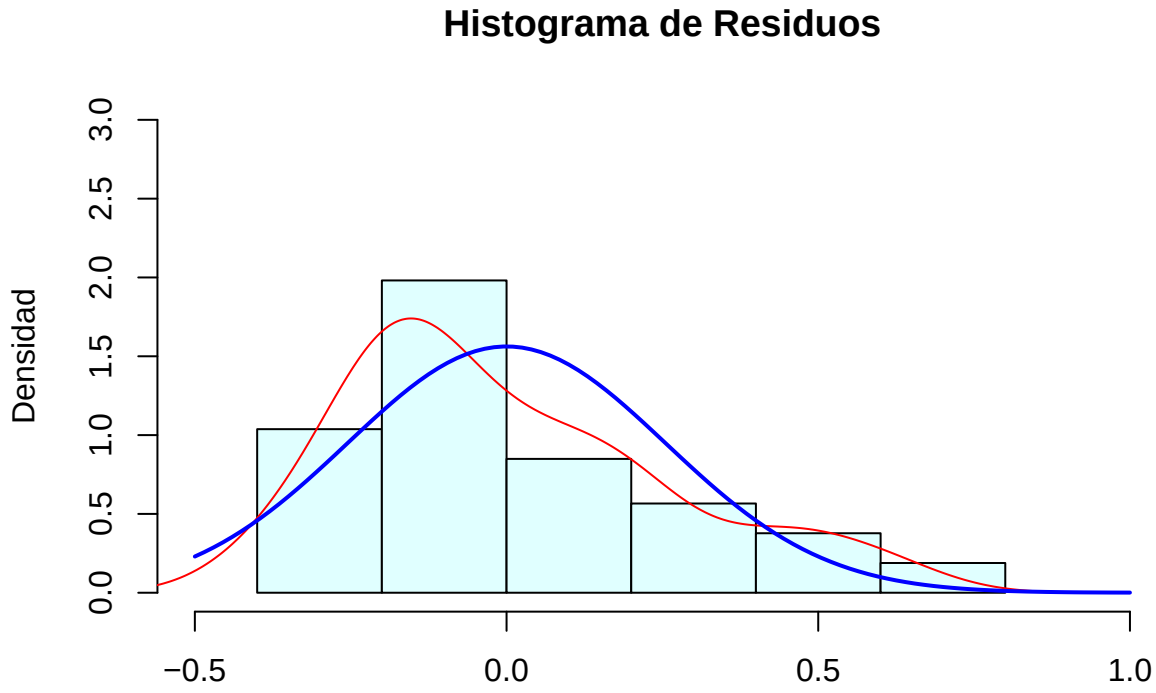
Intervalos de confianza

```
##                     2.5 %       97.5 %
## (Intercept)  0.6392783659  0.849887688
## X3          -0.0100848532 -0.002889577
## X5          -0.0009770002  0.009643095
## X6          -0.0057427822 -0.000326232
```

