

MomentoRetroalimentacionMod1

Amy Murakami Tsutsumi - A01750185

2022-09-08

Este portafolio de implementación tiene el propósito de utilizar herramientas estadísticas vistas en el módulo uno para poder construir un modelo que pueda contestar la pregunta de investigación establecida. El problema consiste en la contaminación por mercurio de los peces que se encuentran en agua dulce, por lo tanto, se utilizará un dataset con información de 53 lagos de Florida. La pregunta que se debe contestar a lo largo de la implementación es la siguiente: ¿cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida?

Exploración de la base de datos

Leer datos

Se almacenarán todas las variables excepto X1 y X2 que son el número de identificación y nombre del lago para realizar la exploración y análisis de datos.

```
db=read.csv("mercurio.csv") #leer la base de datos
X3 = db$X3 # Alcalinidad (mg/l de carbonato de calcio)
X4 = db$X4 # PH
X5 = db$X5 # Calcio (mg/l)
X6 = db$X6 # Clorofila (mg/l)
X7 = db$X7 # Concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces
X8 = db$X8 # Número de peces estudiados en el lago
X9 = db$X9 # Mínimo de la concentración de mercurio en cada grupo de peces
X10 = db$X10 # Máximo de la concentración de mercurio en cada grupo de peces
X11 = db$X11 # Estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o p
X12 = db$X12 # Indicador de la edad de los peces (0: jóvenes; 1: maduros)
```

```
dbNum = db[3:11]
dbNum
```

##	X3	X4	X5	X6	X7	X8	X9	X10	X11
## 1	5.9	6.1	3.0	0.7	1.23	5	0.85	1.43	1.53
## 2	3.5	5.1	1.9	3.2	1.33	7	0.92	1.90	1.33
## 3	116.0	9.1	44.1	128.3	0.04	6	0.04	0.06	0.04
## 4	39.4	6.9	16.4	3.5	0.44	12	0.13	0.84	0.44
## 5	2.5	4.6	2.9	1.8	1.20	12	0.69	1.50	1.33
## 6	19.6	7.3	4.5	44.1	0.27	14	0.04	0.48	0.25
## 7	5.2	5.4	2.8	3.4	0.48	10	0.30	0.72	0.45
## 8	71.4	8.1	55.2	33.7	0.19	12	0.08	0.38	0.16
## 9	26.4	5.8	9.2	1.6	0.83	24	0.26	1.40	0.72

## 10	4.8	6.4	4.6	22.5	0.81	12	0.41	1.47	0.81
## 11	6.6	5.4	2.7	14.9	0.71	12	0.52	0.86	0.71
## 12	16.5	7.2	13.8	4.0	0.50	12	0.10	0.73	0.51
## 13	25.4	7.2	25.2	11.6	0.49	7	0.26	1.01	0.54
## 14	7.1	5.8	5.2	5.8	1.16	43	0.50	2.03	1.00
## 15	128.0	7.6	86.5	71.1	0.05	11	0.04	0.11	0.05
## 16	83.7	8.2	66.5	78.6	0.15	10	0.12	0.18	0.15
## 17	108.5	8.7	35.6	80.1	0.19	40	0.07	0.43	0.19
## 18	61.3	7.8	57.4	13.9	0.77	6	0.32	1.50	0.49
## 19	6.4	5.8	4.0	4.6	1.08	10	0.64	1.33	1.02
## 20	31.0	6.7	15.0	17.0	0.98	6	0.67	1.44	0.70
## 21	7.5	4.4	2.0	9.6	0.63	12	0.33	0.93	0.45
## 22	17.3	6.7	10.7	9.5	0.56	12	0.37	0.94	0.59
## 23	12.6	6.1	3.7	21.0	0.41	12	0.25	0.61	0.41
## 24	7.0	6.9	6.3	32.1	0.73	12	0.33	2.04	0.81
## 25	10.5	5.5	6.3	1.6	0.34	10	0.25	0.62	0.42
## 26	30.0	6.9	13.9	21.5	0.59	36	0.23	1.12	0.53
## 27	55.4	7.3	15.9	24.7	0.34	10	0.17	0.52	0.31
## 28	3.9	4.5	3.3	7.0	0.84	8	0.59	1.38	0.87
## 29	5.5	4.8	1.7	14.8	0.50	11	0.31	0.84	0.50
## 30	6.3	5.8	3.3	0.7	0.34	10	0.19	0.69	0.47
## 31	67.0	7.8	58.6	43.8	0.28	10	0.16	0.59	0.25
## 32	28.8	7.4	10.2	32.7	0.34	10	0.16	0.65	0.41
## 33	5.8	3.6	1.6	3.2	0.87	12	0.31	1.90	0.87
## 34	4.5	4.4	1.1	3.2	0.56	13	0.25	1.02	0.56
## 35	119.1	7.9	38.4	16.1	0.17	12	0.07	0.30	0.16
## 36	25.4	7.1	8.8	45.2	0.18	13	0.09	0.29	0.16
## 37	106.5	6.8	90.7	16.5	0.19	13	0.05	0.37	0.23
## 38	53.0	8.4	45.6	152.4	0.04	4	0.04	0.06	0.04
## 39	8.5	7.0	2.5	12.8	0.49	12	0.31	0.63	0.56
## 40	87.6	7.5	85.5	20.1	1.10	10	0.79	1.41	0.89
## 41	114.0	7.0	72.6	6.4	0.16	14	0.04	0.26	0.18
## 42	97.5	6.8	45.5	6.2	0.10	12	0.05	0.26	0.19
## 43	11.8	5.9	24.2	1.6	0.48	10	0.27	1.05	0.44
## 44	66.5	8.3	26.0	68.2	0.21	12	0.05	0.48	0.16
## 45	16.0	6.7	41.2	24.1	0.86	12	0.36	1.40	0.67
## 46	5.0	6.2	23.6	9.6	0.52	12	0.31	0.95	0.55
## 47	25.6	6.2	12.6	27.7	0.65	44	0.30	1.10	0.58
## 48	81.5	8.9	20.5	9.6	0.27	6	0.04	0.40	0.27
## 49	1.2	4.3	2.1	6.4	0.94	10	0.59	1.24	0.98
## 50	34.0	7.0	13.1	4.6	0.40	12	0.08	0.90	0.31
## 51	15.5	6.9	5.2	16.5	0.43	11	0.23	0.69	0.43
## 52	17.3	5.2	3.0	2.6	0.25	12	0.15	0.40	0.28
## 53	71.8	7.9	20.5	8.8	0.27	12	0.15	0.51	0.25

Exploración de la base de datos

1. Calcula medidas estadísticas

Variables cuantitativas

Las variables cuantitativas que se utilizarán para realizar los cálculos son la X3, X4, X5, X6, X7, X8, X9, X10 Y X11.

```
library(modeest)
summary(dbNum)
```

Medidas de tendencia central: promedio, media, mediana y moda de los datos.

```
##           X3           X4           X5           X6
## Min.      : 1.20    Min.    :3.600    Min.      : 1.1    Min.      : 0.70
## 1st Qu.: 6.60    1st Qu.:5.800    1st Qu.: 3.3    1st Qu.: 4.60
## Median : 19.60    Median :6.800    Median :12.6    Median : 12.80
## Mean     : 37.53    Mean     :6.591    Mean     :22.2    Mean     : 23.12
## 3rd Qu.: 66.50    3rd Qu.:7.400    3rd Qu.:35.6    3rd Qu.: 24.70
## Max.     :128.00    Max.      :9.100    Max.      :90.7    Max.     :152.40
##           X7           X8           X9           X10
## Min.      :0.0400    Min.      : 4.00    Min.      :0.0400    Min.      :0.0600
## 1st Qu.:0.2700    1st Qu.:10.00    1st Qu.:0.0900    1st Qu.:0.4800
## Median :0.4800    Median :12.00    Median :0.2500    Median :0.8400
## Mean     :0.5272    Mean      :13.06    Mean      :0.2798    Mean      :0.8745
## 3rd Qu.:0.7700    3rd Qu.:12.00    3rd Qu.:0.3300    3rd Qu.:1.3300
## Max.     :1.3300    Max.      :44.00    Max.      :0.9200    Max.      :2.0400
##           X11
## Min.      :0.0400
## 1st Qu.:0.2500
## Median :0.4500
## Mean     :0.5132
## 3rd Qu.:0.7000
## Max.     :1.5300
```

```
print("Moda: ")
```

```
## [1] "Moda: "
```

```
modeX3 = mlv(X3, method = "mfv")[1] #Moda X3
sprintf("Moda de X3: %s", modeX3)
```

```
## [1] "Moda de X3: 17.3"
```

```
modeX4 = mlv(X4, method = "mfv")[1] #Moda X4
sprintf("Moda de X4: %s", modeX4)
```

```
## [1] "Moda de X4: 5.8"
```

```
modeX5 = mlv(X5, method = "mfv")[1] #Moda X5
sprintf("Moda de X5: %s", modeX5)
```

```
## [1] "Moda de X5: 3"
```

```
modeX6 = mlv(X6, method = "mfv")[1] #Moda X6
sprintf("Moda de X6: %s", modeX6)
```

```
## [1] "Moda de X6: 1.6"
```

```
modeX7 = mlv(X7, method = "mfv")[1] #Moda X7
sprintf("Moda de X7: %s", modeX7)
```

```
## [1] "Moda de X7: 0.34"
```

```
modeX8 = mlv(X8, method = "mfv")[1] #Moda X8
sprintf("Moda de X8: %s", modeX8)
```

```
## [1] "Moda de X8: 12"
```

```
modeX9 = mlv(X9, method = "mfv")[1] #Moda X9
sprintf("Moda de X9: %s", modeX9)
```

```
## [1] "Moda de X9: 0.04"
```

```
modeX10 = mlv(X10, method = "mfv")[1] #Moda X10
sprintf("Moda de X10: %s", modeX10)
```

```
## [1] "Moda de X10: 0.06"
```

```
modeX11 = mlv(X11, method = "mfv")[1] #Moda X11
sprintf("Moda de X11: %s", modeX11)
```

```
## [1] "Moda de X11: 0.16"
```

```
summary(dbNum)
```

Medidas de dispersión: rango: máximo - mínimo, varianza, desviación estándar.

##	X3	X4	X5	X6
## Min.	: 1.20	Min. :3.600	Min. : 1.1	Min. : 0.70
## 1st Qu.:	6.60	1st Qu.:5.800	1st Qu.: 3.3	1st Qu.: 4.60
## Median	: 19.60	Median :6.800	Median :12.6	Median : 12.80

```
## Mean : 37.53 Mean :6.591 Mean :22.2 Mean : 23.12
## 3rd Qu.: 66.50 3rd Qu.:7.400 3rd Qu.:35.6 3rd Qu.: 24.70
## Max. :128.00 Max. :9.100 Max. :90.7 Max. :152.40
## X7 X8 X9 X10
## Min. :0.0400 Min. : 4.00 Min. :0.0400 Min. :0.0600
## 1st Qu.:0.2700 1st Qu.:10.00 1st Qu.:0.0900 1st Qu.:0.4800
## Median :0.4800 Median :12.00 Median :0.2500 Median :0.8400
## Mean :0.5272 Mean :13.06 Mean :0.2798 Mean :0.8745
## 3rd Qu.:0.7700 3rd Qu.:12.00 3rd Qu.:0.3300 3rd Qu.:1.3300
## Max. :1.3300 Max. :44.00 Max. :0.9200 Max. :2.0400
## X11
## Min. :0.0400
## 1st Qu.:0.2500
## Median :0.4500
## Mean :0.5132
## 3rd Qu.:0.7000
## Max. :1.5300
```

```
print("Varianza: ")
```

```
## [1] "Varianza: "
```

```
apply(dbNum, 2, var)
```

```
## X3 X4 X5 X6 X7 X8
## 1.459509e+03 1.660102e+00 6.216333e+02 9.496457e+02 1.163053e-01 7.328520e+01
## X9 X10 X11
## 5.125958e-02 2.725329e-01 1.147376e-01
```

```
print("")
```

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

```
print("Desviación estándar: ")
```

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

```
apply(dbNum, 2, sd)
```

```
## X3 X4 X5 X6 X7 X8 X9
## 38.2035267 1.2884493 24.9325744 30.8163214 0.3410356 8.5606773 0.2264058
## X10 X11
## 0.5220469 0.3387294
```

Variables cualitativas

La variable cualitativa que se utilizará es X12 que es el indicador de la edad de los peces.

```
print("Tabla de distribución de frecuencia de X12:")
```

Tabla de distribución de frecuencia y moda

```
## [1] "Tabla de distribución de frecuencia de X12:"
```

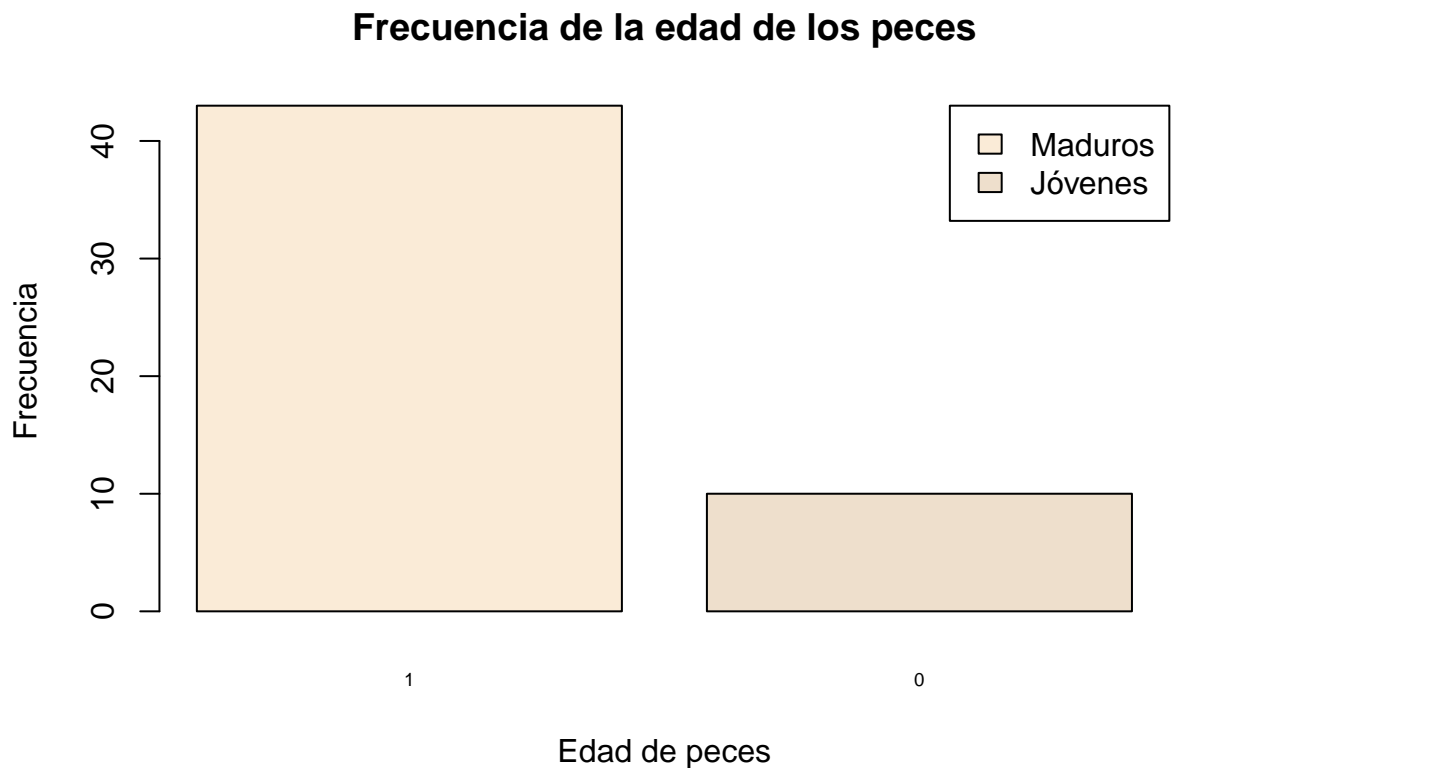
```
X12_table = table(X12)
print(X12_table)
```

```
## X12
##  0  1
## 10 43
```

```
modeX12 = mlv(X12, method = "mfv")[1] #Moda
sprintf("Moda de X12: %s", modeX12)
```

```
## [1] "Moda de X12: 1"
```

```
sorted_table = sort(X12_table, decreasing = TRUE)[1:2]
barplot(sorted_table, width = 1, cex.names = 0.6, xlab="Edad de peces", ylab="Frecuencia", col = c("ant", "ant"))
```



En la tabla y gráfica anterior de distribución de frecuencia se muestra la variable que contiene la edad de los peces. Se puede observar que existen 43 peces maduros y 10 jóvenes.

2. Explora los datos usando herramientas de visualización

Variables cuantitativas:

```
print("Cuartiles de X3")
```

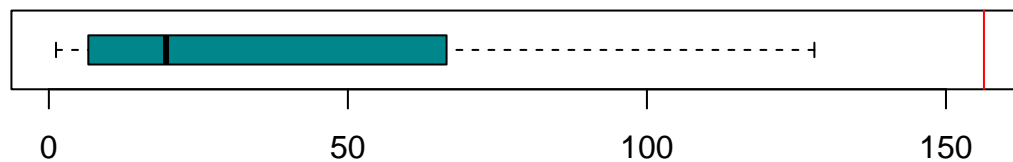
Medidas de posición: cuartiles, outlier (valores atípicos), boxplots

```
## [1] "Cuartiles de X3"
```

```
q1_c=quantile(X3,0.25) #Cuantil 1
q3_c = quantile(X3, 0.75) #Cuantil 3
ri_c= IQR(X3) #Rango intercuartílico
y2 = q3_c+1.5*ri_c
par(mfrow=c(2,1)) #Matriz de gráficos de 2x1
boxplot(X3,horizontal=TRUE,ylim=c(0,y2),main="Boxplot de X3 (alcalinidad)", col="turquoise4")
abline(v=q3_c+1.5*ri_c,col="red") #línea vertical en el límite de los datos atípicos
X = db[X3<q3_c+1.5*ri_c,c("X3")] #Quitar datos atípicos de la matriz M en la variable X
summary(X)
```

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

Boxplot de X3 (alcalinidad)



En la gráfica anterior podemos observar que se tiene una distribución de sesgo a la derecha, ya que la

mayoría de los datos se concentran en la parte izquierda de la distribución. Por lo tanto, es una distribución asimétrica.

