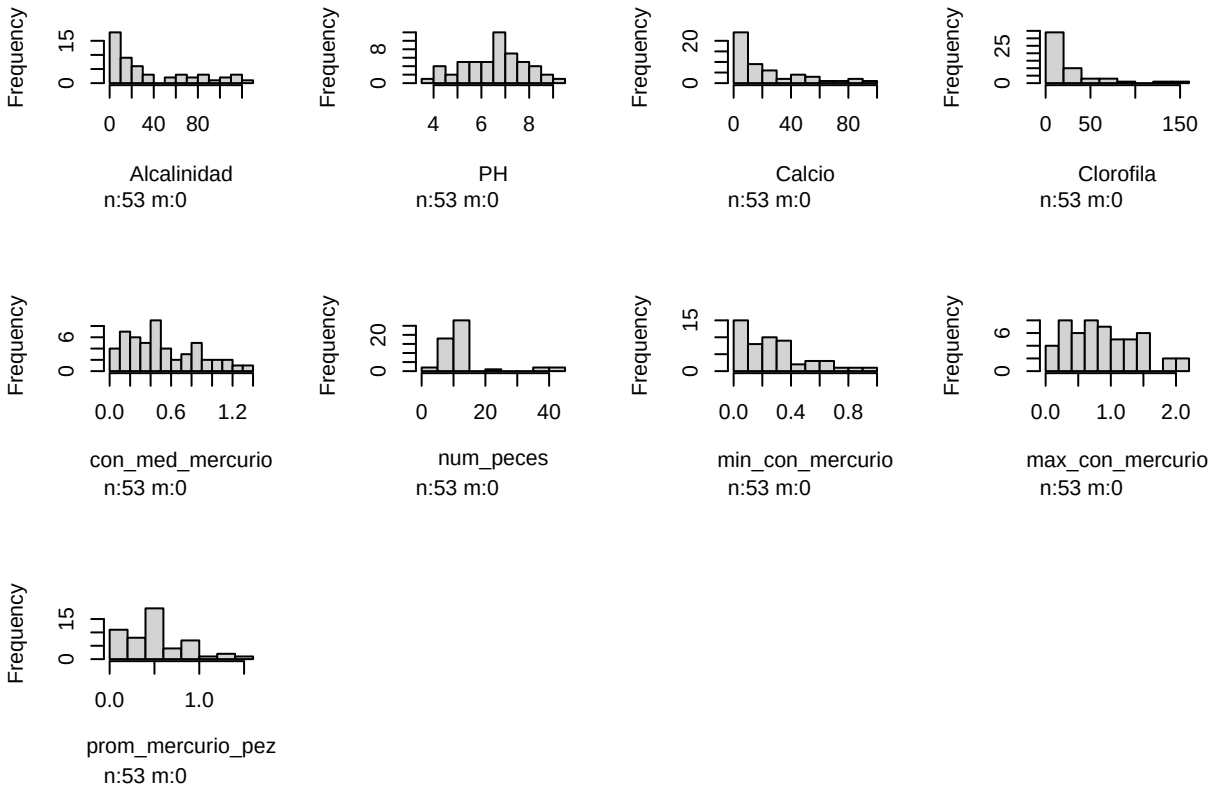
Momento de Retroalimentacion 2 - Modulo 1