Verificación de supuestos

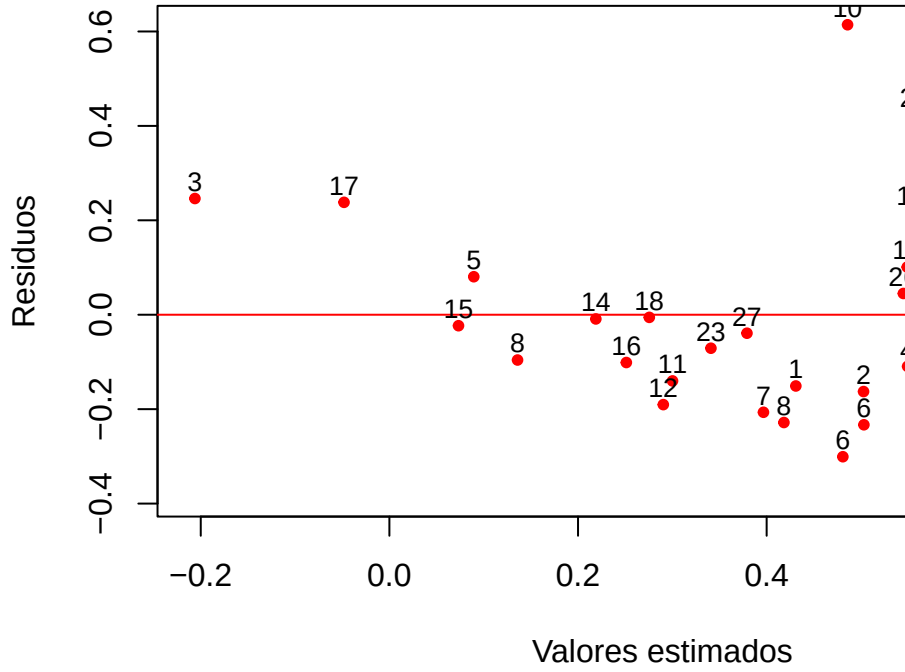


Normalidad

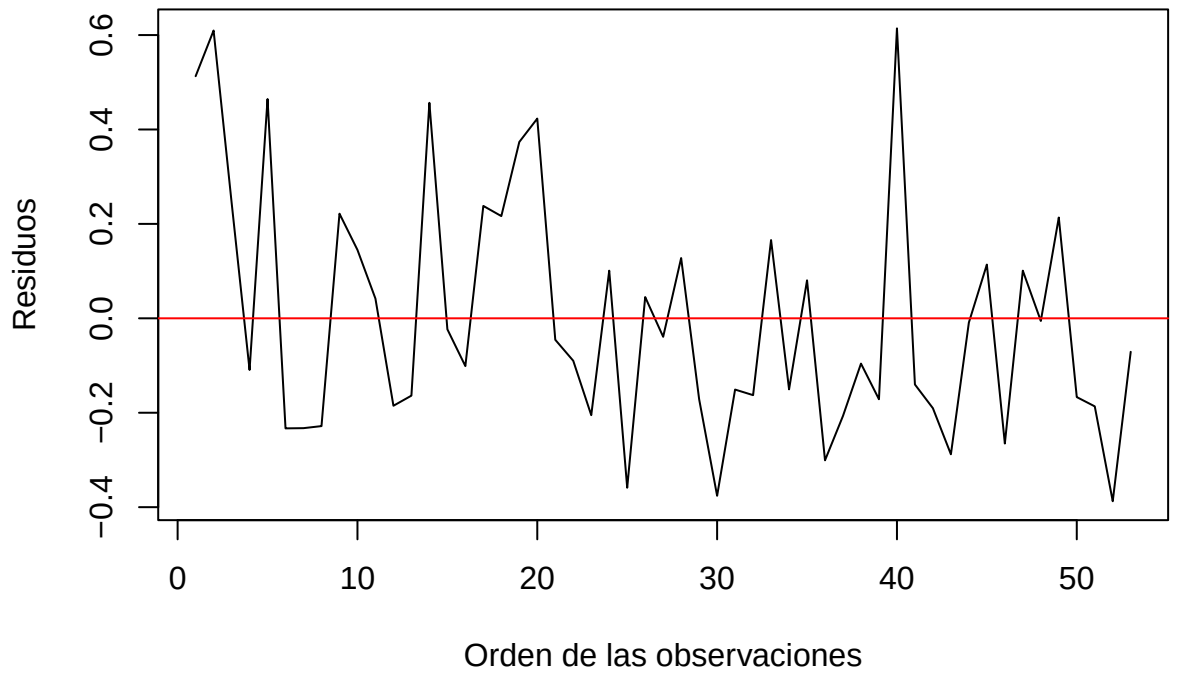


```
##
##  Shapiro-Wilk normality test
##
## data:  E
## W = 0.93258, p-value = 0.005116
```

En la normalidad, se nos muestra que la probabilidad es ideal con respecto al modelo.



Homocedasticidad y modelo apropiado
En la gráfica anterior, verificamos el modelo ya que se muestra cumple con los supuestos.



Independencia

En la independencia. vemos que nuestra autocorrelación en los errores son negativos, es decir, se observa una alternancia muy marcada de residuos positivos y negativos.