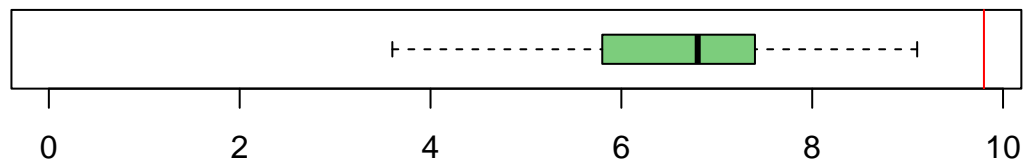
```
print("Cuartiles de X4")
```

```
## [1] "Cuartiles de X4"
```

```
q1_c=quantile(X4,0.25) #Cuantil 1
q3_c = quantile(X4, 0.75) #Cuantil 3
ri_c= IQR(X4) #Rango intercuartílico
y2 = q3_c+1.5*ri_c
par(mfrow=c(2,1)) #Matriz de gráficos de 2x1
boxplot(X4,horizontal=TRUE,ylim=c(0,y2),main="Boxplot de X4 (PH)", col="palegreen3")
abline(v=q3_c+1.5*ri_c,col="red") #línea vertical en el límite de los datos atípicos
X = db[X4<q3_c+1.5*ri_c,c("X4")] #Quitar datos atípicos de la matriz M en la variable X
summary(X)
```

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

Boxplot de X4 (PH)



En la gráfica anterior podemos observar que se tiene una distribución simétrica.

```
print("Cuartiles de X5")
```

```
## [1] "Cuartiles de X5"
```



```

q1_c=quantile(X5,0.25) #Cuantil 1
q3_c = quantile(X5, 0.75) #Cuantil 3
ri_c= IQR(X5) #Rango intercuartílico
y2 = q3_c+1.5*ri_c
par(mfrow=c(2,1)) #Matriz de gráficos de 2x1
boxplot(X5,horizontal=TRUE,ylim=c(0,y2),main="Boxplot de X5 (calcio)", col="lightpink2")
abline(v=q3_c+1.5*ri_c,col="red") #línea vertical en el límite de los datos atípicos
X = db[X5<q3_c+1.5*ri_c,c("X5")] #Quitar datos atípicos de la matriz M en la variable X
summary(X)

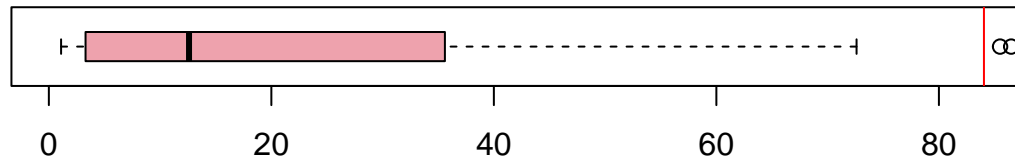
```

```

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.10   3.30   10.45   18.28   24.95   72.60

```

Boxplot de X5 (calcio)



En la gráfica anterior podemos observar que se tiene una distribución asimétrica con sesgo a la derecha, ya que la mayoría de los datos se concentran en la parte izquierda de la distribución. Asimismo, se tienen datos atípicos que van más allá del límite derecho.

```

print("Cuartiles de X6")

```

```

## [1] "Cuartiles de X6"

```

```

q1_c=quantile(X6,0.25) #Cuantil 1
q3_c = quantile(X6, 0.75) #Cuantil 3
ri_c= IQR(X6) #Rango intercuartílico
y2 = q3_c+1.5*ri_c

```

```

par(mfrow=c(2,1)) #Matriz de gráficos de 2x1
boxplot(X6,horizontal=TRUE,ylim=c(0,y2),main="Boxplot de X6 (clorofila)", col="sienna2")
abline(v=q3_c+1.5*ri_c,col="red") #línea vertical en el límite de los datos atípicos
X = db[X6<q3_c+1.5*ri_c,c("X6")] #Quitar datos atípicos de la matriz M en la variable X
summary(X)

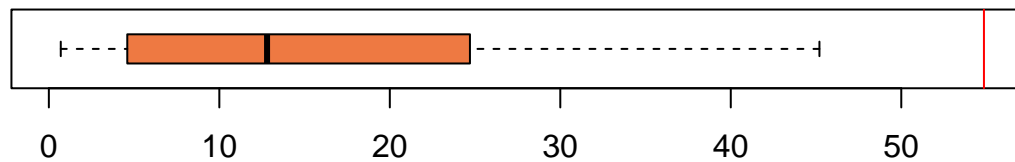
```

```

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

```

Boxplot de X6 (clorofila)



En la gráfica anterior podemos observar que se tiene una distribución de sesgo a la derecha, ya que la mayoría de los datos se concentran en la parte izquierda de la distribución. Por lo tanto, es una distribución asimétrica.

```

print("Cuartiles de X7")

```

```

## [1] "Cuartiles de X7"

```

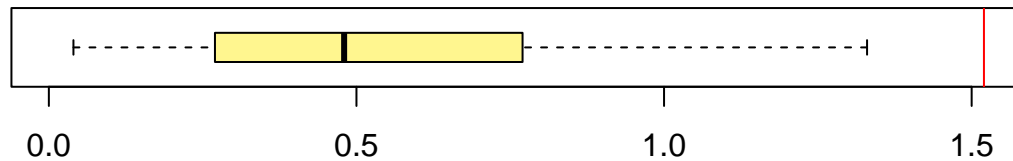
```

q1_c=quantile(X7,0.25) #Cuantil 1
q3_c = quantile(X7, 0.75) #Cuantil 3
ri_c= IQR(X7) #Rango intercuartílico
y2 = q3_c+1.5*ri_c
par(mfrow=c(2,1)) #Matriz de gráficos de 2x1
boxplot(X7,horizontal=TRUE,ylim=c(0,y2),main="Boxplot de X7 (concentración de mercurio en el tejido muscular)", col="sienna2")
abline(v=q3_c+1.5*ri_c,col="red") #línea vertical en el límite de los datos atípicos
X = db[X7<q3_c+1.5*ri_c,c("X7")] #Quitar datos atípicos de la matriz M en la variable X
summary(X)

```

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

Boxplot de X7 (concentración de mercurio en el tejido muscular)



En la gráfica anterior podemos observar que se tiene una distribución de sesgo a la derecha, ya que la mayoría de los datos se concentran en la parte izquierda de la distribución. Por lo tanto, es una distribución asimétrica.

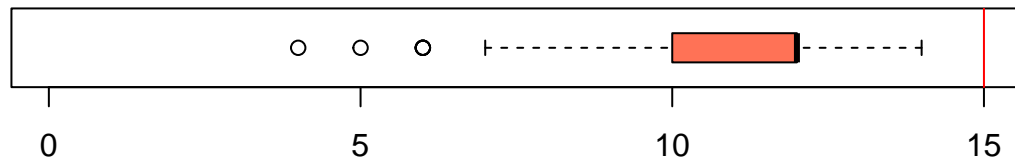
```
print("Cuartiles de X8")
```

```
## [1] "Cuartiles de X8"
```

```
q1_c=quantile(X8,0.25) #Cuantil 1
q3_c = quantile(X8, 0.75) #Cuantil 3
ri_c= IQR(X8) #Rango intercuartílico
y2 = q3_c+1.5*ri_c
par(mfrow=c(2,1)) #Matriz de gráficos de 2x1
boxplot(X8,horizontal=TRUE,ylim=c(0,y2),main="Boxplot de X8 (peces estudiados en el lago)", col="coral1",
abline(v=q3_c+1.5*ri_c,col="red") #línea vertical en el límite de los datos atípicos
X = db[X8<q3_c+1.5*ri_c,c("X8")] #Quitar datos atípicos de la matriz M en la variable X
summary(X)
```

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

Boxplot de X8 (peces estudiados en el lago)



En la gráfica anterior podemos observar que se tiene una distribución asimétrica.

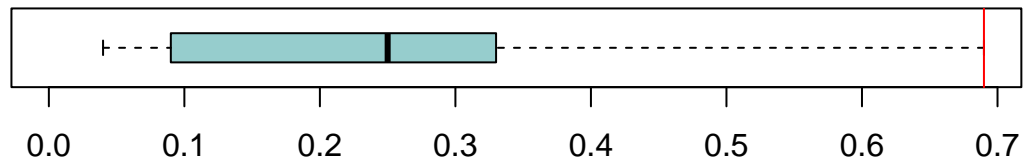
```
print("Cuartiles de X9")
```

```
## [1] "Cuartiles de X9"
```

```
q1_c=quantile(X9,0.25) #Cuantil 1
q3_c = quantile(X9, 0.75) #Cuantil 3
ri_c= IQR(X9) #Rango intercuartílico
y2 = q3_c+1.5*ri_c
par(mfrow=c(2,1)) #Matriz de gráficos de 2x1
boxplot(X9,horizontal=TRUE,ylim=c(0,y2),main="Boxplot de X9 (mínimo de la concentración de mercurio)", col="red")
abline(v=q3_c+1.5*ri_c,col="red") #línea vertical en el límite de los datos atípicos
X = db[X9<q3_c+1.5*ri_c,c("X9")] #Quitar datos atípicos de la matriz M en la variable X
summary(X)
```

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

Boxplot de X9 (mínimo de la concentración de mercurio)



En la gráfica anterior podemos observar que se tiene una distribución de sesgo a la derecha, ya que la mayoría de los datos se concentran en la parte izquierda de la distribución. Por lo tanto, es una distribución asimétrica.

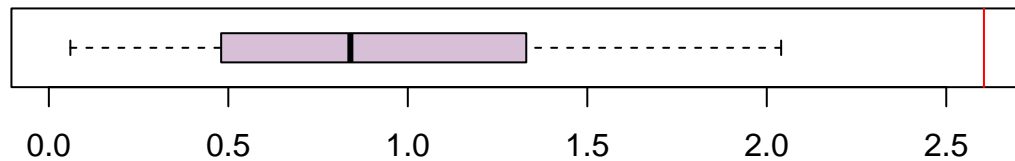
```
print("Cuartiles de X10")
```

```
## [1] "Cuartiles de X10"
```

```
q1_c=quantile(X10,0.25) #Cuantil 1
q3_c = quantile(X10, 0.75) #Cuantil 3
ri_c= IQR(X10) #Rango intercuartílico
y2 = q3_c+1.5*ri_c
par(mfrow=c(2,1)) #Matriz de gráficos de 2x1
boxplot(X10,horizontal=TRUE,ylim=c(0,y2),main="Boxplot de X10 (estimación de la concentración de mercurio)")
abline(v=q3_c+1.5*ri_c,col="red") #línea vertical en el límite de los datos atípicos
X = db[X10<q3_c+1.5*ri_c,c("X10")] #Quitar datos atípicos de la matriz M en la variable X
summary(X)
```

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

Boxplot de X10 (estimación de la concentración de mercurio)



En la gráfica anterior podemos observar que se tiene una distribución de sesgo a la derecha, ya que la mayoría de los datos se concentran en la parte izquierda de la distribución. Por lo tanto, es una distribución asimétrica.

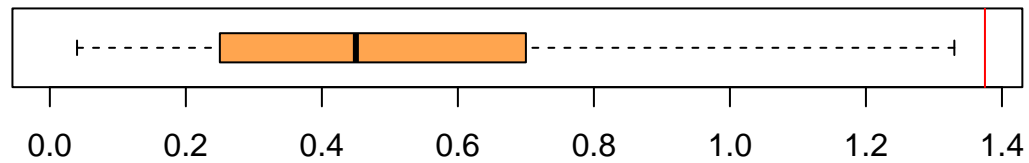
```
print("Cuartiles de X11")
```

```
## [1] "Cuartiles de X11"
```

```
q1_c=quantile(X11,0.25) #Cuantil 1
q3_c = quantile(X11, 0.75) #Cuantil 3
ri_c= IQR(X11) #Rango intercuartílico
y2 = q3_c+1.5*ri_c
par(mfrow=c(2,1)) #Matriz de gráficos de 2x1
boxplot(X11,horizontal=TRUE,ylim=c(0,y2),main="Boxplot de X11 (máximo de la concentración de mercurio)")
abline(v=q3_c+1.5*ri_c,col="red") #línea vertical en el límite de los datos atípicos
X = db[X11<q3_c+1.5*ri_c,c("X11")] #Quitar datos atípicos de la matriz M en la variable X
summary(X)
```

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

Boxplot de X11 (máximo de la concentración de mercurio)



En la gráfica anterior podemos observar que se tiene una distribución de sesgo a la derecha, ya que la mayoría de los datos se concentran en la parte izquierda de la distribución. Por lo tanto, es una distribución asimétrica.

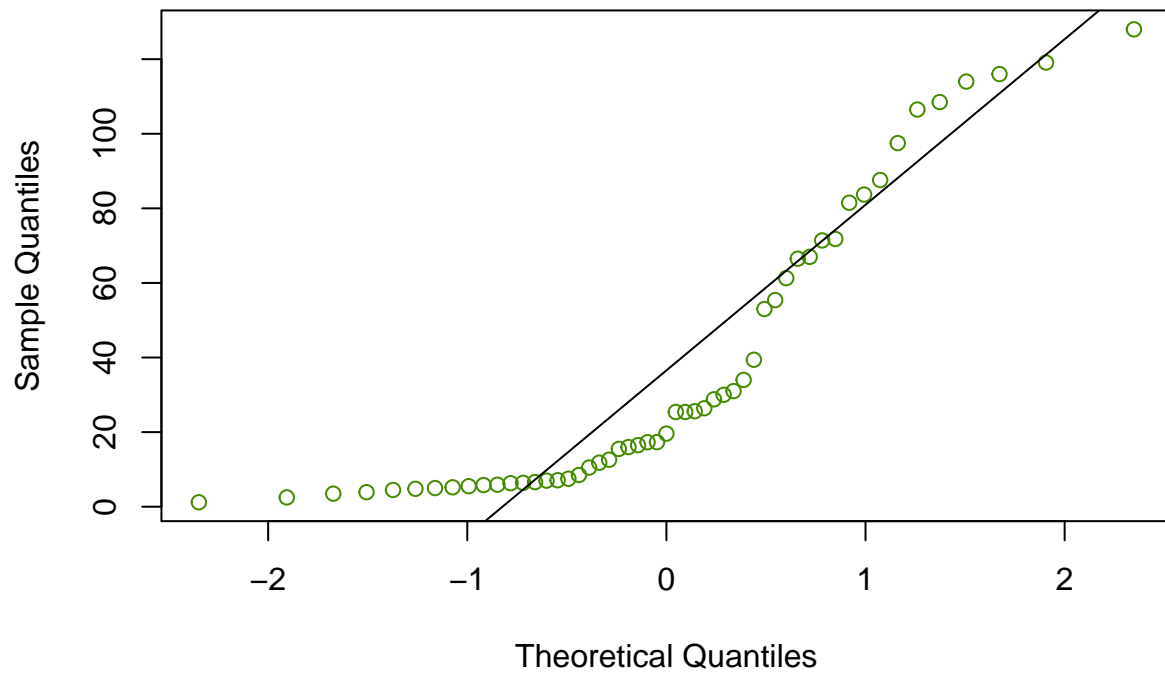
```
print("Cuartiles de X3")
```

Análisis de distribución de los datos (Histogramas). Identificar si tiene forma simétrica o asimétrica

```
## [1] "Cuartiles de X3"
```

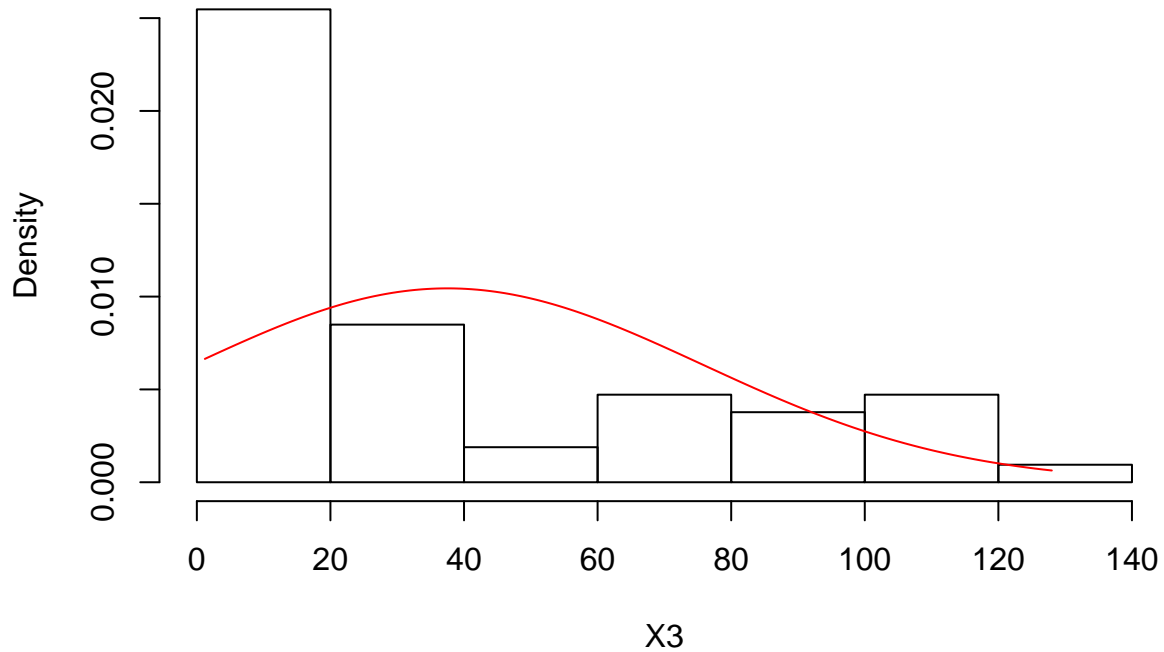
```
qqnorm(X3, main = "Normal Q-Q Plot de X3 (alcalinidad)", col="chartreuse4")  
qqline(X3)
```

Normal Q-Q Plot de X3 (alcalinidad)



```
hist(X3,prob=TRUE,col=0, main = "Histograma de X3 (alcalinidad)")  
x=seq(min(X3),max(X3),0.1)  
y=dnorm(x,mean(X3),sd(X3))  
lines(x,y,col="red")
```


Histograma de X3 (alcalinidad)



```
library(moments)
```

```
##  
## Attaching package: 'moments'  
  
## The following object is masked from 'package:modeest':  
##  
##      skewness
```

```
skewness(X3)
```

```
## [1] 0.9959715
```

```
kurtosis(X3)
```