Facundo Vecchi A01283666

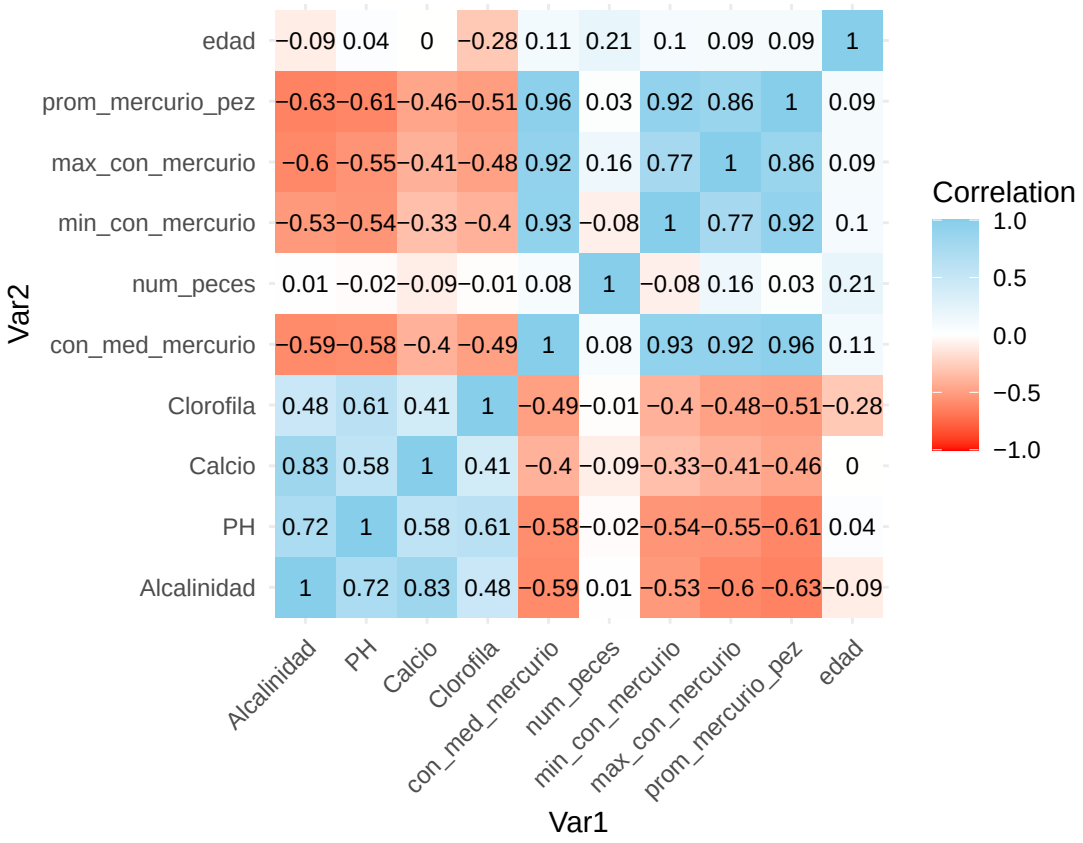
8 de septiembre de 2022

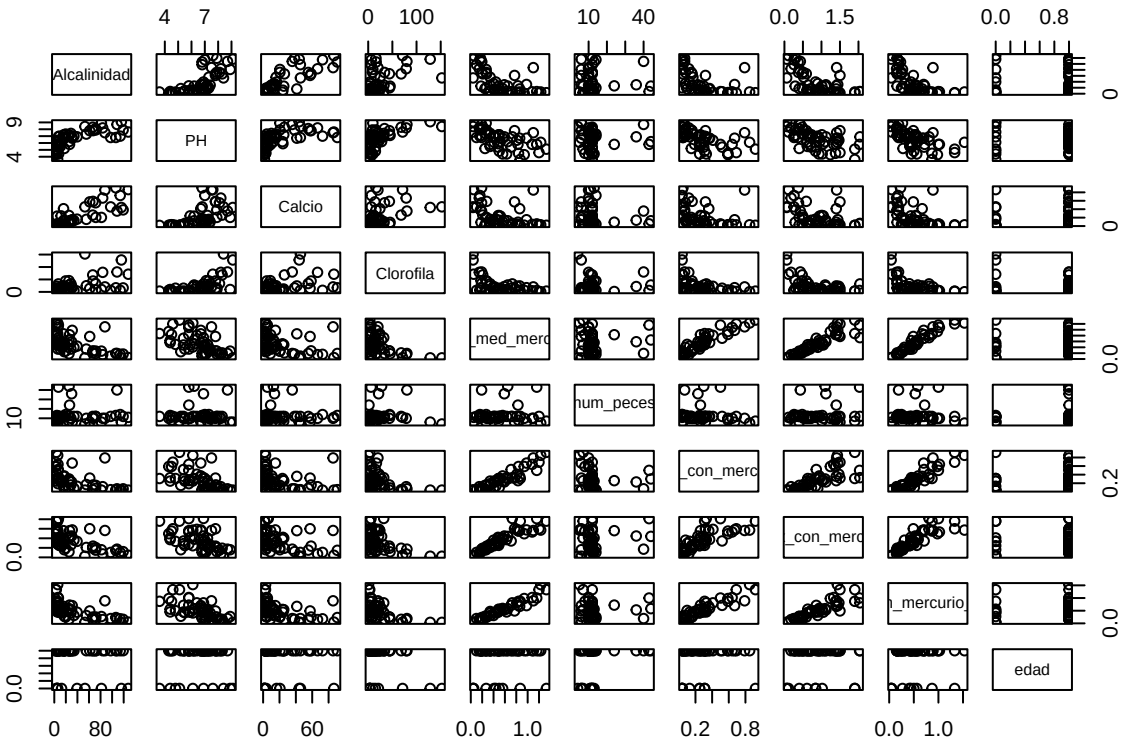
Histogramas de las variables



Podemos observar que varias de las variables tienen un sesgo a la derecha, lo que indica que la mayoria de los datos se encuentran en la parte izquierda de la distribucion. Esto puede deberse a que los datos fueron tomados de una poblacion que no es normal, o que la muestra no es representativa de la poblacion.

Matriz de correlacion





Podemos observar que todas las variables a exepcion de num_peces y edad tienen correlacion moderadas o altas con otras variables. Esto indica que se tendran que eliminar algunas variables para evitar multicolinealidad. Como sabemos que la variable con_med_mercurio es la que queremos predecir, nos quedaremos con las variables que tengan una correlacion alta con esta. Esto nos deja con las variables Alcalinidad, PH, Calcio, Clorofila, min_con_mercurio, max_con_mercurio y prom_mercurio_pez. Analizando las correlaciones entre estas variables, podemos ver que las variables min_con_mercurio, max_con_mercurio y prom_mercurio_pez tienen una correlacion alta entre si, por lo que nos quedaremos con la variable prom_mercurio_pez. Esto nos deja con las variables Alcalinidad, PH, Calcio, Clorofila y prom_mercurio_pez, de las cuales Alcanilidad, PH, Calcio y clorofila tienen una correlacion alta entre si, por lo que nos quedaremos con la variable Alcanilidad. Esto nos deja con las variables Alcalinidad y prom_mercurio_pez. Ya que las variables num_peces y edad tienen una correlacion baja con la variable con_med_mercurio, y no tienen una correlacion alta entre si, tambien las excluiremos.