Prueba de autocorrelación para verificar independencia: H0: rho=0

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

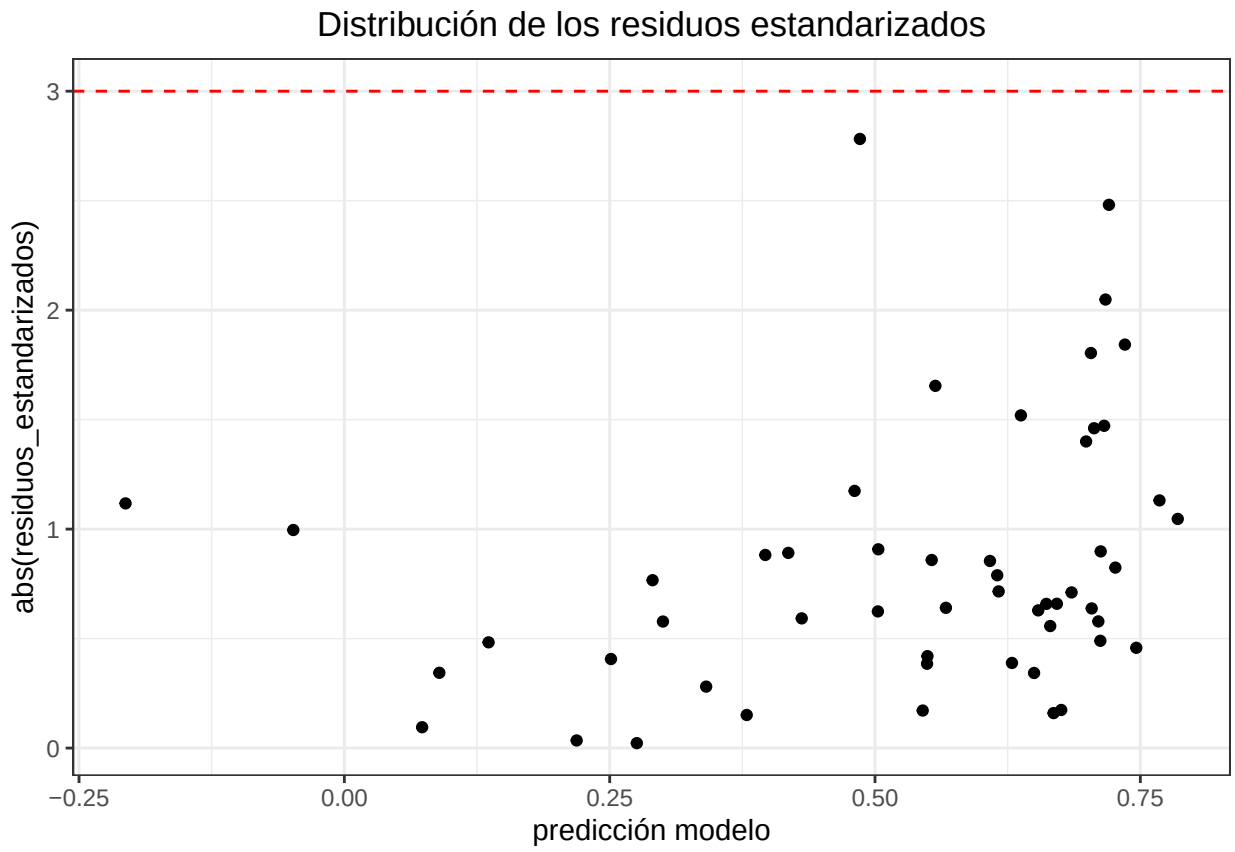
```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1660837      1.588784   0.132
##  Alternative hypothesis: rho != 0
```

Datos atípicos o influyentes

Datos atípicos Se estandarizan los residuos y se observa si hay distancias mayores a 3.