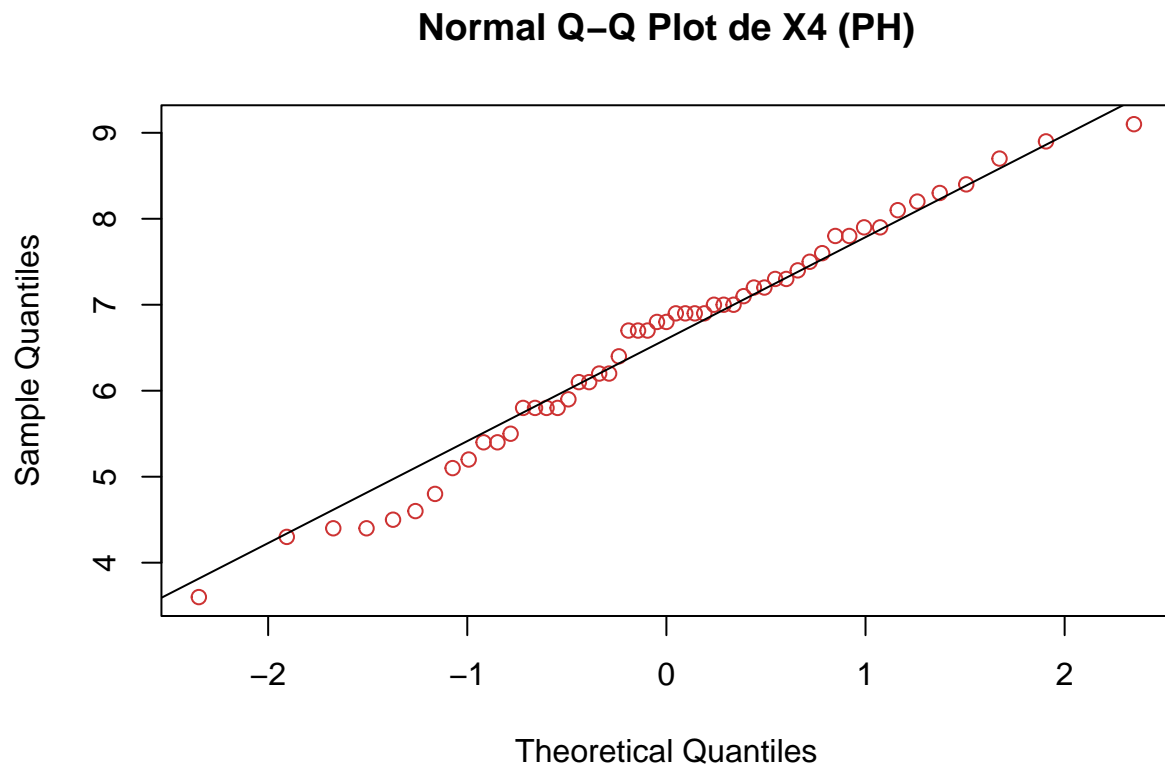
```
## [1] 2.627688
```

La gráfica de Q-Q Plot muestra que tiene una distribución con colas delgadas, es decir, tiene una alta curtosis y una distribución Leptocúrtica. Por otro lado, en el histograma se muestra una distribución asimétrica con sesgo a la derecha.

```
print("Cuartiles de X4")
```

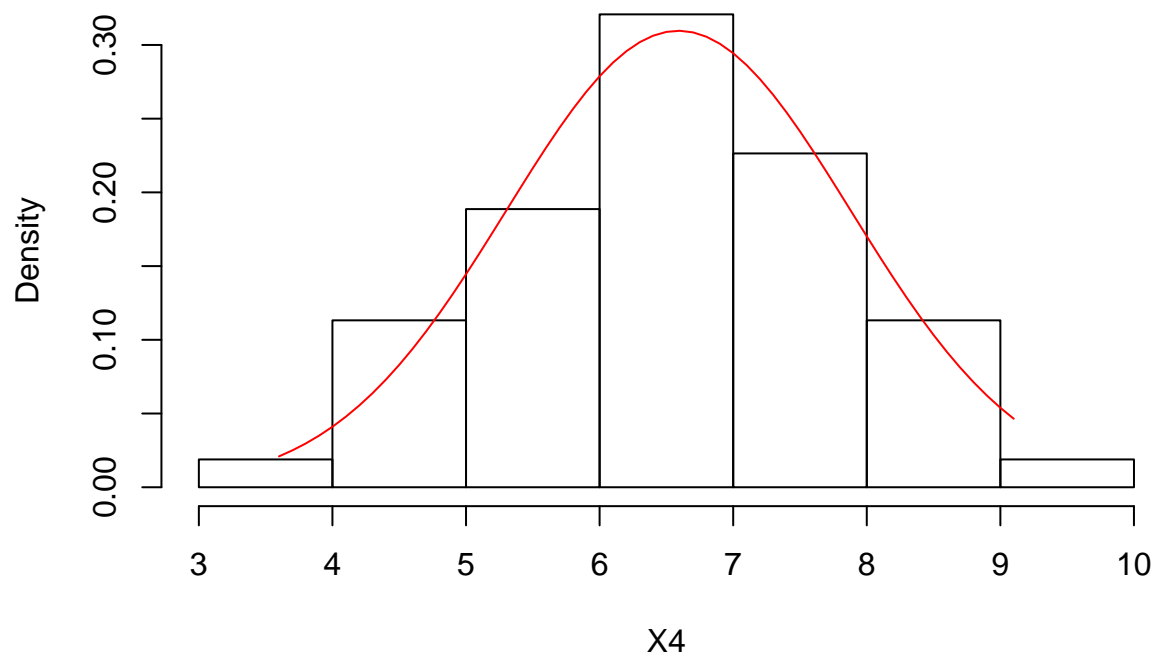
```
## [1] "Cuartiles de X4"
```

```
qqnorm(X4, main = "Normal Q-Q Plot de X4 (PH)", col="brown3")  
qqline(X4)
```



```
hist(X4,prob=TRUE,col=0, main = "Histograma de X4 (PH)")  
x=seq(min(X4),max(X4),0.1)  
y=dnorm(x,mean(X4),sd(X4))  
lines(x,y,col="red")
```

Histograma de X4 (PH)



```
skewness(X4)
```

```
## [1] -0.2530037
```

```
kurtosis(X4)
```

```
## [1] 2.468301
```

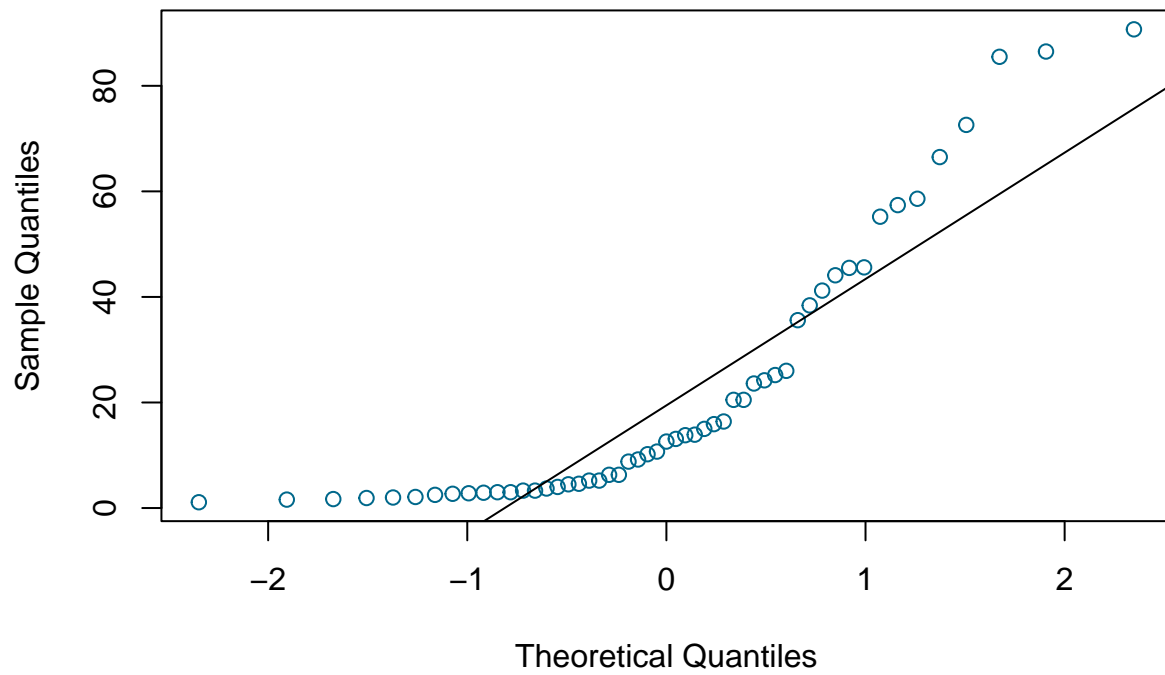
La gráfica de Q-Q Plot muestra que tiene una distribución casi ideal. Por otro lado, en el histograma se muestra una distribución simétrica.

```
print("Cuartiles de X5")
```

```
## [1] "Cuartiles de X5"
```

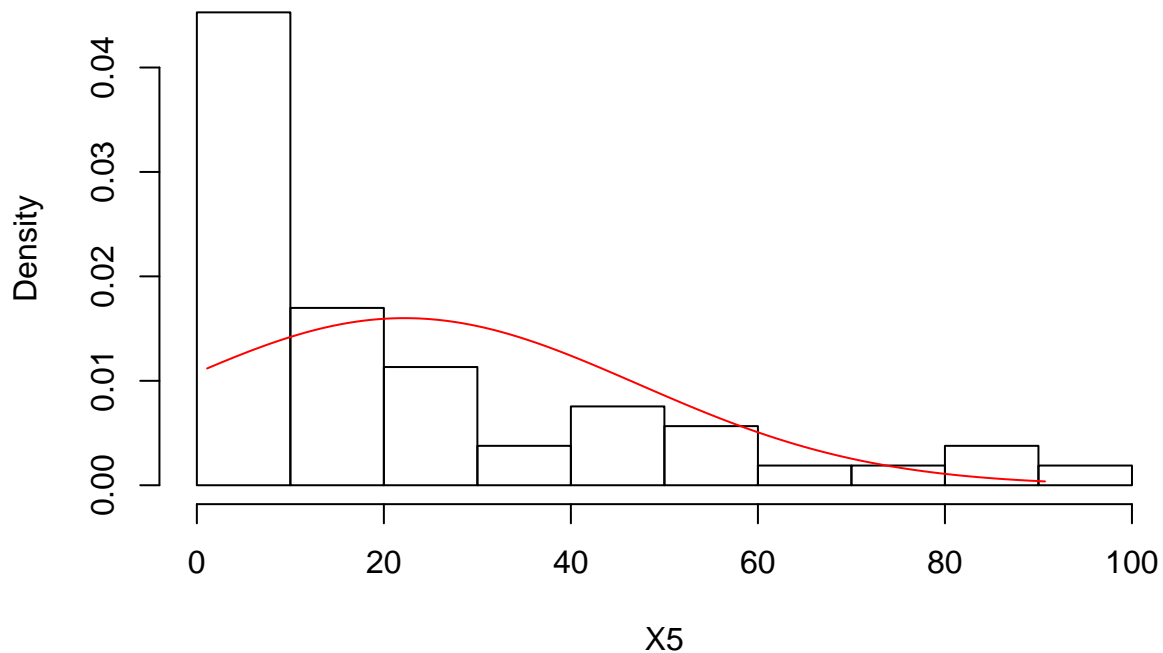
```
qqnorm(X5, main = "Normal Q-Q Plot de X5 (calcio)", col="deepskyblue4")  
qqline(X5)
```

Normal Q-Q Plot de X5 (calcio)



```
hist(X5,prob=TRUE,col=0, main = "Histograma de X5 (calcio)")
x=seq(min(X5),max(X5),0.1)
y=dnorm(x,mean(X5),sd(X5))
lines(x,y,col="red")
```

Histograma de X5 (calcio)



```
skewness(X5)
```

```
## [1] 1.342399
```

```
kurtosis(X5)
```

```
## [1] 3.753335
```

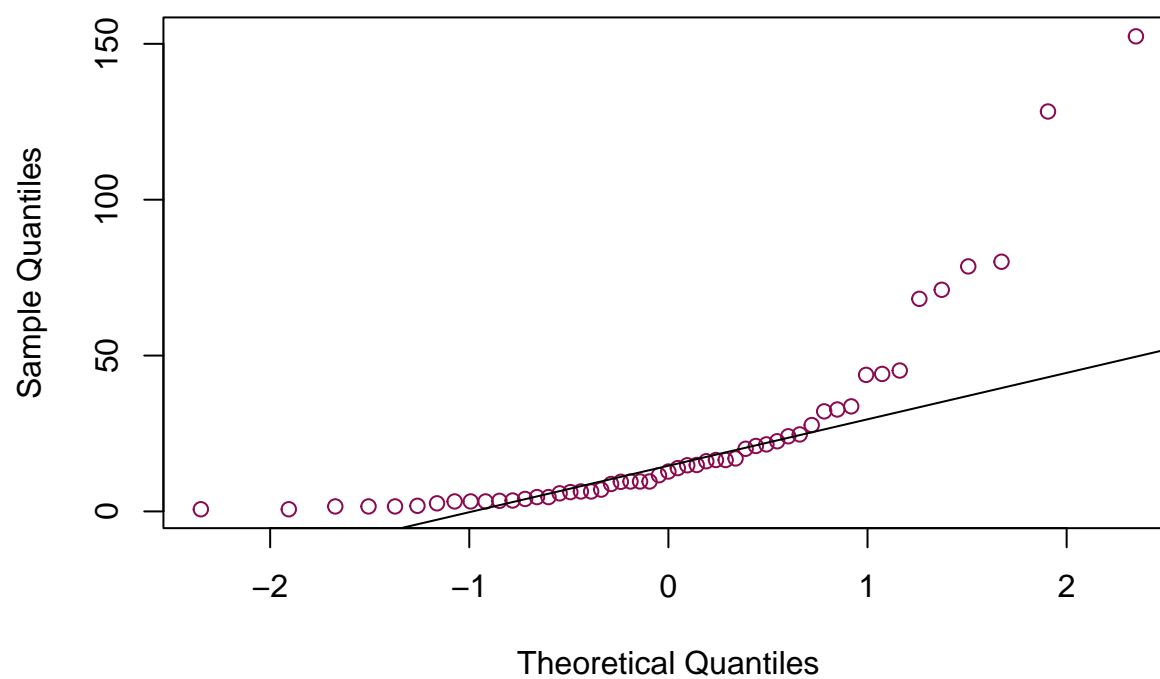
La gráfica de Q-Q Plot muestra que tiene una distribución con colas gruesas, es decir, tiene una baja curtosis y una distribución platycúrtica. Por otro lado, en el histograma se muestra una distribución asimétrica con sesgo a la derecha.

```
print("Cuartiles de X6")
```

```
## [1] "Cuartiles de X6"
```

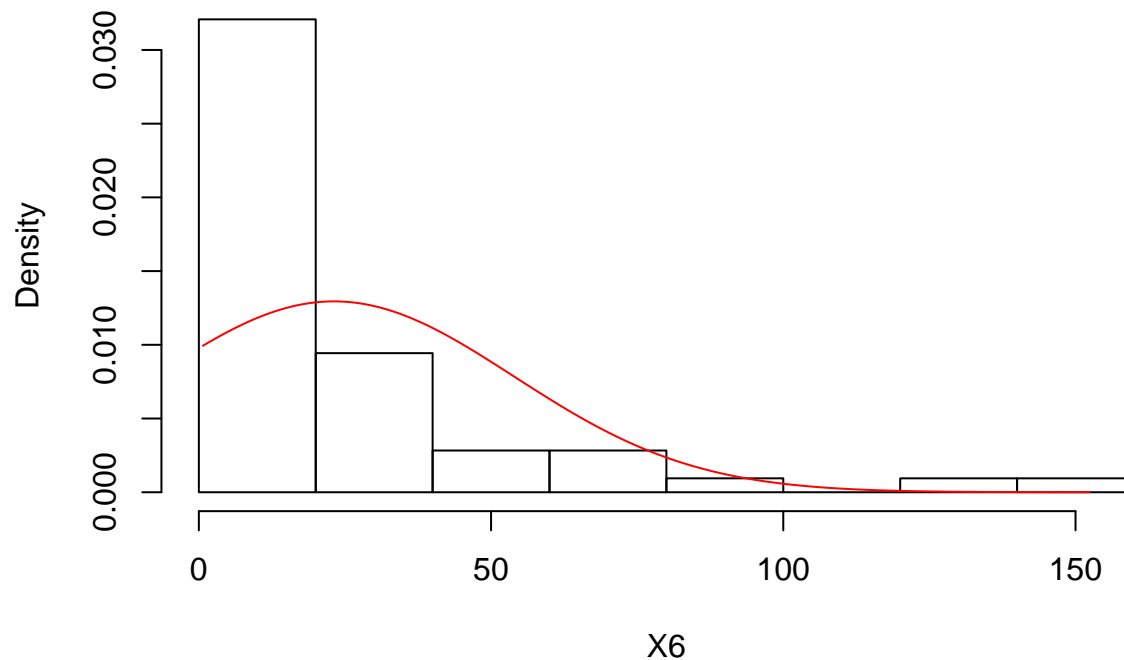
```
qqnorm(X6, main = "Normal Q-Q Plot de X6 (clorofila)", col="deeppink4")  
qqline(X6)
```

Normal Q-Q Plot de X6 (clorofila)



```
hist(X6,prob=TRUE,col=0, main = "Histograma de X6 (clorofila)")
x=seq(min(X6),max(X6),0.1)
y=dnorm(x,mean(X6),sd(X6))
lines(x,y,col="red")
```

Histograma de X6 (clorofila)



```
skewness(X6)
```

```
## [1] 2.482998
```

```
kurtosis(X6)
```

```
## [1] 9.457748
```