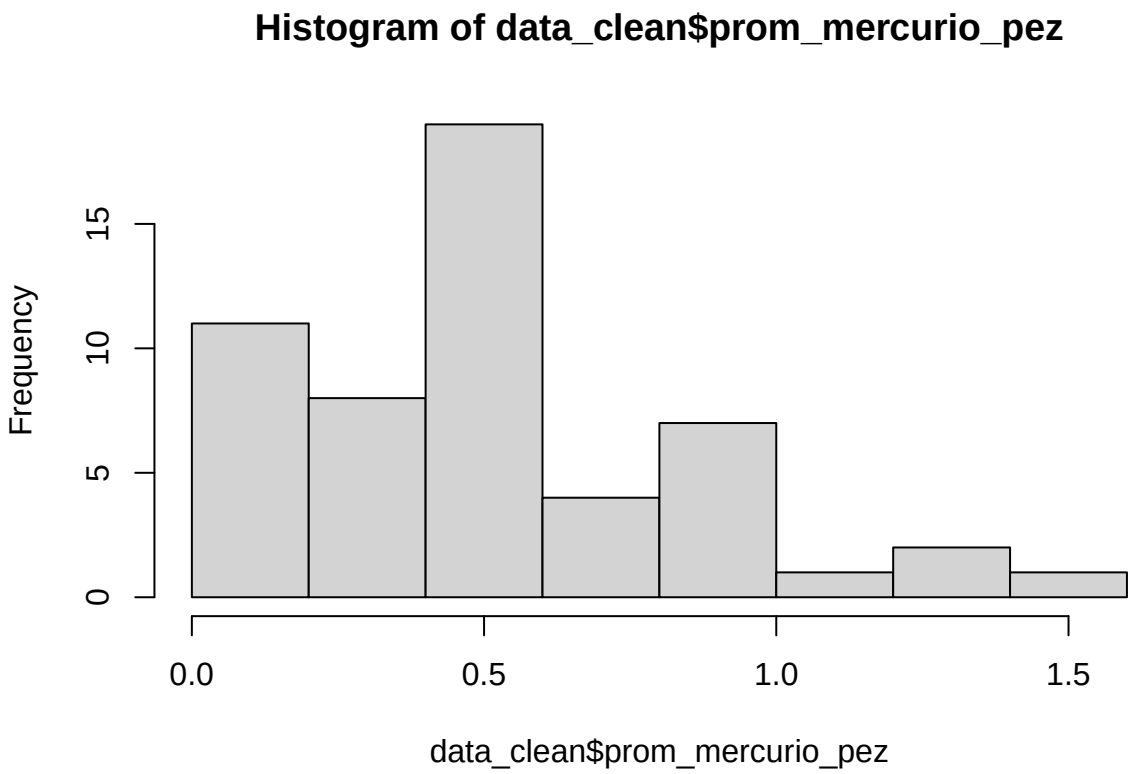
Normalidad de las variables

Chequeo de normalidad

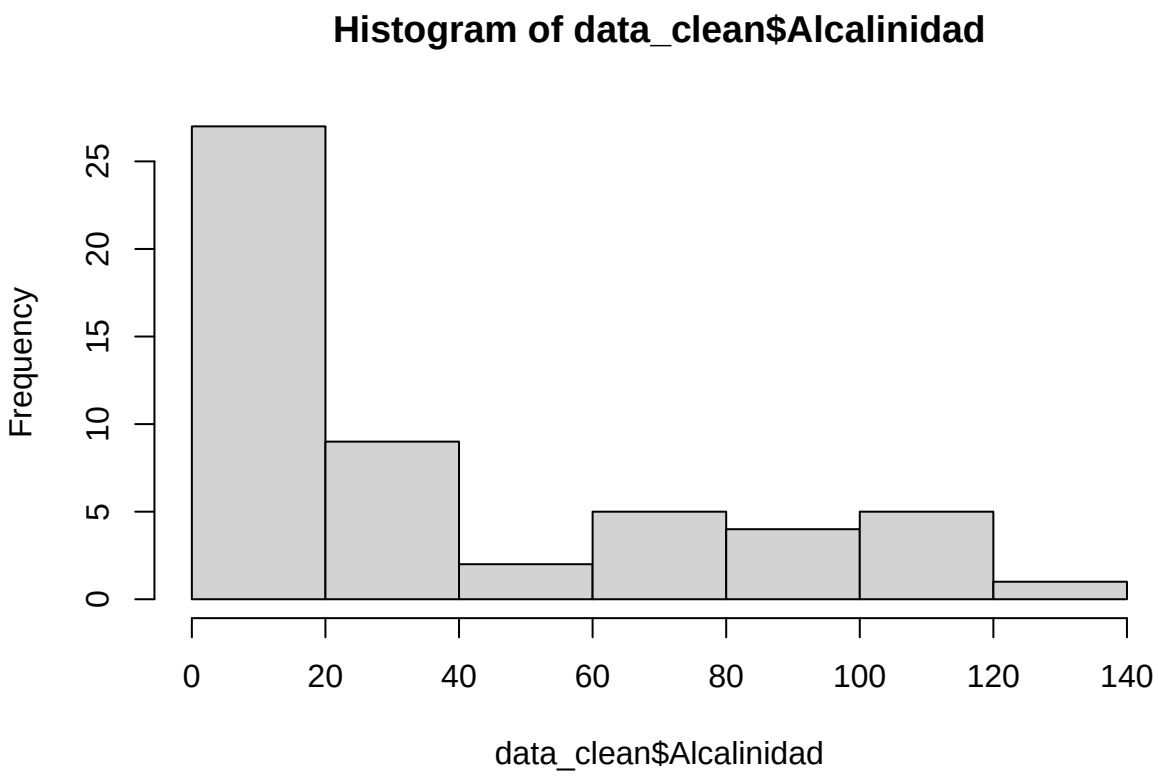
Hipotesis Shapiro-Wilk:
H0: los datos provienen de una distribucion normal
H1: los datos no provienen de una distribucion normal