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:Hmisc':
##
##     src, summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```



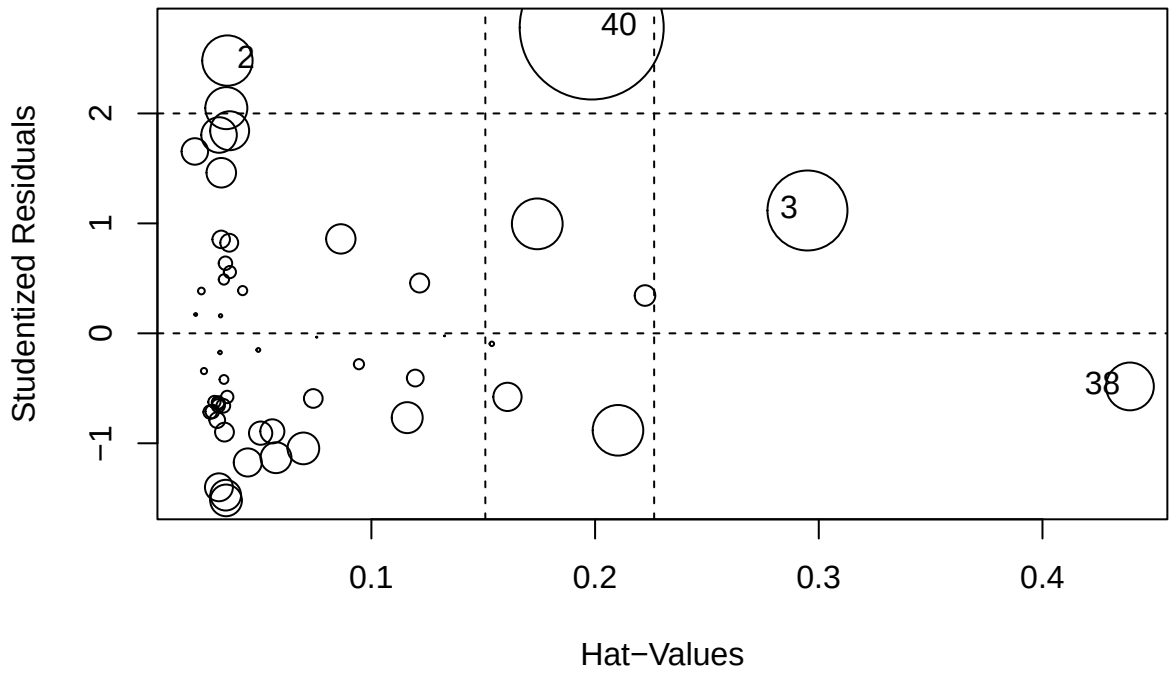

```
## integer(0)
```

En los datos atípicos se visualiza que las observaciones de medias de residuos no están alejados más allá del 4, que es el valor absoluto del residuo estandarizado, por lo que no hay datos atípicos.