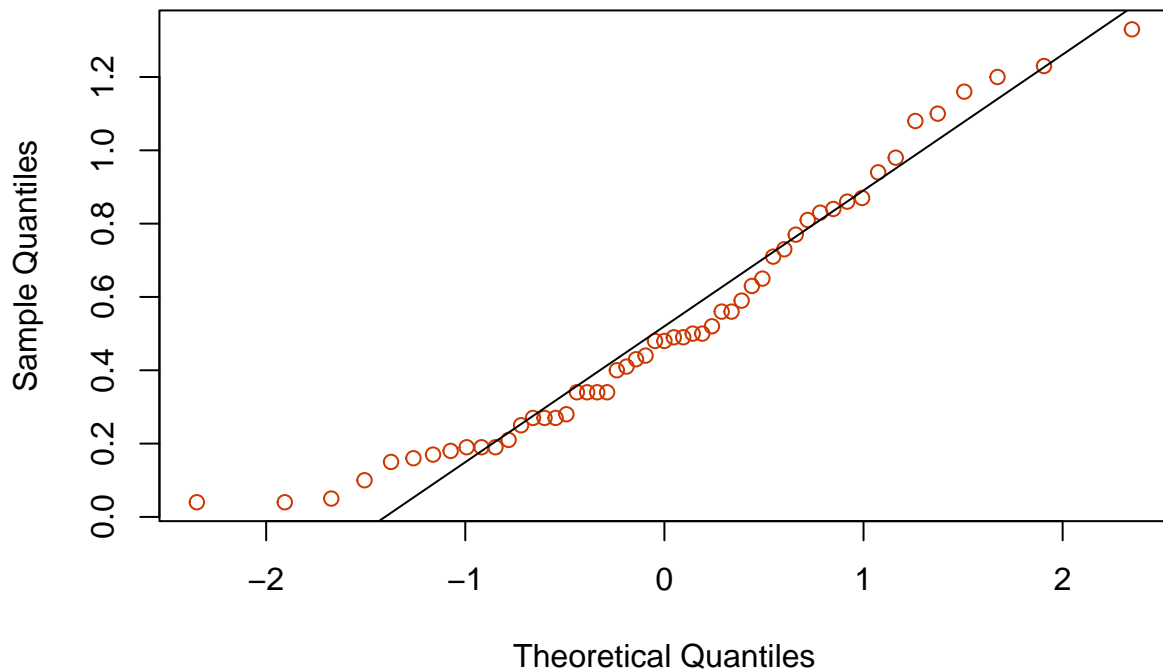
La gráfica de Q-Q Plot muestra que tiene una distribución con colas gruesas, es decir, tiene una baja curtosis y una distribución platycúrtica. Por otro lado, en el histograma se muestra una distribución asimétrica con sesgo a la derecha.

```
print("Cuartiles de X7")
```

```
## [1] "Cuartiles de X7"
```

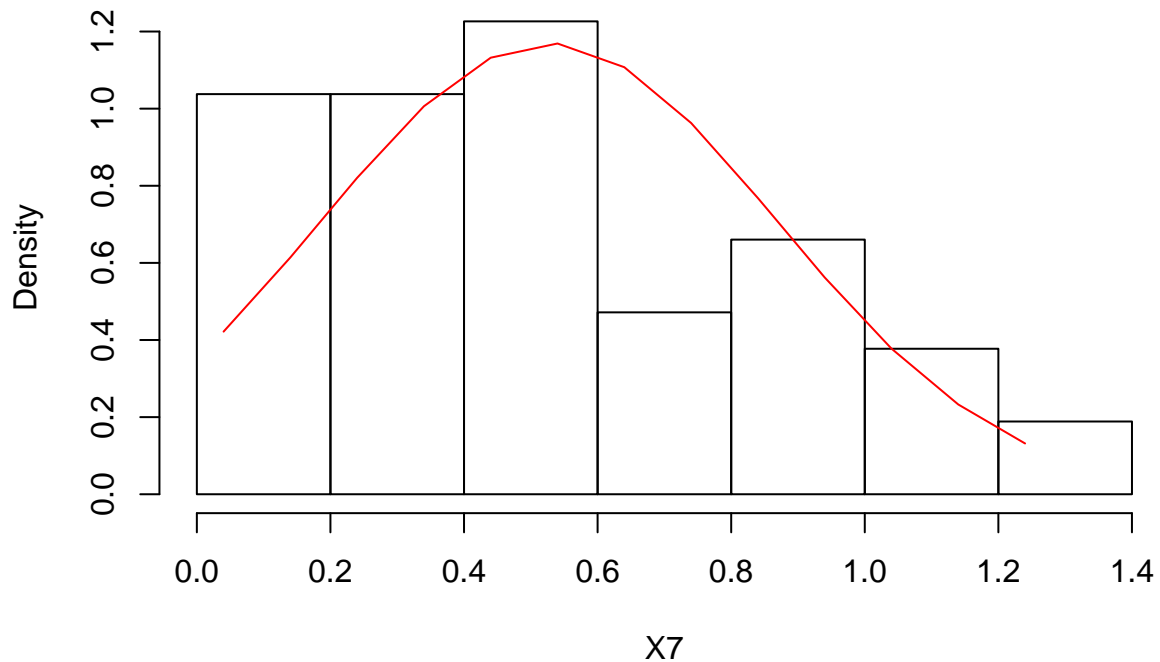
```
qqnorm(X7, main = "Normal Q-Q Plot de X7 (concentración de mercurio en el tejido muscular)", col="orange", las=1)  
qqline(X7)
```

Normal Q-Q Plot de X7 (concentración de mercurio en el tejido muscular)



```
hist(X7,prob=TRUE,col=0, main = "Histograma de X7 (concentración de mercurio en el tejido muscular)")
x=seq(min(X7),max(X7),0.1)
y=dnorm(x,mean(X7),sd(X7))
lines(x,y,col="red")
```


Histograma de X7 (concentración de mercurio en el tejido muscular)



```
skewness(X7)
```

```
## [1] 0.6159853
```

```
kurtosis(X7)
```

```
## [1] 2.460721
```

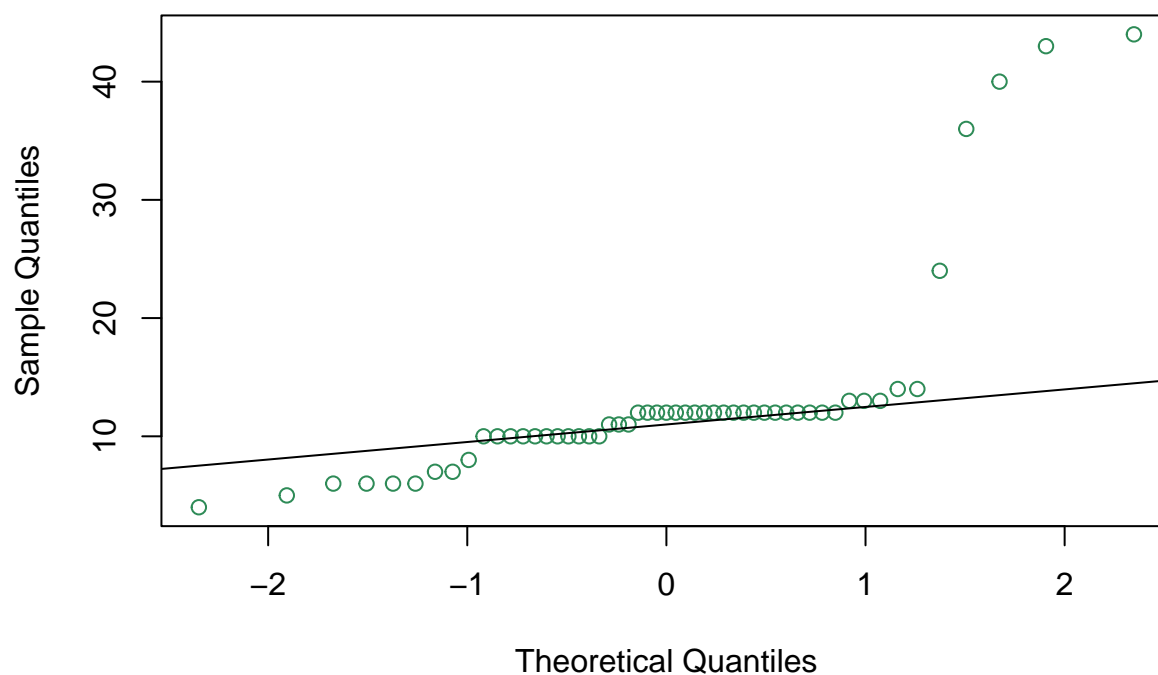
La gráfica de Q-Q Plot muestra que tiene una distribución con colas delgadas, es decir, tiene una alta curtosis y una distribución Leptocúrtica. Por otro lado, en el histograma se muestra una distribución casi simétrica con ligero sesgo a la derecha.

```
print("Cuartiles de X8")
```

```
## [1] "Cuartiles de X8"
```

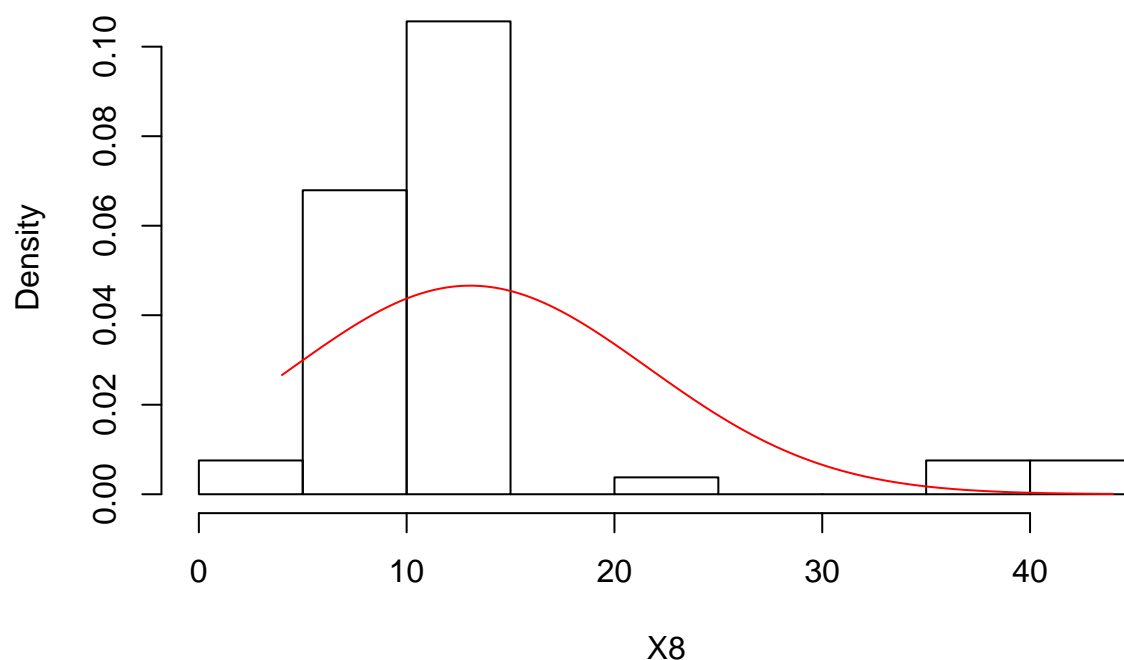
```
qqnorm(X8, main = "Normal Q-Q Plot de X8 (peces estudiados en el lago)", col="seagreen")  
qqline(X8)
```

Normal Q–Q Plot de X8 (peces estudiados en el lago)



```
hist(X8,prob=TRUE,col=0, main = "Histograma de X8 (peces estudiados en el lago)")  
x=seq(min(X8),max(X8),0.1)  
y=dnorm(x,mean(X8),sd(X8))  
lines(x,y,col="red")
```

Histograma de X8 (peces estudiados en el lago)



```
skewness(X8)
```

```
## [1] 2.655682
```

```
kurtosis(X8)
```

```
## [1] 9.358775
```

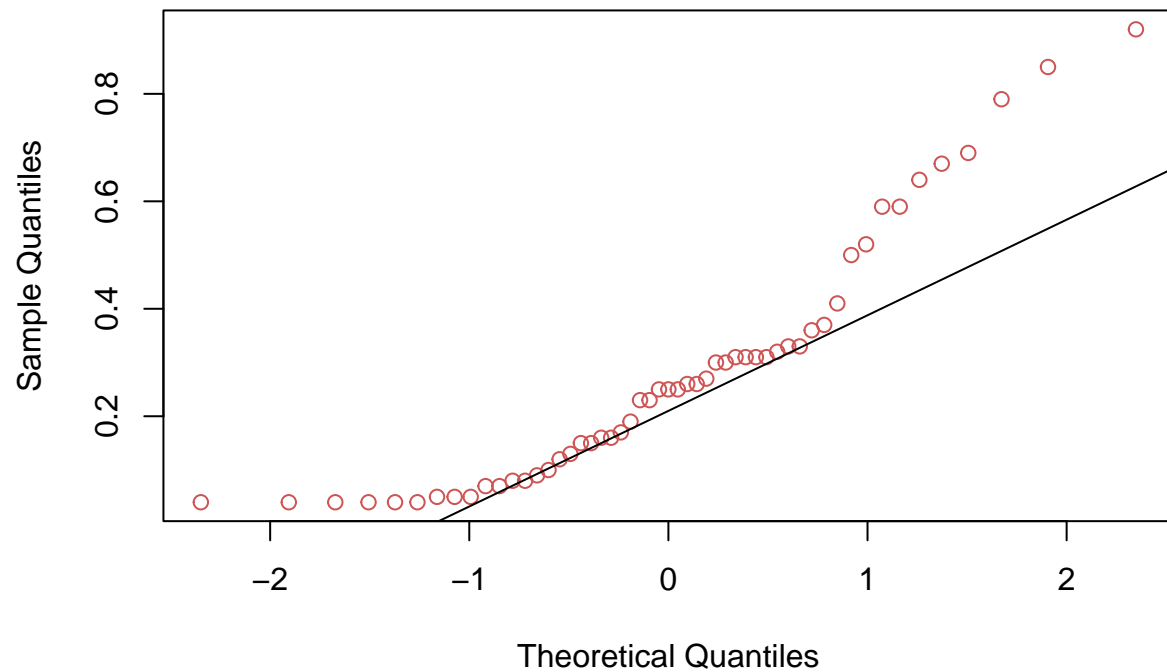
La gráfica de Q-Q Plot muestra que tiene una distribución con colas gruesas, es decir, tiene una baja curtosis y una distribución platycúrtica. Por otro lado, en el histograma se muestra una distribución asimétrica con sesgo a la derecha.

```
print("Cuartiles de X9")
```

```
## [1] "Cuartiles de X9"
```

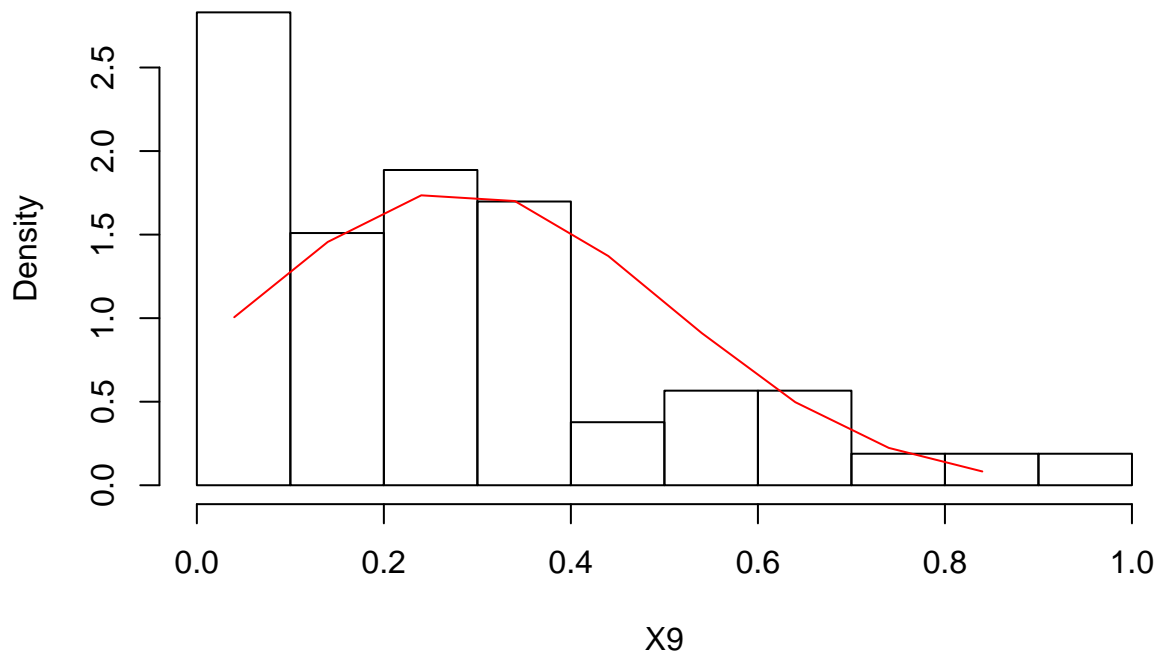
```
qqnorm(X9, main = "Normal Q-Q Plot de X9 (mínimo de la concentración de mercurio)", col="indianred3")  
qqline(X9)
```

Normal Q-Q Plot de X9 (mínimo de la concentración de mercurio)



```
hist(X9,prob=TRUE,col=0, main = "Histograma de X9 (mínimo de la concentración de mercurio)")
x=seq(min(X9),max(X9),0.1)
y=dnorm(x,mean(X9),sd(X9))
lines(x,y,col="red")
```

Histograma de X9 (mínimo de la concentración de mercurio)



```
skewness(X9)
```

```
## [1] 1.104008
```

```
kurtosis(X9)
```

```
## [1] 3.538346
```

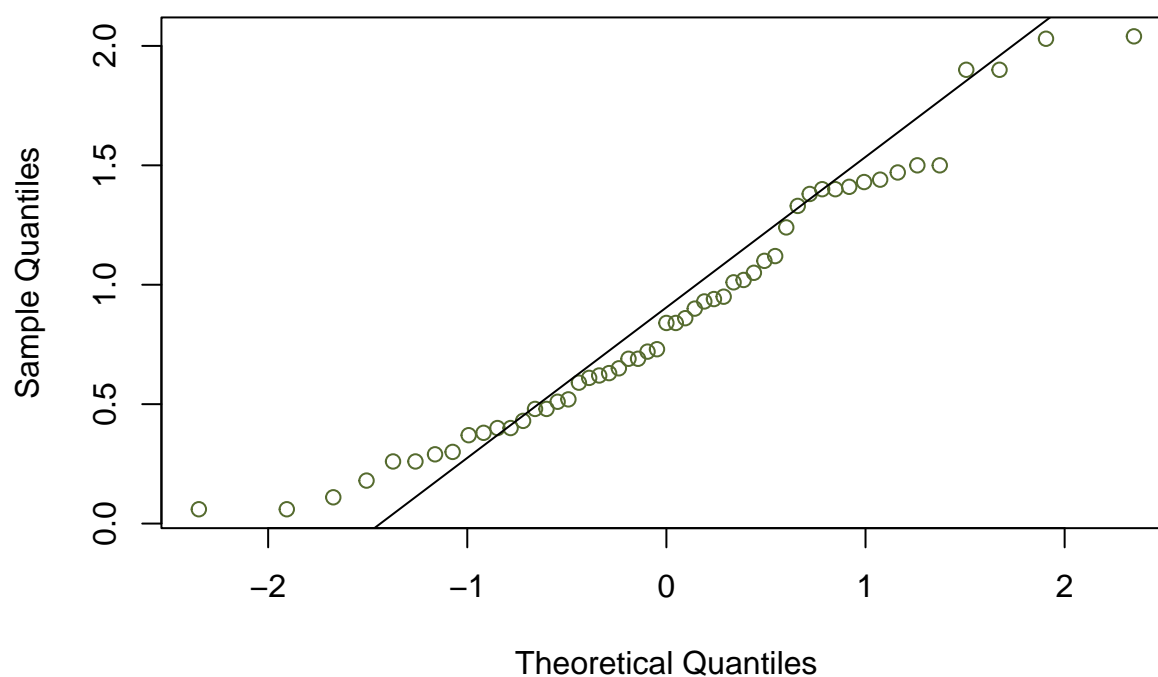
La gráfica de Q-Q Plot muestra que tiene una distribución con colas gruesas, es decir, tiene una baja curtosis y una distribución platicúrtica. Por otro lado, en el histograma se muestra una distribución asimétrica con sesgo a la derecha.

```
print("Cuartiles de X10")
```

```
## [1] "Cuartiles de X10"
```