Reglas de decision:

- alpha = 0.05
- Si p-value < alpha, se rechaza H0 y se acepta H1
- Si p-value > alpha, se rechaza H1 y se acepta H0



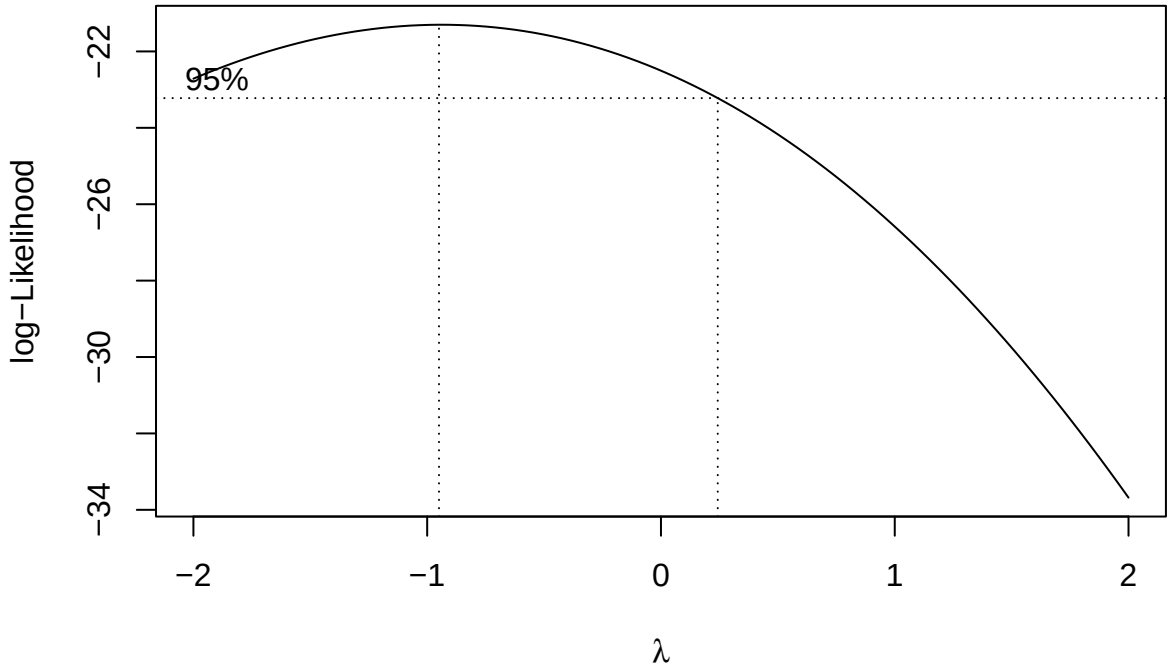

```
##
##  Shapiro-Wilk normality test
##
## data:  data_clean$prom_mercurio_pez
## W = 0.92582, p-value = 0.002782
```




```
##
##  Shapiro-Wilk normality test
##
## data:  data_clean$Alcalinidad
## W = 0.8203, p-value = 1.537e-06
```

Como se puede observar en las graficas y en ambas pruebas del shapiro test, ambas variables no provienen de una distribucion normal. Lo que nos obliga a normalizarlas.