Datos influyentes

```
## Potentially influential observations of
##   lm(formula = X7 ~ X3 + X5 + X6, data = medidas) :
##
##    dfb.1_ dfb.X3 dfb.X5  dfb.X6 dffit   cov.r   cook.d hat
## 2   0.48  -0.11  -0.04   -0.09   0.48    0.70_*  0.05   0.04
## 3  -0.22   0.28  -0.29    0.52   0.72    1.39_*  0.13   0.29_*
## 15  0.02   0.00  -0.02   -0.01  -0.04    1.28_*  0.00   0.15
## 35 -0.02   0.17  -0.11   -0.07   0.18    1.38_*  0.01   0.22
## 37  0.07   0.07  -0.31    0.18  -0.46    1.29_*  0.05   0.21
## 38  0.06   0.17  -0.09   -0.40  -0.43    1.90_*  0.05   0.44_*
## 40 -0.11  -0.50   1.14_* -0.38   1.38_*  0.75_*  0.42   0.20
## 41  0.03  -0.09  -0.06    0.15  -0.25    1.26_*  0.02   0.16
## 48  0.00  -0.01   0.01    0.00  -0.01    1.25_*  0.00   0.13


## Influence measures of
##   lm(formula = X7 ~ X3 + X5 + X6, data = medidas) :
##
##       dfb.1_   dfb.X3    dfb.X5   dfb.X6    dffit cov.r   cook.d    hat inf
## 1   0.388970 -0.06468 -0.039438 -0.10827  0.39101 0.805 3.59e-02 0.0352
## 2   0.476052 -0.10616 -0.043719 -0.08697  0.47696 0.695 5.15e-02 0.0356   *
## 3  -0.222093  0.28467 -0.290188  0.51806  0.72269 1.390 1.30e-01 0.2949   *
## 4  -0.050777 -0.03679  0.028122  0.04376 -0.07886 1.108 1.58e-03 0.0341
## 5   0.358090 -0.10303 -0.003135 -0.07513  0.35939 0.858 3.08e-02 0.0366
## 6  -0.106736  0.00905  0.077856 -0.13501 -0.20935 1.068 1.10e-02 0.0505
## 7  -0.168885  0.03466  0.015983  0.03359 -0.16938 1.052 7.20e-03 0.0343
## 8  -0.000803  0.04359 -0.134767  0.01498 -0.21654 1.077 1.18e-02 0.0557
## 9   0.128410  0.05020 -0.058758 -0.07517  0.15752 1.057 6.24e-03 0.0329
## 10  0.092138 -0.04622  0.000161  0.03486  0.10886 1.099 3.01e-03 0.0368
## 11  0.027551 -0.00706 -0.004475  0.00319  0.02932 1.120 2.19e-04 0.0326
## 12 -0.113077  0.03581 -0.022954  0.04127 -0.12261 1.072 3.80e-03 0.0289
## 13 -0.079432  0.05788 -0.062119  0.02397 -0.11338 1.085 3.25e-03 0.0315
## 14  0.324494 -0.08702 -0.007267 -0.05235  0.32752 0.863 2.56e-02 0.0319
## 15  0.016324 -0.00374 -0.015564 -0.00657 -0.04060 1.282 4.21e-04 0.1538   *
## 16  0.034233  0.05030 -0.081520 -0.08160 -0.14990 1.217 5.71e-03 0.1196
## 17 -0.102253  0.32584 -0.281250  0.17347  0.45733 1.212 5.23e-02 0.1742
## 18  0.033523 -0.09567  0.203278 -0.08968  0.26404 1.118 1.75e-02 0.0863
## 19  0.268051 -0.06087 -0.017937 -0.04918  0.26942 0.944 1.77e-02 0.0329
## 20  0.197518  0.03595 -0.060583 -0.02930  0.24278 0.889 1.42e-02 0.0211
## 21 -0.030911  0.00432  0.006855  0.00159 -0.03181 1.119 2.58e-04 0.0324
## 22 -0.052783  0.00940  0.001515  0.01032 -0.05511 1.103 7.73e-04 0.0252
## 23 -0.122204  0.01528  0.040693 -0.03176 -0.14141 1.064 5.04e-03 0.0311
## 24  0.058190 -0.03705  0.001675  0.04215  0.08185 1.120 1.70e-03 0.0425
## 25 -0.249601  0.03884  0.015308  0.08053 -0.25395 0.956 1.58e-02 0.0318
## 26  0.020025  0.00299 -0.007356  0.00132  0.02538 1.107 1.64e-04 0.0215
## 27 -0.010863 -0.02630  0.024816  0.00358 -0.03446 1.140 3.03e-04 0.0495
## 28  0.090883 -0.02778 -0.002940 -0.00800  0.09201 1.102 2.15e-03 0.0341
## 29 -0.115528  0.02872  0.021206 -0.01434 -0.12326 1.084 3.84e-03 0.0338
## 30 -0.278097  0.04597  0.027058  0.07899 -0.27978 0.943 1.91e-02 0.0349
## 31  0.004002  0.07299 -0.125312 -0.02624 -0.16750 1.139 7.11e-03 0.0740
## 32 -0.068435 -0.01467  0.047262 -0.04320 -0.10944 1.084 3.03e-03 0.0298
## 33  0.120124 -0.01556 -0.021264 -0.02504  0.12102 1.088 3.71e-03 0.0348
## 34 -0.110458  0.01698  0.018111  0.02124 -0.11104 1.095 3.13e-03 0.0355
## 35 -0.015672  0.16987 -0.108245 -0.07402  0.18390 1.383 8.61e-03 0.2223   *
## 36 -0.120720  0.00731  0.086380 -0.16680 -0.25421 1.015 1.60e-02 0.0448
```

```
## 37  0.065182  0.06774 -0.313523  0.17655 -0.45503 1.289 5.20e-02 0.2102   *
## 38  0.061805  0.16772 -0.086795 -0.40496 -0.42728 1.899 4.64e-02 0.4391   *
## 39 -0.113188  0.01743  0.026472 -0.00363 -0.11885 1.082 3.57e-03 0.0316
## 40 -0.110028 -0.49826  1.135482 -0.38357  1.38433 0.745 4.21e-01 0.1985   *
## 41  0.031899 -0.08908 -0.059832  0.15064 -0.25298 1.259 1.62e-02 0.1608   *
## 42  0.000794 -0.20023  0.071853  0.16526 -0.27777 1.170 1.95e-02 0.1160
## 43 -0.184014  0.18795 -0.188752  0.07401 -0.27905 1.037 1.94e-02 0.0574
## 44  0.000190 -0.00378  0.004500 -0.00640 -0.00995 1.175 2.53e-05 0.0756
## 45  0.054380 -0.14636  0.150082  0.01941  0.17037 1.215 7.38e-03 0.1216
## 46 -0.173793  0.23206 -0.208028  0.01075 -0.28623 1.066 2.04e-02 0.0696
## 47  0.045465 -0.00424 -0.012677  0.01864  0.06040 1.099 9.28e-04 0.0240
## 48 -0.001014 -0.00801  0.006119  0.00361 -0.00881 1.252 1.98e-05 0.1328   *
## 49  0.158099 -0.05348 -0.002747 -0.01164  0.16034 1.065 6.47e-03 0.0365
## 50 -0.083993 -0.04831  0.045527  0.05603 -0.11580 1.084 3.39e-03 0.0316
## 51 -0.110028  0.00368  0.036853 -0.00788 -0.12129 1.071 3.71e-03 0.0279
## 52 -0.257310 -0.06797  0.124997  0.09761 -0.28953 0.933 2.04e-02 0.0350
## 53 -0.017163 -0.07863  0.059160  0.04006 -0.09064 1.191 2.09e-03 0.0944
```




```
##       StudRes        Hat      CookD
## 2   2.4809441 0.03564225 0.05145857
## 3   1.1173717 0.29494242 0.12991250
## 38 -0.4829154 0.43910418 0.04636791
## 40  2.7817319 0.19849694 0.42117600
```