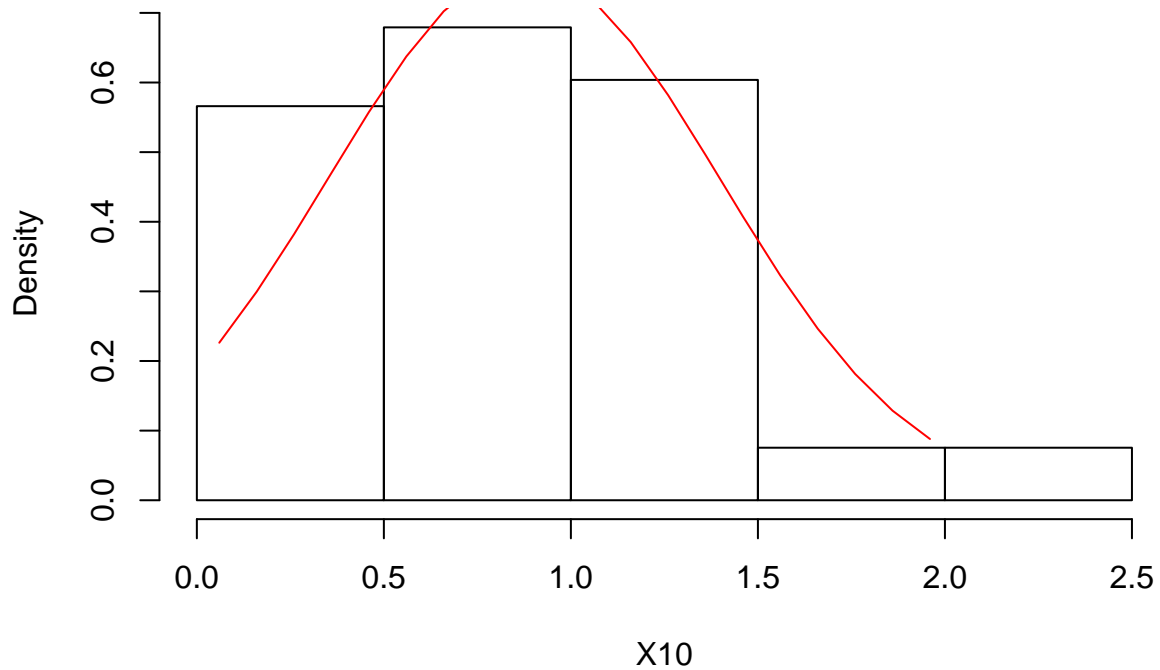
```
qqnorm(X10, main = "Normal Q-Q Plot de X10 (máximo de la concentración de mercurio)", col="darkolivegreen4",  
qqline(X10))
```

Normal Q-Q Plot de X10 (máximo de la concentración de mercurio)



```
hist(X10,prob=TRUE,col=0, main = "Histograma de X10 (máximo de la concentración de mercurio)")
x=seq(min(X10),max(X10),0.1)
y=dnorm(x,mean(X10),sd(X10))
lines(x,y,col="red")
```

Histograma de X10 (máximo de la concentración de mercurio)



```
skewness(X10)
```

```
## [1] 0.4780584
```

```
kurtosis(X10)
```

```
## [1] 2.421257
```

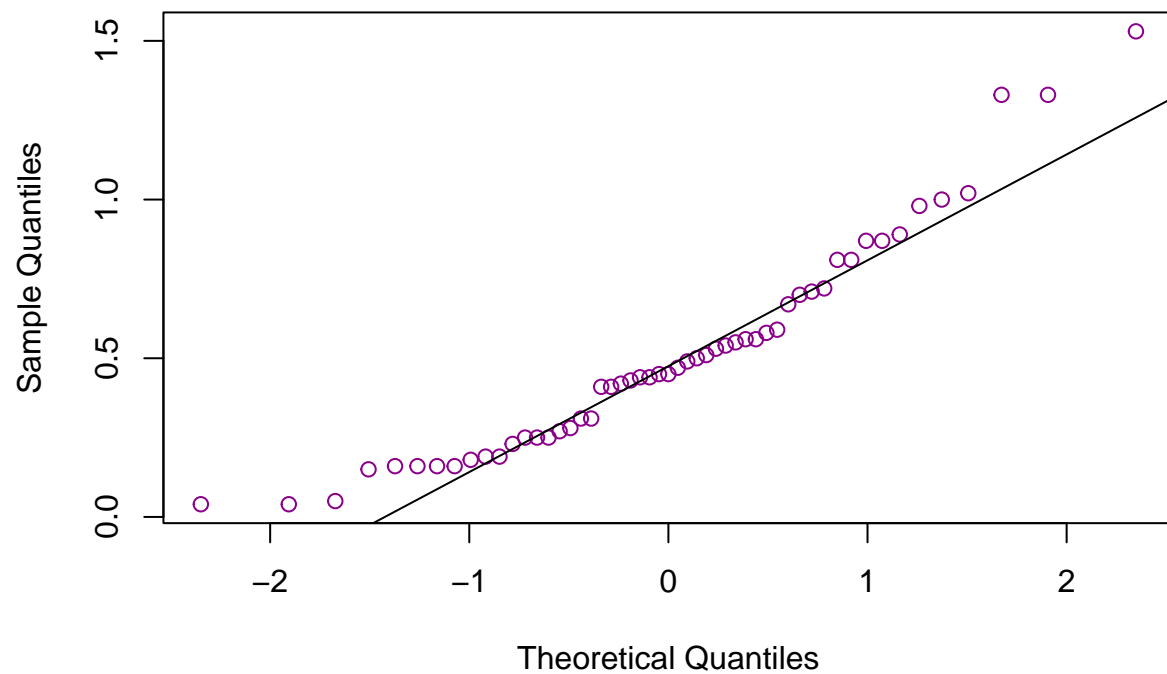
La gráfica de Q-Q Plot muestra que tiene una distribución con colas delgadas, es decir, tiene una alta curtosis y una distribución leptocúrtica. Por otro lado, en el histograma se muestra una distribución casi simétrica con un ligero sesgo a la derecha.

```
print("Cuartiles de X11")
```

```
## [1] "Cuartiles de X11"
```

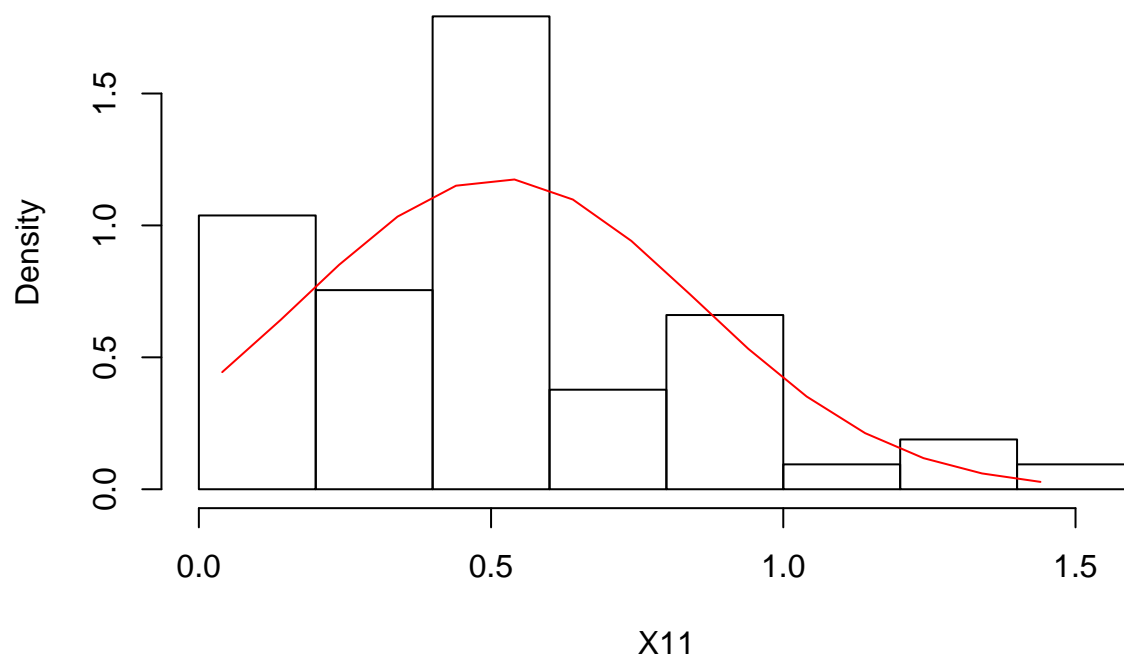
```
qqnorm(X11, main = "Normal Q-Q Plot de X11 (estimación de la concentración de mercurio)", col="darkmagenta",  
qqline(X11))
```

Normal Q-Q Plot de X11 (estimación de la concentración de mercuri



```
hist(X11,prob=TRUE,col=0, main = "Histograma de X11 (estimación de la concentración de mercurio)")  
x=seq(min(X11),max(X11),0.1)  
y=dnorm(x,mean(X11),sd(X11))  
lines(x,y,col="red")
```


Histograma de X11 (estimación de la concentración de mercurio)



```
skewness(X11)
```

```
## [1] 0.9723853
```

```
kurtosis(X11)
```

```
## [1] 3.712108
```

La gráfica de Q-Q Plot muestra que tiene una distribución con colas gruesas, es decir, tiene una baja curtosis y una distribución platicúrtica. Por otro lado, en el histograma se muestra una distribución asimétrica con sesgo a la derecha.

Variables categóricas

```
db_mercurio_cnt = dbNum
db_mercurio_cnt$X13 <- with(db_mercurio_cnt, ifelse(X7 > 0.5, 1, 0))
db_mercurio_cnt_table = table(db_mercurio_cnt$X13)
print("Tabla de Distribución de Lagos que Superaron los 0.5 mg de Hg/Kg: ")
```

Distribución de los datos (diagramas de barras, diagramas de pastel)

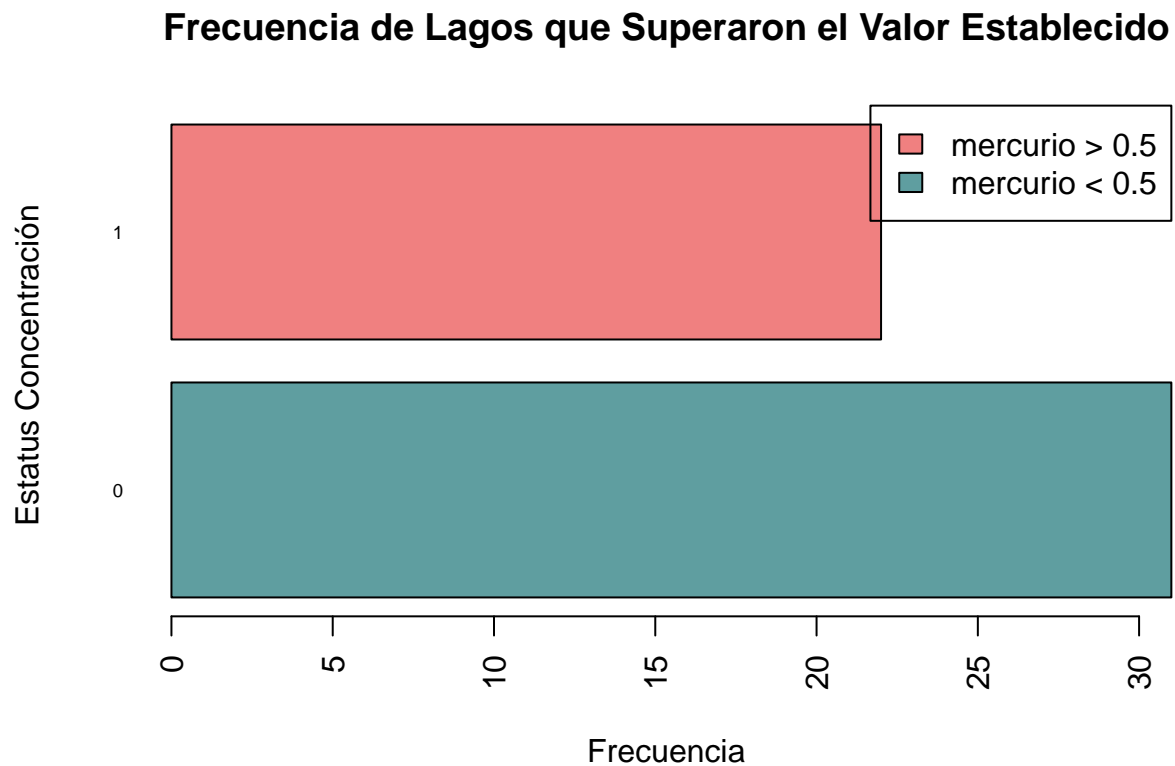
```
## [1] "Tabla de Distribución de Lagos que Superaron los 0.5 mg de Hg/Kg: "
```

```
db_mercurio_cnt_table
```

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

```
# Gráfica de Frecuencia
```

```
barplot(db_mercurio_cnt_table, width = 1, cex.names = 0.6, col = c("cadetblue", "lightcoral"), main = "Frecuencia de Lagos que Superaron el Valor Establecido")
```



La gráfica anterior muestra la frecuencia de lagos que superaron el valor establecido de 0.5 mg de Hg/kg. Por lo tanto, la barra rosa muestra que 22 lagos no son adecuados para la pesca ya que supera la cantidad de concentración de mercurio. Por otro lado, 31 lagos tienen una concentración de mercurio menor a 0.5.

```
db_ph_cnt = dbNum  
db_ph_cnt$X4 <- with(db_ph_cnt, ifelse(X4 < 7.0, "Ácido", ifelse(X4 == 7.0, "Neutro", "Alcalino")))  
db_ph_cnt_table = table(db_ph_cnt$X4)  
print("Tabla de Distribución del PH: ")
```

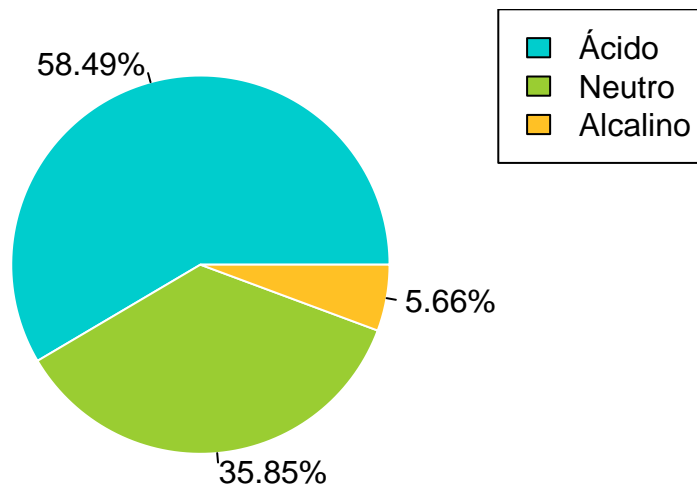
```
## [1] "Tabla de Distribución del PH: "
```

```
db_ph_cnt_table
```

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

```
#Gráfica de pie
colors <- c("cyan3", "yellowgreen", "goldenrod1")
pie(db_ph_cnt_table, border="white", col = colors, main = "Gráfica del PH", labels = paste0(round(100 *
legend("topright", c("Ácido", "Neutro", "Alcalino"), fill=colors)
```

Gráfica del PH



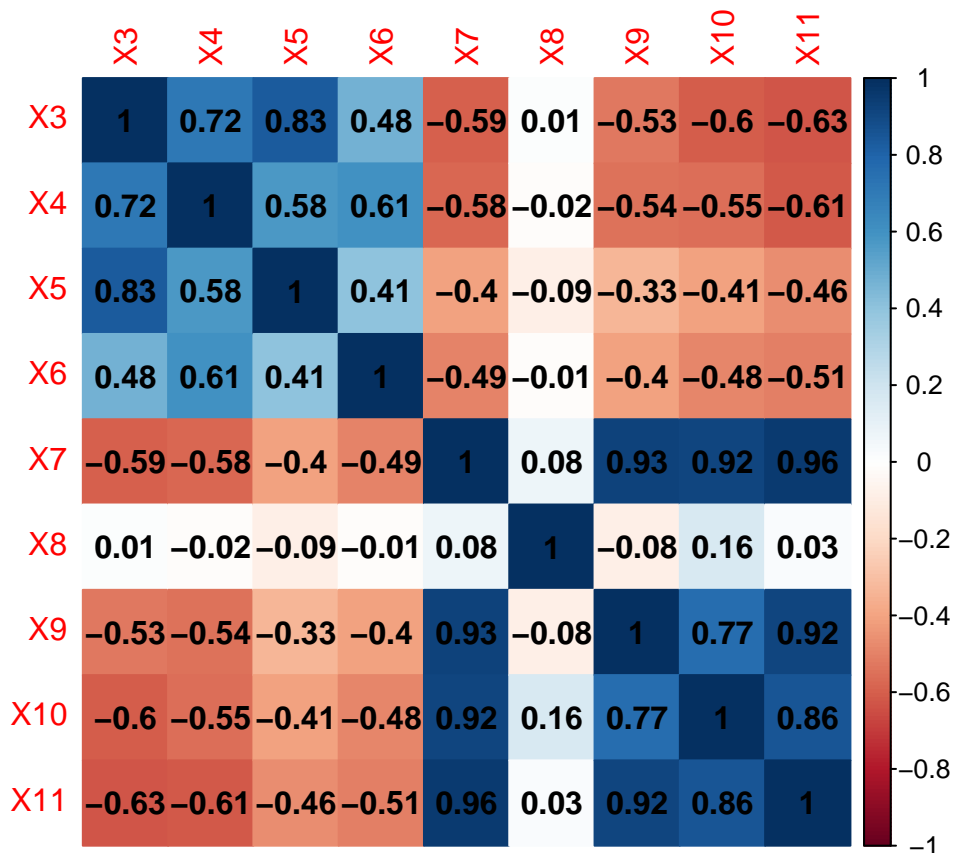
En el diagrama anterior se muestra el nivel de PH de los lagos divididos por ácido (menor a 7), neutro (igual a 7) y alcalino (mayor a 7).