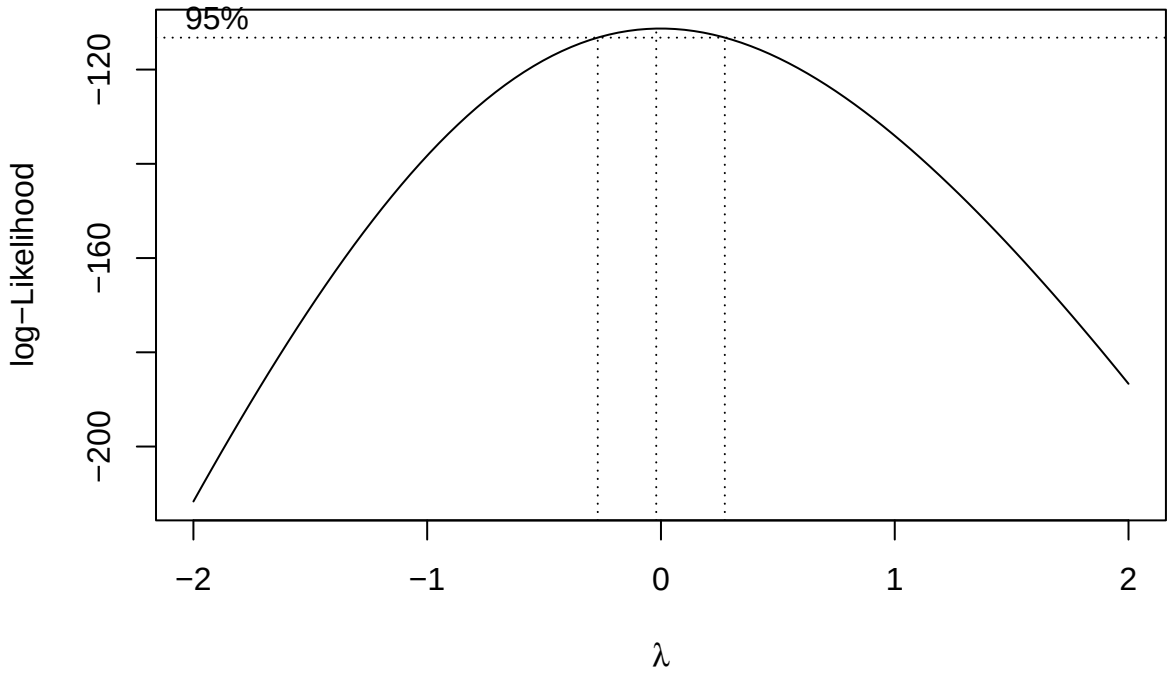
Normalizacion de prom_mercurio_pez con Box-Cox




```
## [1] -0.9494949
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pmp_t
## W = 0.9794, p-value = 0.4879
```

Normalizacion de Alcanilidad con Box-Cox




```
## [1] -0.02020202
```

```
##
##  Shapiro-Wilk normality test
##
## data:  a_t
## W = 0.94374, p-value = 0.01461
```

Como se puede observar por las pruebas de shapiro, solo la variable Alcalinidad sigue sin ser normal. Debido a esto, optaremos por no usarla en el modelo. Lo que nos deja con la variable independiente de prom_mercurio_pez.

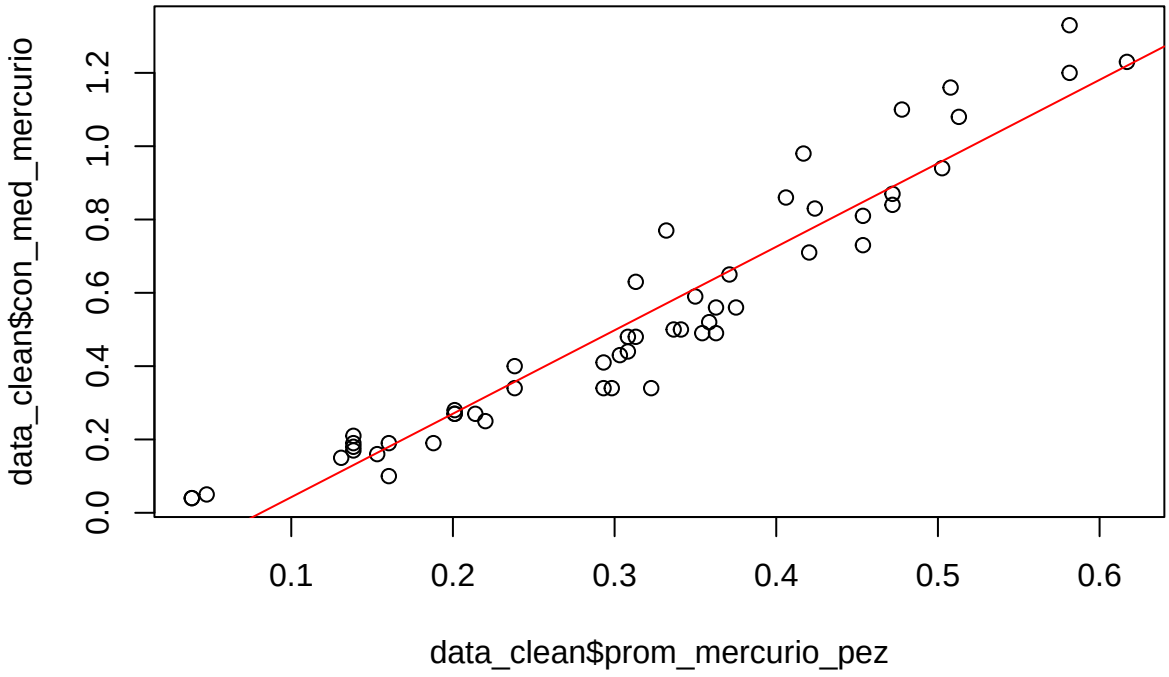
Regresion lineal

```
##
## Call:
## lm(formula = con_med_mercurio ~ prom_mercurio_pez, data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20952 -0.07537 -0.01720  0.05992  0.21605
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.18511    0.03433  -5.393  1.8e-06 ***
## prom_mercurio_pez  2.27680    0.09996  22.776  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.103 on 51 degrees of freedom
## Multiple R-squared:  0.9105, Adjusted R-squared:  0.9087
## F-statistic: 518.8 on 1 and 51 DF,  p-value: < 2.2e-16
```

Ecuacion de la regresion lineal

```
## con_med_mercurio =  -0.1851  + 2.2768 * prom_mercurio_pez
```



Validacion del modelo

Pruebas de hipotesis

Aunque el modelo ya nos indica que la variable prom_mercurio_pez es significativa, realizaremos las pruebas de hipotesis para asegurarnos de que el modelo es correcto.