Con esta gráfica, logramos visualizar todos los puntos anteriormente calculados con respecto a la detección de datos influyentes, los cuales son los puntos que tienen impacto en las estimativas del modelo.

Conclusión

De acuerdo con el problema, se buscaba investigar cuales eran los principales factores que influían e la contaminación de lagos y su afectación a los peces y seres humanos. Después de realizar el método anova, primero se buscó si es que la concentración media de mercurio existente en los lagos, afectaba de importante manera a los peces respecto a su edad, ya fueran jóvenes o maduros. Sin embargo, el mismo método nos arrojó que no había la evidencia suficiente para poder llegar a una conclusión de ese tipo, lo que nos da a entender que sin importar la edad del pez, este se ve afectado de misma manera por el mercurio contaminante. Por otro lado, buscamos la evidencia para poder suponer que la concentración promedio del mercurio en lagos fuera dañino en la salud humana, y aunque en nuestra gráfica de exploración de datos se arrojaban que eran más los que no superaban la normatividad de referencia, mediante el método ANOVA se encontraron datos significativamente relevantes que pudieran indicar una evidencia de afección ya que, si hay valores existentes en algunos datos que pueden afectar a un ser humano a través de la pesca. Finalmente, respondiendo a la pregunta sobre cuáles son los principales factores que influyen en el nivel de contaminación, gracias a la regresión lineal múltiple, concluimos que son la **alcalinidad, calcio y clorofila**, ya que son las variables que conformaban el mejor modelo de regresión, y que más afluencia tenían con la concentración media del mercurio.

Anexo

Link a github: https://github.com/A01749373/CreacionModeloEstadistico-A01749373