3. Explora la correlación entre las variables. Identifica cuáles son las correlaciones más fuertes y qué sentido tiene relacionarlas.

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(cor(dbNum), method="color", addCoef.col = "black")
```



Al analizar la matriz podemos notar que las correlaciones que más destacan son las que tiene la variable X7 con X9 (0.93), X10 (0.92) y X11 (0.96) con una correlación fuerte positiva. Es probable que esto se deba a que los datos de todas estas variables miden la concentración media de mercurio. Asimismo, la matriz muestra que la variable X7 es la que tiene una correlación significativa ya sea positiva o negativa con las otras variables. Por lo tanto, esta será la variable principal para realizar los modelos más adelante.

A pesar de que las variables X9, X10 y X11 estén fuertemente correlacionadas con la X7, considero que para el análisis de datos y creación de modelos se utilizarán otras variables. Debido al supuesto que establece que ninguna variable independiente se encuentra altamente correlacionada con otra variable del modelo. Al elegir variables con alta correlación puede ocurrir un problema de multicolinealidad en modelos de regresión.

Análisis de datos y pregunta base

Las dos herramientas estadísticas para realizar el análisis de datos y proporcionar información para contestar la pregunta base son el modelo ANOVA y la regresión múltiple.

ANOVA

Se utilizará el modelo ANOVA con el propósito de contestar la siguiente pregunta: ¿habrá diferencia significativa entre la concentración de mercurio y la edad de los peces? Por lo tanto, para realizarlo se utilizarán los datos de X12 (indicador de la edad de los peces) y X7 (concentración media de mercurio en el tejido muscular de los peces). Primero, se acomodarán los datos de X7 en otro dataframe según los valores de edad, es decir, primero los que tengan valor 0 (jóvenes) y después los que tengan valor 1 (maduros). Después

se creará la variable edad que contiene los factores categóricos de X12 para poder realizar el ANOVA con un nivel de significación de 0.05.

```
db_mercurio_num = db[3:12]
media_mercurio_j = db_mercurio_num[db_mercurio_num$X12 == 0, ]$X7
media_mercurio_m = db_mercurio_num[db_mercurio_num$X12 == 1, ]$X7

print("jóvenes")
```

```
## [1] "jóvenes"
```

```
media_mercurio_j
```

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

```
len_media_mercurio_j = length(media_mercurio_j)
print(len_media_mercurio_j)
```

```
## [1] 10
```

```
print("maduros")
```

```
## [1] "maduros"
```

```
media_mercurio_m
```

```
## [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
```

```
len_media_mercurio_m = length(media_mercurio_m)
print(len_media_mercurio_m)
```

```
## [1] 43
```

```
media_mercurio = c(media_mercurio_j, media_mercurio_m)
print("media mercurio")
```

```
## [1] "media mercurio"
```

```
media_mercurio
```

```
## [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
```

```
edad = c(rep("J", len_media_mercurio_j), rep("M", len_media_mercurio_m))
edad = factor(edad)

anova = aov(media_mercurio ~ edad)
summary(anova)
```

```
##           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
```

Ya que el valor de p (0.438) es mayor a 0.05 entonces encontramos que no existe un efecto significativo de la variable edad. Además, no aparecen los valores de significancia en el summary. Entonces se puede concluir que la relación de edad y concentración media no es estadísticamente significativa ya que no se cuenta con suficiente información para detectar una diferencia significativa. Por lo tanto, se utilizará el modelo ANOVA con otros datos y el mismo nivel de significación de 0.05.

Ahora, se implementará el modelo con los datos de frecuencia de lagos que superaron el valor establecido (0.5). Esto se realizará con el fin de contestar la pregunta: ¿hay evidencia para suponer que la concentración promedio de mercurio en los lagos es dañino para la salud humana? Considera que las normativas de referencia para evaluar los niveles máximos de Hg (Reglamento 34687-MAG y los reglamentos internacionales CE 1881/2006 y Codex Standard 193-1995) establecen que la concentración promedio de mercurio en productos de la pesca no debe superar los 0.5 mg de Hg/kg.

Por lo tanto, valores que sean mayores a 0.5 no son adecuados para la pesca, de lo contrario son adecuados.

```
#db_mercurio_num = db[3:12]
media_mercurio_menor = dbNum[dbNum$X7 <= 0.5, ]$X7
media_mercurio_mayor = dbNum[dbNum$X7 > 0.5, ]$X7

print("menor")
```

```
## [1] "menor"
```

```
media_mercurio_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
```

```
len_media_mercurio_menor = length(media_mercurio_menor)
print(len_media_mercurio_menor)
```

```
## [1] 31
```

```
print("mayor")
```

```
## [1] "mayor"
```

```
media_mercurio_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
```

```
len_media_mercurio_mayor = length(media_mercurio_mayor)
print(len_media_mercurio_mayor)
```

```
## [1] 22
```

```
media_mercurio = c(media_mercurio_menor, media_mercurio_mayor)
print("media mercurio")
```

```
## [1] "media mercurio"
```

```
media_mercurio
```

```
## [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
```

```
nivel = c(rep("Menor", len_media_mercurio_menor), rep("Mayor", len_media_mercurio_mayor))
nivel = factor(nivel)
nivel
```

```
## [1] Menor Menor Menor Menor Menor Menor Menor Menor Menor Menor Menor Menor Menor
## [13] Menor Menor Menor Menor Menor Menor Menor Menor Menor Menor Menor Menor Menor
## [25] Menor Menor Menor Menor Menor Menor Menor Menor Mayor Mayor Mayor Mayor Mayor
## [37] Mayor Mayor Mayor Mayor Mayor Mayor Mayor Mayor Mayor Mayor Mayor Mayor Mayor
## [49] Mayor Mayor Mayor Mayor Mayor
## Levels: Mayor Menor
```

```
anova = aov(media_mercurio ~ nivel)
summary(anova)
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## nivel         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
```

Para realizar el ANOVA primero se agregó la variable dependiente numérica (media_mercurio) y luego la variable independiente categórica (nivel). Además, en la tabla anterior podemos notar que el valor de $p = 9.68e-15$ es menor al valor de significancia establecido, por lo tanto, se continuará con el análisis.

```
m = tapply(media_mercurio, nivel, mean)
s = tapply(media_mercurio, nivel, sd)
n = tapply(media_mercurio, nivel, length)
print("Medias del nivel")
```

```
## [1] "Medias del nivel"
```

```
m
```

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

```
print("Desviación estándar del nivel")
```

```
## [1] "Desviación estándar del nivel"
```

```
s
```

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

```
print("Tamaño de la muestra por nivel")
```

```
## [1] "Tamaño de la muestra por nivel"
```

```
n
```

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

Intervalos de confianza

```
sm = s/sqrt(n)
E=abs(qt(0.025,n-1))*sm
In=m-E
Sup=m+E
In
```

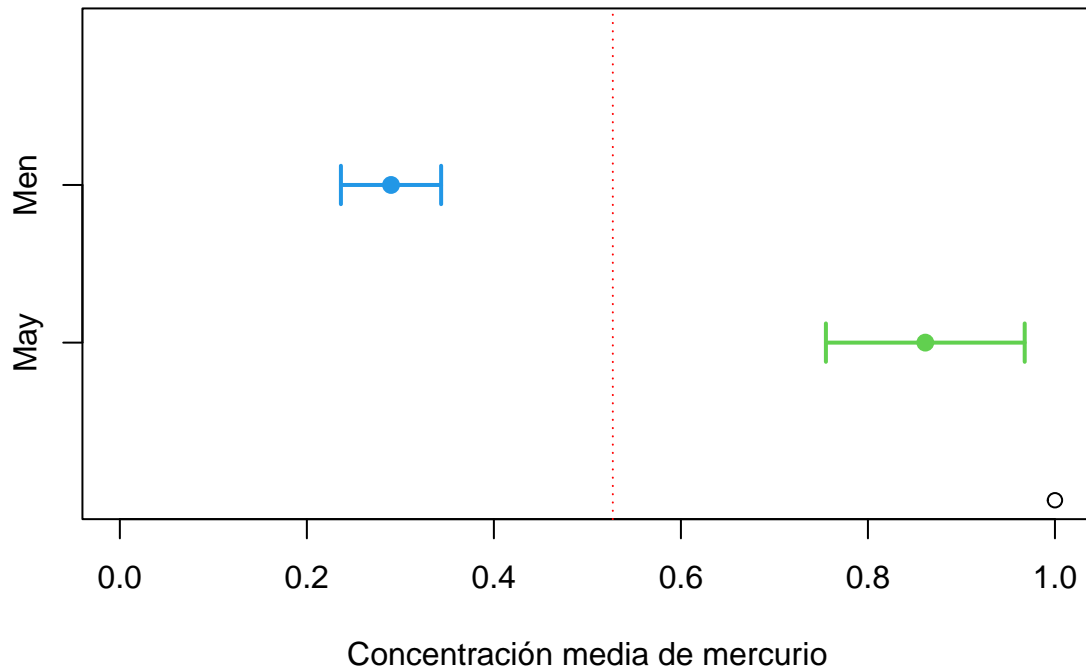
```
##      Mayor      Menor
## 0.7550654 0.2364250
```

```
Sup
```

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

```
plot(0,ylim=c(0,3),xlim=c(0,1), yaxt="n", ylab="",xlab="Concentración media de mercurio",main="Concentr
axis(2,at=c(1:2),labels=c("May","Men"))
for(i in 1:2){
arrows(In[i],i,Sup[i],i, angle=90, code=3, length = 0.1, lwd = 2,col=i+2)
points(m[i], i, pch=19, cex=1.1,col=i+2)}
abline(v=mean(media_mercurio),lty=3,col="red")
```

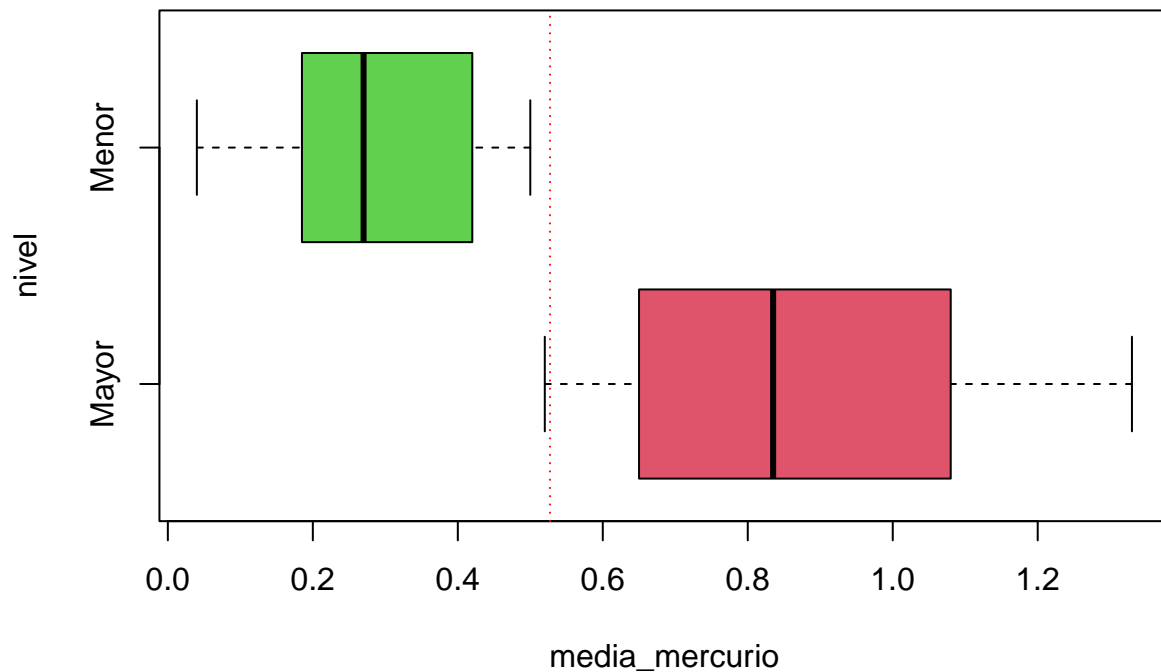

Concentración media de mercurio por nivel



La gráfica anterior muestra la relación entre los niveles que son mayores o menores con la concentración media de mercurio. Se puede observar que la concentración de datos de la clasificación “menor” se encuentra entre 0.2 y 0.4, mientras que el grupo “mayor” se encuentra entre 0.7 y 1.

BoxPlot

```
boxplot(media_mercurio ~ nivel, col = 2:5, horizontal = TRUE)
abline(v = mean(media_mercurio), lty = 3, col = "red")
```



El boxplot muestra la distribución de datos en ambos grupos (menor y mayor), este indica que para el grupo “menor” la concentración de datos esta ligeramente a la derecha mientras que en el grupo “mayor” los datos se concentran a la izquierda.

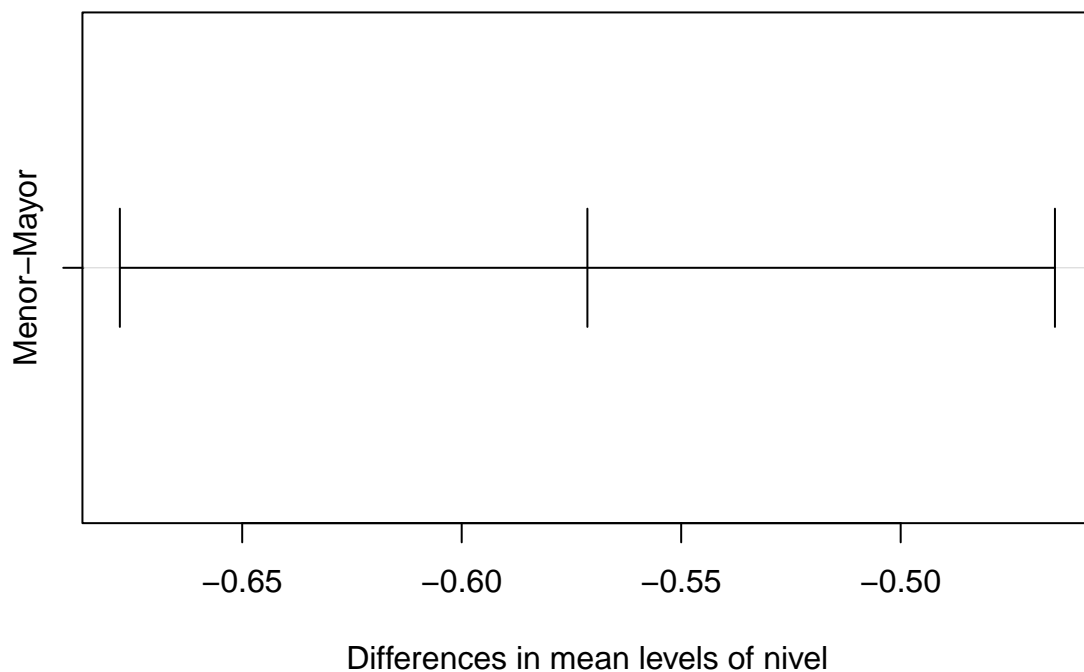
Prueba de Tukey

```
Tu=TukeyHSD(anova)
Tu
```

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

```
plot(TukeyHSD(anova))
```

95% family-wise confidence level



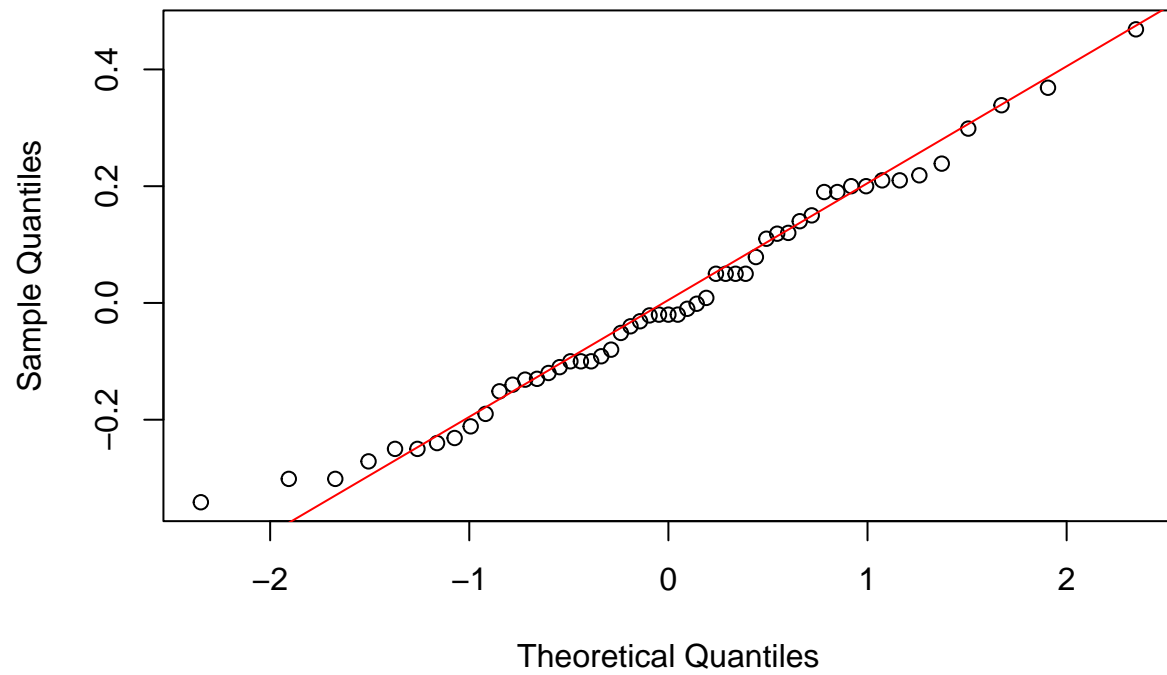
En la prueba de Tukey únicamente se muestra un par de datos, ya que no se tienen otros grupos. Por lo tanto, no se pueden realizar comparación ya que solo se tiene un par de datos. Sin embargo, se puede observar el intervalo de confianza del par menor-mayor.

Verificación de supuestos

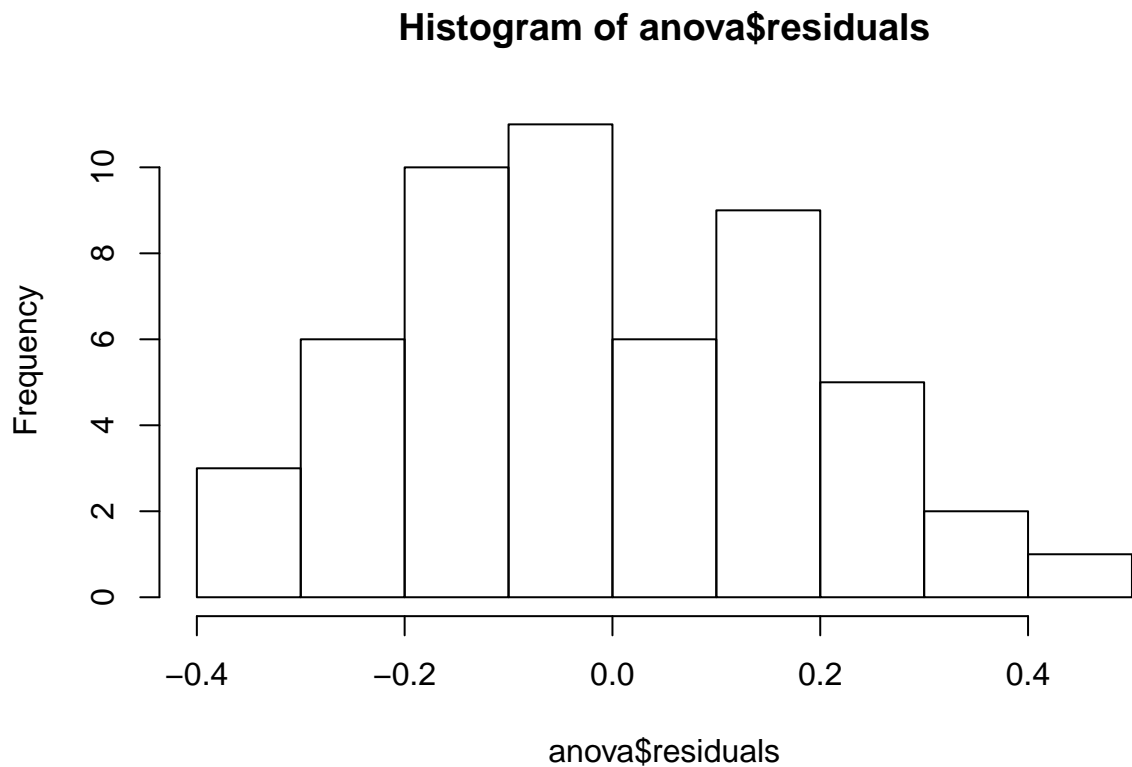
Normalidad