Hipotesis:
h0: beta1 = 0
h1: beta1 != 0
Reglas de decision:
- alpha = 0.05
- Si p-value < alpha, se rechaza H0 y se acepta H1
- Si p-value > alpha, se rechaza H1 y se acepta H0
- Si t* > t, se rechaza H0 y se acepta H1
- Si t* < t, se rechaza H1 y se acepta H0

```
## La variable prom_mercurio_pez es significativa. (t* > t0 & p < alpha)
##  t* = 22.7765 , t0 = 2.0076
##  p-value = 2.179351e-28 , alpha = 0.05
```

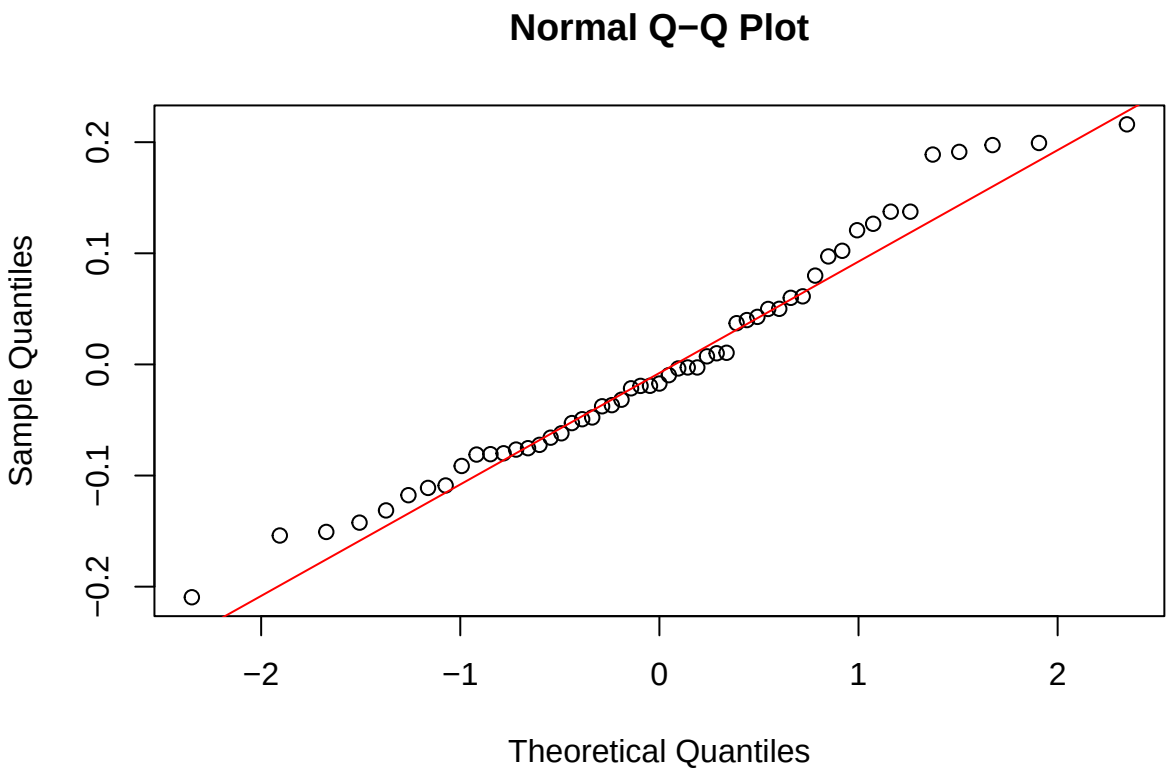
En este caso al solo tener una variable independiente, solo existe la hipotesis para B1. Como podemos observar, la variable prom_mercurio_pez es significativa, ya que el p-value es menor que alpha y la t* es mayor que t0. Confirmando asi que la variable prom_mercurio_pez es significativa para explicar la variable con_med_mercurio.

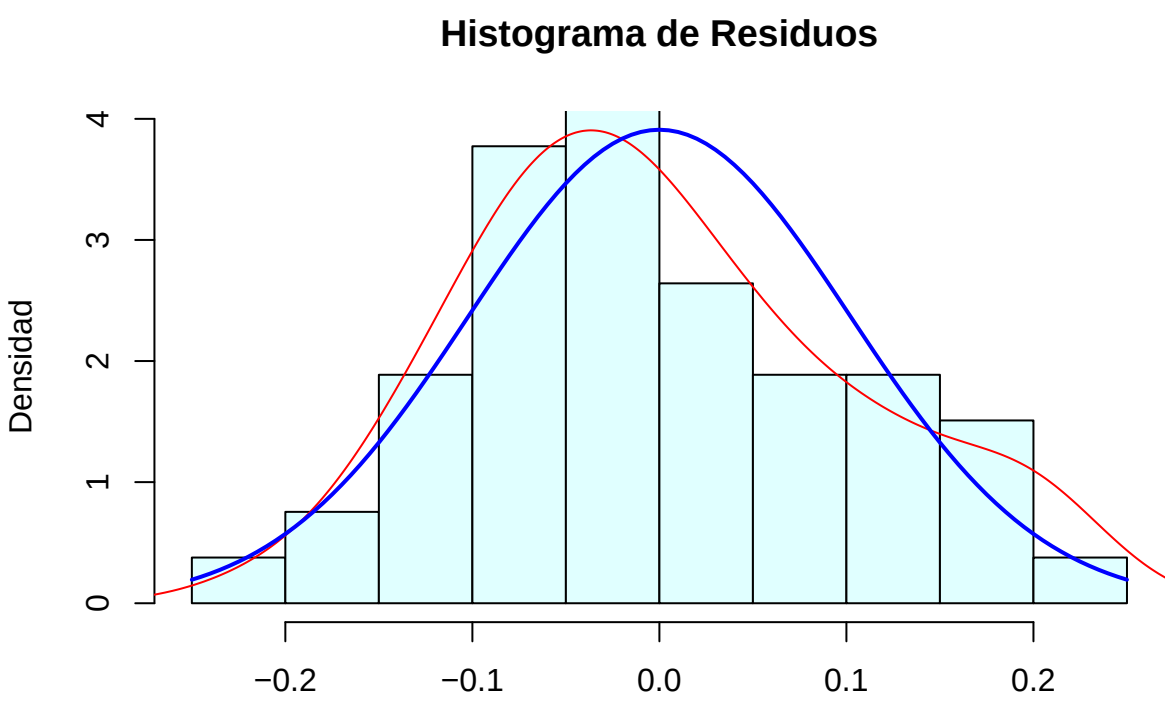
Verificación de supuestos

Normalidad de los residuos

Hipotesis:
H0: miu = 0
H1: miu != 0
Reglas de decision:
- alpha = 0.05
- Si p-value < alpha, se rechaza H0 y se acepta H1
- Si p-value > alpha, se rechaza H1 y se acepta H0

Hipotesis Shapiro-Wilk:
H0: los datos provienen de una distribucion normal
H1: los datos no provienen de una distribucion normal
Reglas de decision:
- alpha = 0.05
- Si p-value < alpha, se rechaza H0 y se acepta H1
- Si p-value > alpha, se rechaza H1 y se acepta H0


Normal Q−Q Plot
Sample Quantiles
−0.2
−0.1
0.0
0.1
0.2
−2
−1
0
1
2
Theoretical Quantiles




```
##
##  Shapiro-Wilk normality test
##
## data:  E
## W = 0.96932, p-value = 0.1886
```

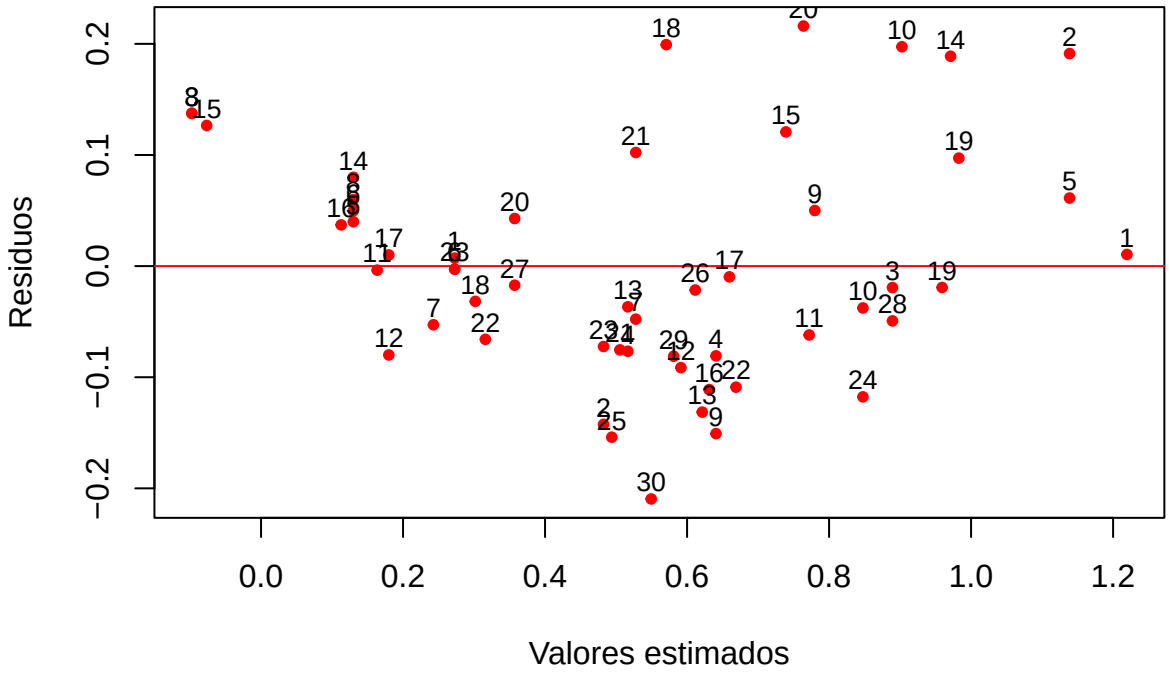
```
##
##  One Sample t-test
##
## data:  E
## t = -1.2136e-16, df = 52, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.02812347  0.02812347
## sample estimates:
##     mean of x
## -1.700822e-18
```

Podemos observar en el qqplot y en el histograma que los residuos siguen una distribución muy cerca a la normal. Al observar los resultados de la prueba de shapiro, se acepta la hipotesis nula y podemos concluir que los residuos siguen una distribución normal. De igual manera podemos observar que tambien se acepta la hipotesis nula en la prueba t de student concluyendo que la media es 0.