```
qqnorm(anova$residuals)
qqline(anova$residuals,col="red")
```

Normal Q-Q Plot



```
hist(anova$residuals, col=0)
```

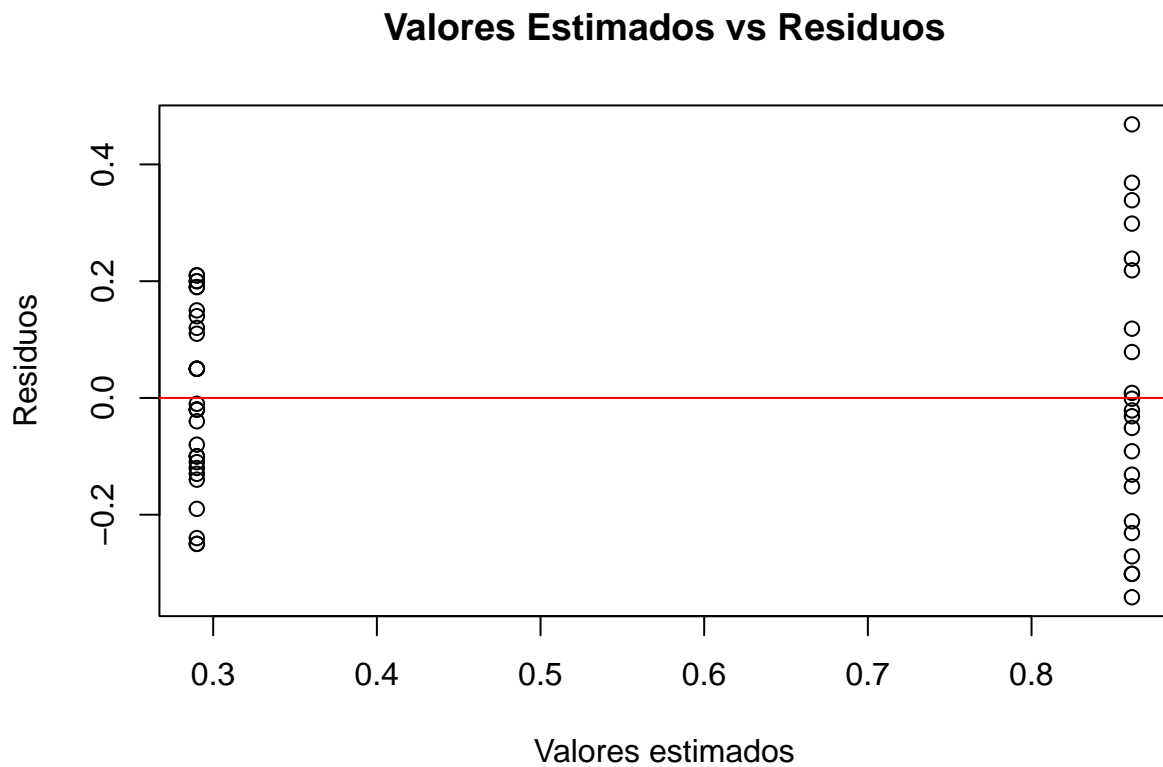


La gráfica del Q-Q plot muestra una distribución normal es ideal. Por otro lado, el histograma muestra una distribución simétrica.

Homocedasticidad

Gráfica Valores estimados vs Residuos

```
plot(anova$fitted.values, anova$residuals, ylab="Residuos", xlab="Valores estimados", main="Valores Estimados vs Residuos")
abline(h=0, col="red")
```



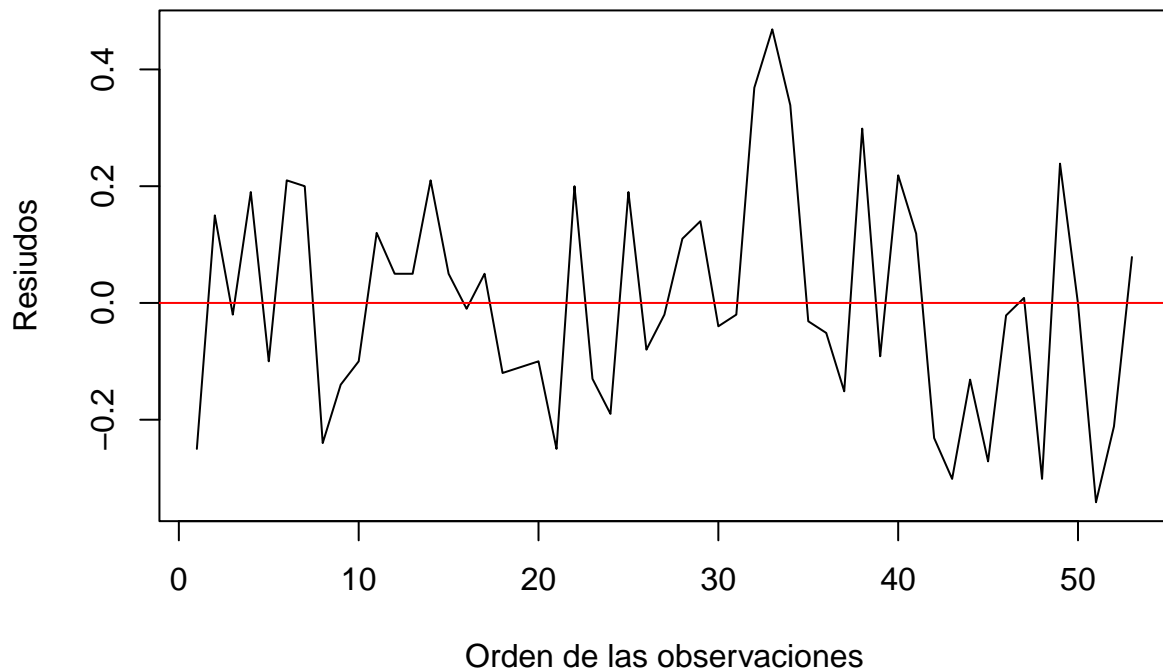
La dispersión de los residuos es constante en toda la gráfica, es decir que cumple con los supuestos.

Independencia

Errores vs Orden de observación

```
plot(c(1:53), anova$residuals, type="l", xlab= "Orden de las observaciones", ylab="Residuos", main="Error vs Orden de observación")
abline(h=0,col="red")
```

Errores vs Orden de Observación



La gráfica tiene una autocorrelación negativa ya que tiene una altercancia muy marcada de residuos positivos y negativos

Regresión Múltiple

Medidas

```
medidas = db[3:7]
medidas
```

##	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
```

Correlación

```
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```



```
## Loading required package: ggplot2
```

```
##
```

```
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
##      format.pval, units
```

```
Rc = rcorr(as.matrix(medidas))
```

```
Rc
```

```
##      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
```

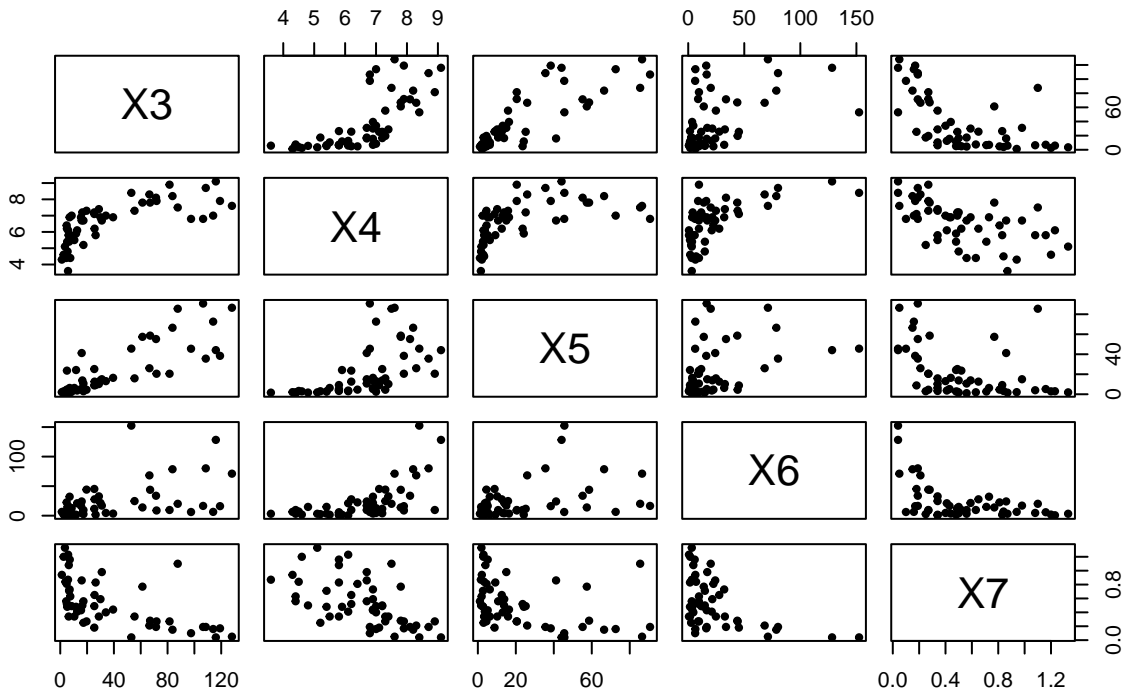
```
## X5 0.0000 0.0000           0.0023 0.0029
```

```
## X6 0.0003 0.0000 0.0023           0.0002
```

```
## X7 0.0000 0.0000 0.0029 0.0002
```

```
pairs(medidas,labels=c("X3","X4","X5","X6","X7"),main="Matriz de dispersión",pch=20)
```

Matriz de dispersión



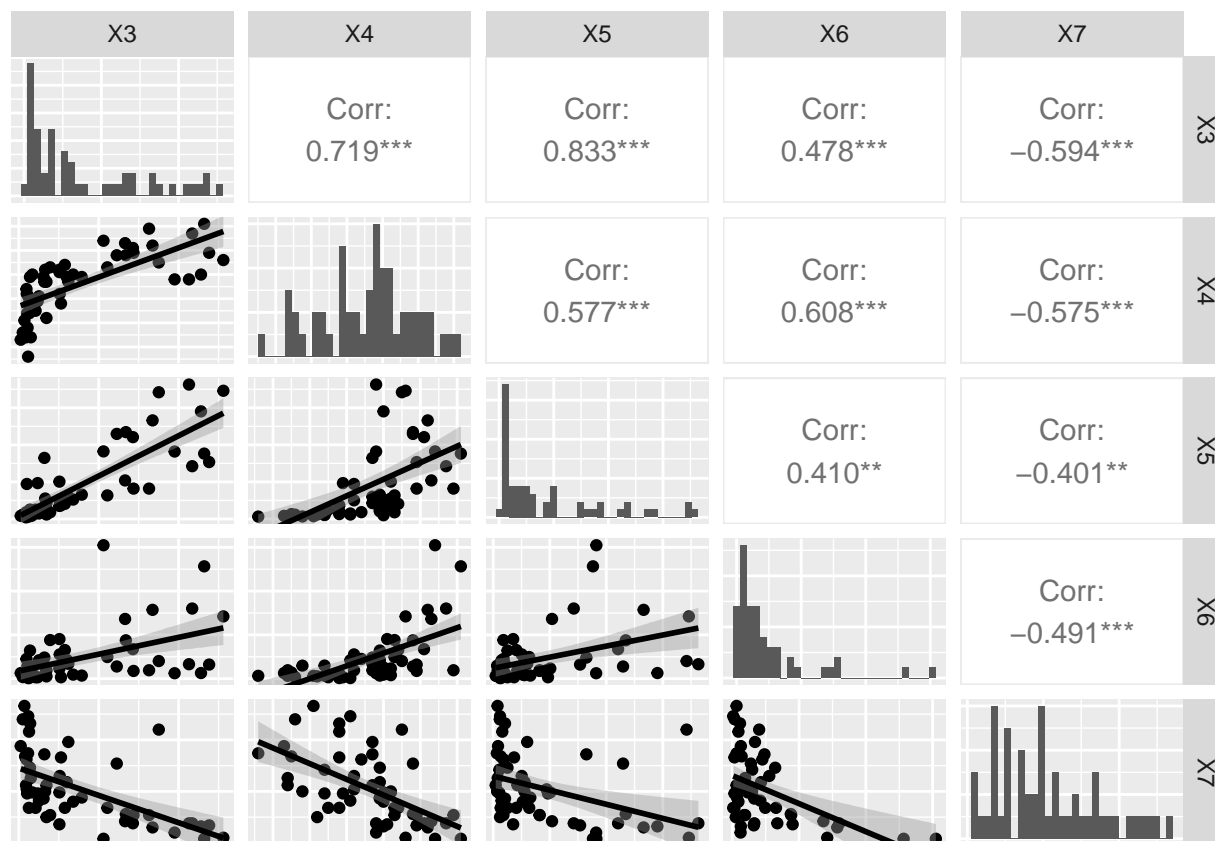
```
library(GGally)
```

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

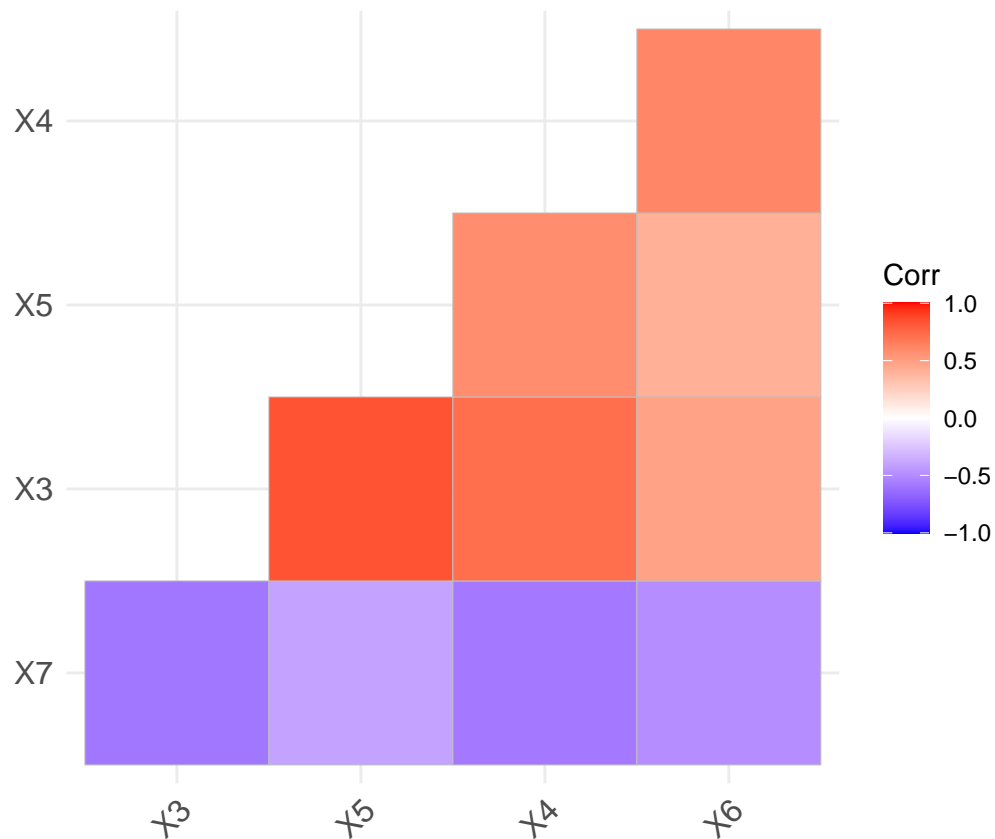
```
ggpairs(medidas, lower = list(continuous = "smooth"),
        diag = list(continuous = "barDiag"), axisLabels = "none")
```

```
## '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'.
```



```
library(ggcorrplot)
library(polycor)
mat_cor <- hetcor(medidas)$correlations #matriz de correlación policorica
ggcorrplot(mat_cor,type="lower",hc.order = T)
```



El modelo

```
R=lm(X7~X3+X4+X5+X6,data=medidas)
summary(R)
```

```
##
## 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
```

Selección del mejor modelo

```
step(R,direction="both",trace=1)
```

```
## 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
```

El mejor modelo

```
R1=lm(X7~X3+X5+X6,data=medidas)
S=summary(R1)
S
```

```
##
## 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
```

Intervalos de confianza

```
confint(R1)
```

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

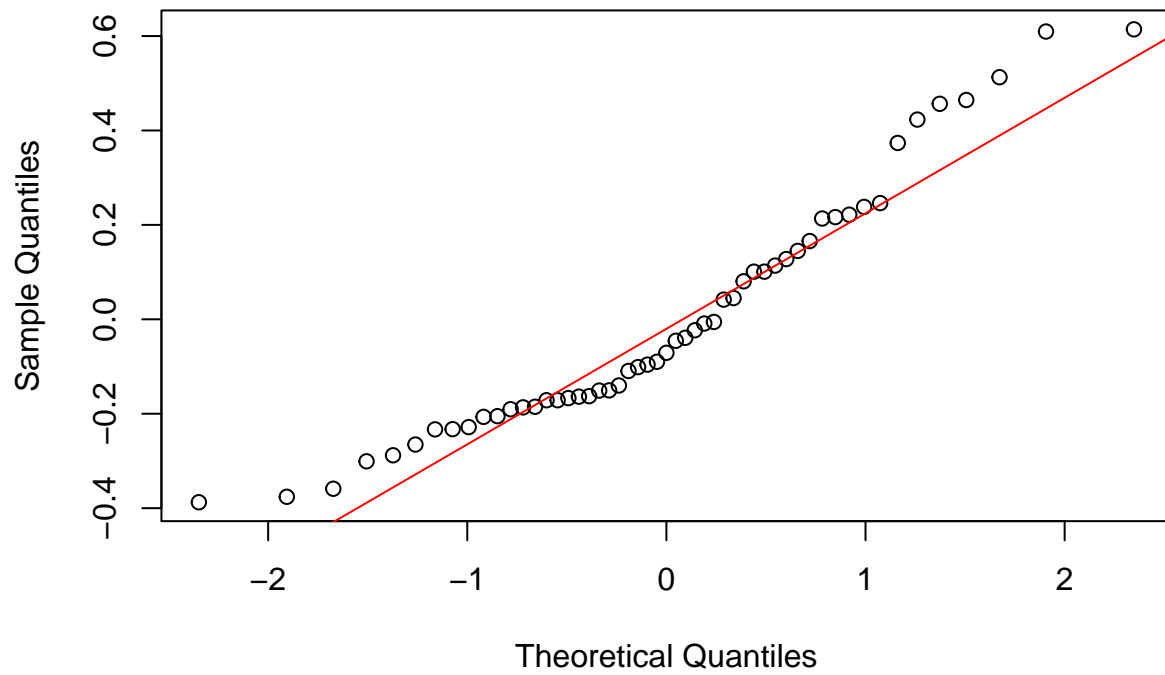
Verificación de supuestos

Normalidad

```
E=R1$residuals
Y=R1$fitted.values

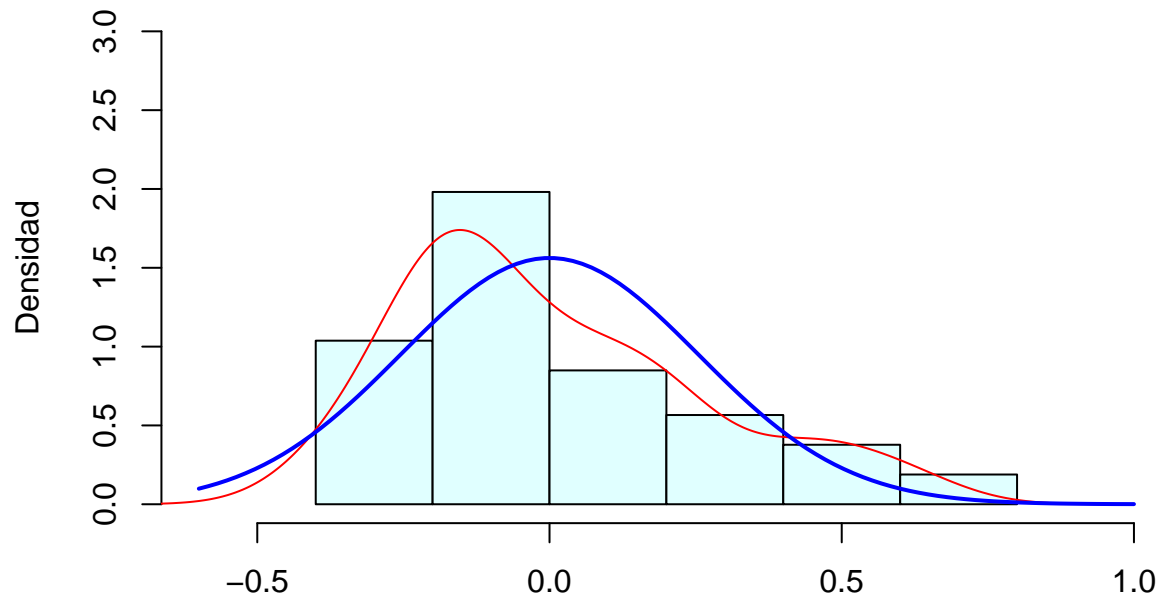
qqnorm(E)
qqline(E,col="red")
```

Normal Q-Q Plot



```
hist(E,col="lightcyan",freq=FALSE,main="Histograma de Residuos",ylim=c(0,3),xlim=c(-0.6,1), xlab="",ylab="")
lines(density(E),col="red")
curve(dnorm(x,mean=mean(E),sd=sd(E)), add=TRUE, col="blue",lwd=2)
```

Histograma de Residuos



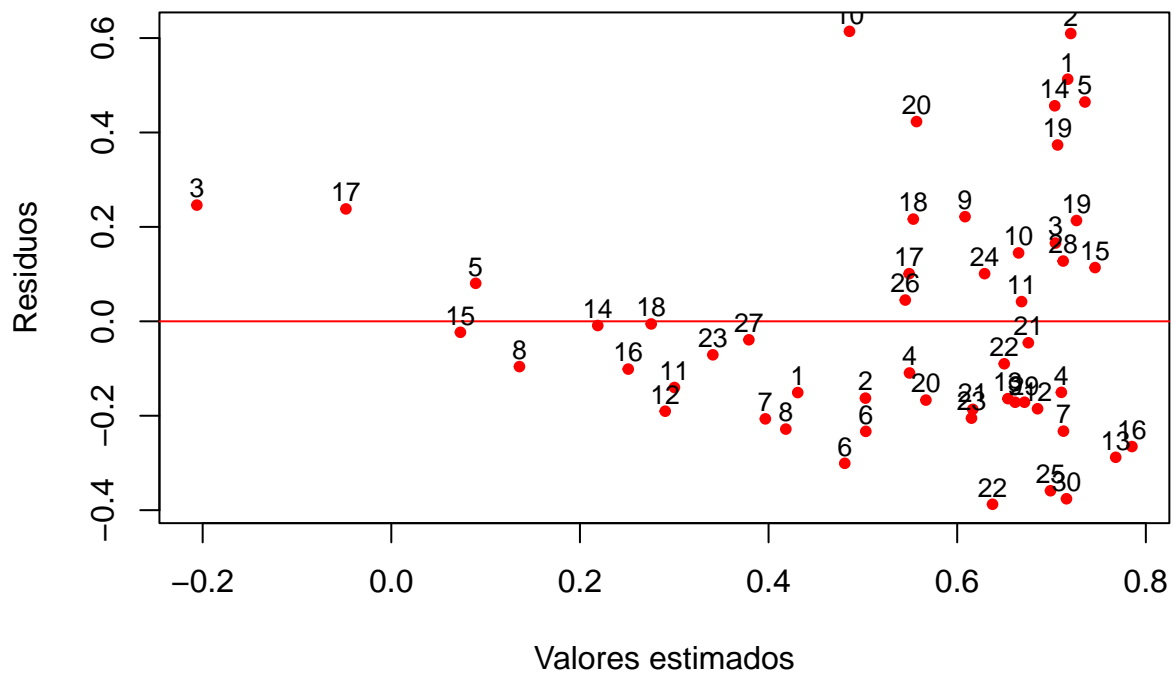
```
shapiro.test(E)
```

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

Homocedasticidad y modelo apropiado

Gráfica Valores estimados vs Residuos

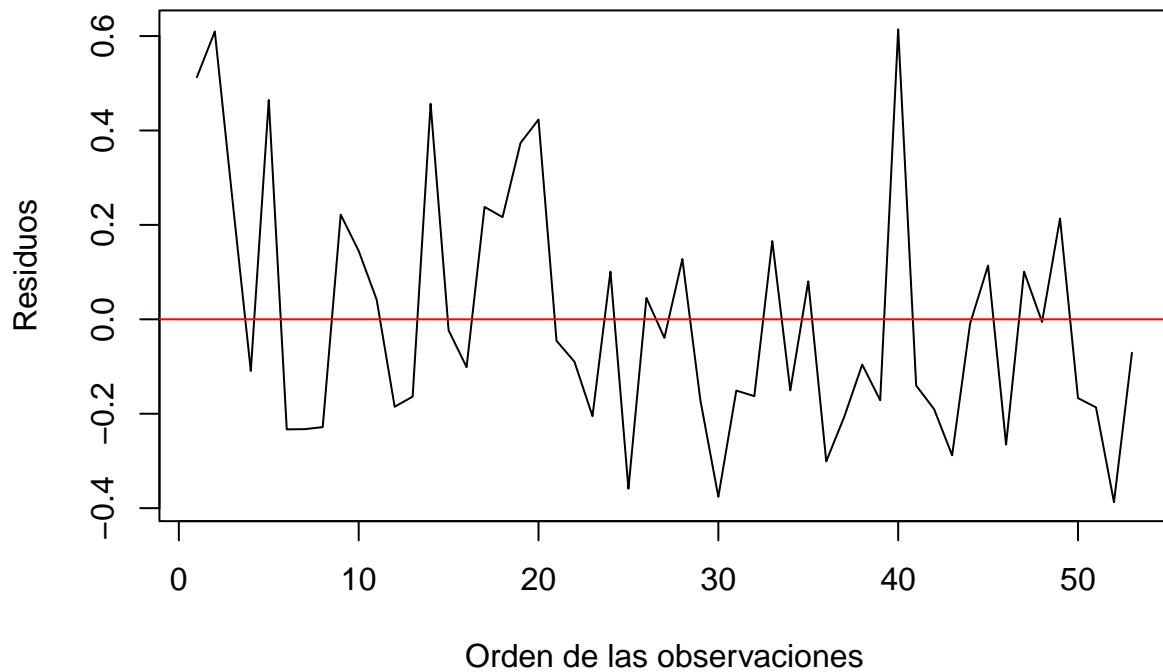
```
plot(Y,E,ylab="Residuos",xlab="Valores estimados",pch=20,col="red")  
abline(h=0,col="red")  
text(Y[,E[,1:30,cex=0.8,pos=3,offset=0.2)
```

Independencia

Errores vs Orden de observación

```
n=length(medidas$X7)
plot(c(1:n),R1$residuals,type="l",xlab="Orden de las observaciones",ylab="Residuos")
abline(h=0,col="red")
```



#Prueba de autocorrelación para verificar independencia: $H_0: \rho=0$

```
library(car)
```

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

```
dwt(R1,alternative="two.sided")
```

```
## lag Autocorrelation D-W Statistic p-value
## 1 0.1660837 1.588784 0.11
## Alternative hypothesis: rho != 0
```

Datos atípicos o influyentes

Datos atípicos

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

```
library(dplyr)
```

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

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

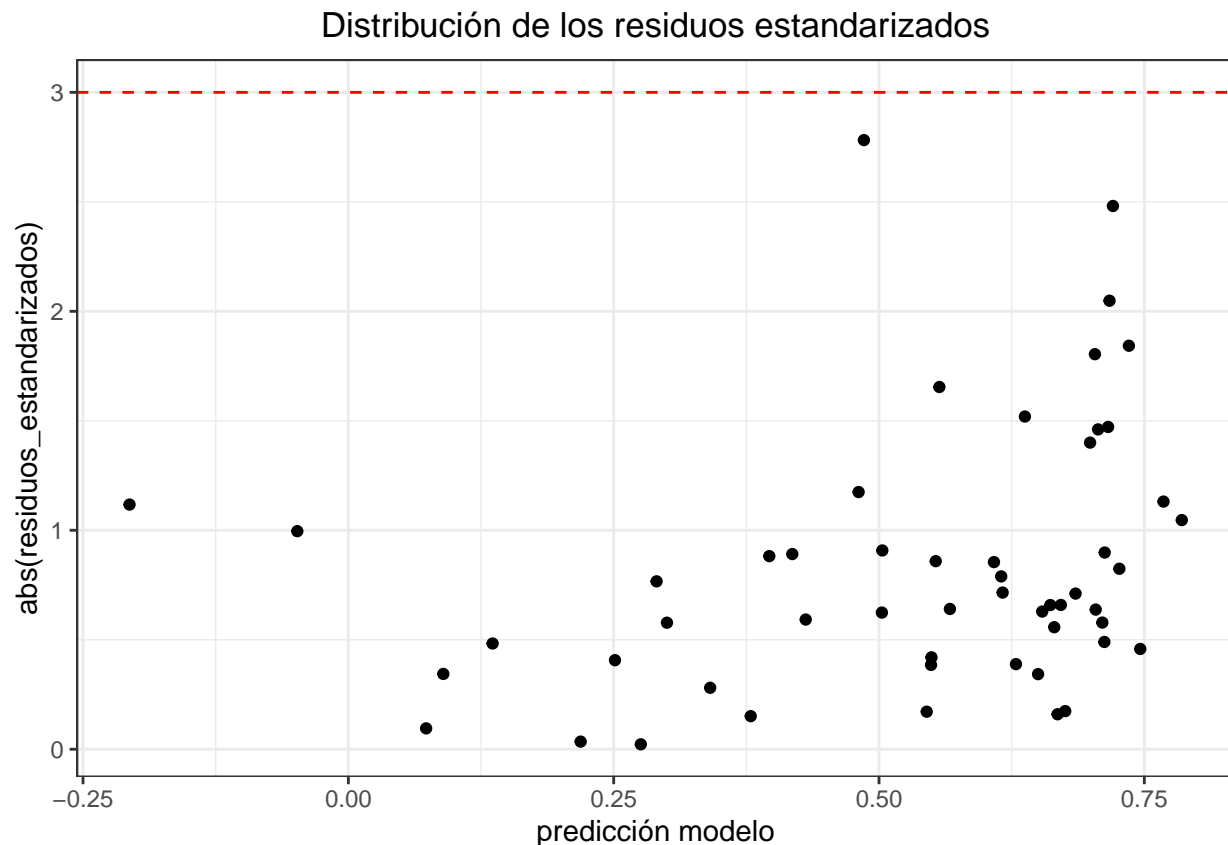
```
## The following objects are masked from 'package:Hmisc':
##
##   src, summarize
```

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

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

```
library(ggplot2)
medidas$residuos_estandarizados <- rstudent(R1) #Introduce una columna en D con los residuos del model

ggplot(data = medidas, aes(x = predict(R1), y = abs(residuos_estandarizados))) +
  geom_hline(yintercept = 3, color = "red", linetype = "dashed") +
  # se identifican en rojo observaciones con residuos estandarizados absolutos > 3
  geom_point(aes(color = ifelse(abs(residuos_estandarizados) > 3, 'red', 'black'))) +
  scale_color_identity() +
  labs(title = "Distribución de los residuos estandarizados", x = "predicción modelo") +
  theme_bw() + theme(plot.title = element_text(hjust = 0.5))
```



```
which(abs(medidas$residuos_estandarizados)>3)
```

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

Datos influyentes

```
summary(influence.measures(R1))
```

```
## 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(R1)
```

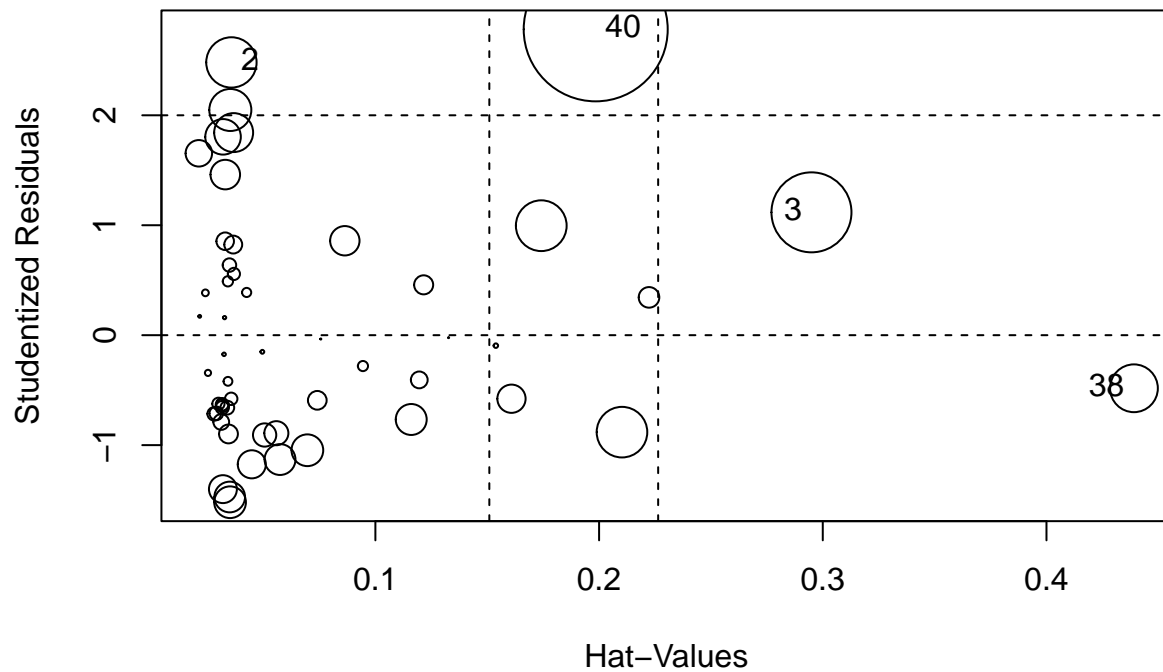
```
## Influence measures of
## lm(formula = X7 ~ X3 + X5 + X6, data = medidas) :
##
##      dfb.1_ dfb.X3 dfb.X5 dfb.X6 dffit cov.r cook.d hat inf
## 1  0.388970 -0.06468 -0.039438 -0.10827 0.39101 0.805 3.59e-02 0.0352
## 2  0.476052 -0.10616 -0.043719 -0.08697 0.47696 0.695 5.15e-02 0.0356 *
## 3 -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
```

Se consideran influyentes aquellas observaciones:

- que tengan leverages mayores a $2.5(p+1)/n = 2.5*3/30 = 2.5/10 = 0.25$
- que tengan distancia de Cook superiores a $4/n$.

```
influencePlot(R1)
```



##	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

Conclusión

- Escribir el Modelo
- Variabilidad explicada por el modelo (coeficiente de determinación)
- Significancia del modelo: Valor p del modelo (F)
- Si se satisfacen todos los supuestos del modelo (indicar ligeros alejamientos)
- Si hay datos atípicos o datos que influyan en el modelo.
- En caso de que se haya hecho el análisis sin datos influyentes, se reportan ambos modelos comparando sus mejoras.