Homocedasticidad y modelo apropiado

H0: existe homocedasticidad
H1: existe heterocedasticidad

Reglas de decision:
- alpha = 0.05
- Si p-value < alpha, se rechaza H0 y se acepta H1
- Si p-value > alpha, se rechaza H1 y se acepta H0

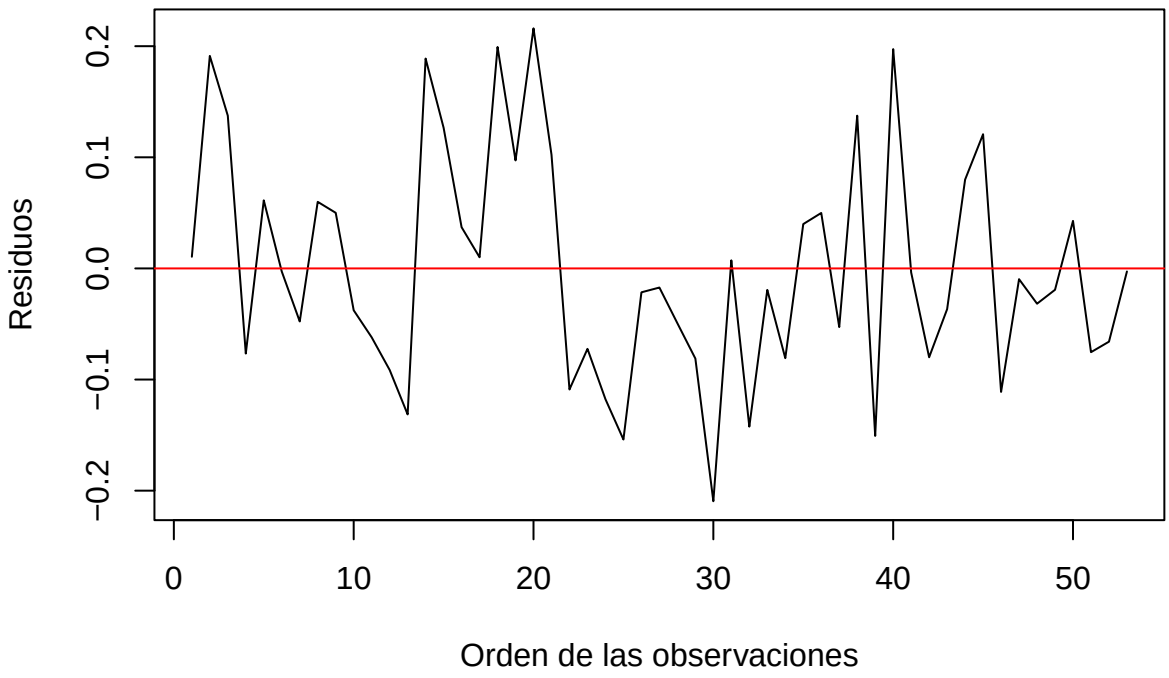



```
##
##  studentized Breusch-Pagan test
##
## data:  rl_best
## BP = 2.5118, df = 1, p-value = 0.113
```

En la grafica podemos observar que los residuos no aparentan seguir algun tipo de patron evidente, al realizar la prueba de Breusch-Pagan podemos observar que el p-value es mayor que alpha, por lo que se acepta la hipotesis nula y podemos concluir que existe homocedasticidad.

Independencia

Hipotesis:
H0: rho = 0
H1: rho != 0
Reglas de decision:
- alpha = 0.05
- Si p-value < alpha, se rechaza H0 y se acepta H1
- Si p-value > alpha, se rechaza H1 y se acepta H0




```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1622996      1.675184   0.254
##  Alternative hypothesis: rho != 0
```

Podemos observar que los residuos no siguen un patron, por lo que podemos decir que los residuos son independientes. Tambien podemos observar que el p-value de la prueba de durbin watson es mayor que alpha, por lo que podemos aceptar la hipotesis nula y decir que los residuos son independientes.

Conclusiones

¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida?
Tras realizar el analisis de regresion lineal, podemos decir que el principal factor que influye en el nivel de contaminacion por mercurio en los peces de los lagos de Florida es el promedio de mercurio en los peces de los lagos.
Ademas de esto podemos concluir que tanto el promedio como el maximo de mercurio en los peces de los lagos serian signficativos dependiendo de lo que se quiera analizar.
Esto se debe a que ambas variables tienen una correlacion alta con la variable dependiente y entre si, posiblemente resultando en modelos de regresion lineal similares al utilizarse individualmente.