Inteligencia Artificial para la Ciencia de Datos

# Reporte final de “Los peces y el mercurio”

Yoceline Aralí Mata Ledezma

A01562116

Instituto Tecnológico  y de Estudios Superiores de Monterrey

Módulo 1. Estadística para la ciencia de datos

Profesora: Blanca Ruiz

18 de septiembre de 2022

# Índice



# INTRODUCCIÓN

Se llevó a cabo un estudio en 53 lagos de Florida con el fin de examinar los factores que influyen en el nivel de contaminación de mercurio, ya que este tipo de contaminación es una amenaza directa contra la salud. A continuación se realiza un análisis de los datos para conocer los principales factores que influyen en el nivel de contaminación.

# EXPLORACIÓN DE LA BASE DE DATOS

## Explora las variables y familiarízate con su significado.

- Identifica la cantidad de datos y variables presentes.
- Clasifica las variables de acuerdo a su tipo y escala de medición.

**Cantidad de datos:** 53 registros no nulos

**Número de variables:** 12

| Variable | Significado | Tipo de variable y escala de medición | Tipo de dato |
|---|---|---|---|
| nombre_lago | nombre del lago | Categórica nominal | String |
| edad | Indicador de la edad de los peces (0: jóvenes; 1: maduros) | Categórica nominal | Bool |
| alcalinidad | Alcalinidad (mg/l de carbonato de calcio) | Cuantitativa de razón | Float |
| ph | PH | Cuantitativa de razón | Float |
| calcio | Calcio (mg/l) | Cuantitativa de razón | Float |
| clorofila | Clorofila (mg/l) | Cuantitativa de razón | Float |
| conc_mercurio | Concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago. | Cuantitativa de razón | Float |
| num_peces | Número de peces estudiados en el lago | Cuantitativa de razón | Int |
| min_concentracion_peces | Mínimo de la concentración de | Cuantitativa de razón | Float |

| | | | |
|---|---|---|---|
| | mercurio en cada grupo de peces | | |
| max_concentracion_peces | Máximo de la concentración de mercurio en cada grupo de peces | Cuantitativa de razón | Float |
| estimacion_concentrracion | Estmación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible) | Cuantitativa de razón | Float |

# Exploración de la base de datos

## Calcula medidas estadísticas

### Variables cuantitativas

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

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

| index | alcalinidad | ph | calcio | clorofila | conc_mercurio | num_peces | min_conc_peces | max_conc_peces | estim_conc |
|---|---|---|---|---|---|---|---|---|---|
| count | 53.00 | 53.00 | 53.00 | 53.00 | 53.00 | 53.00 | 53.00 | 53.00 | 53.00 |
| **mean** | 37.53 | 6.59 | 22.20 | 23.12 | 0.53 | 13.06 | 0.28 | 0.87 | 0.51 |
| **std** | 38.20 | 1.29 | 24.93 | 30.82 | 0.34 | 8.56 | 0.23 | 0.52 | 0.34 |
| **min** | 1.20 | 3.60 | 1.10 | 0.70 | 0.04 | 4.00 | 0.04 | 0.06 | 0.04 |
| **25%** | 6.60 | 5.80 | 3.30 | 4.60 | 0.27 | 10.00 | 0.09 | 0.48 | 0.25 |
| **50%** | 19.60 | 6.80 | 12.60 | 12.80 | 0.48 | 12.00 | 0.25 | 0.84 | 0.45 |
| **75%** | 66.50 | 7.40 | 35.60 | 24.70 | 0.77 | 12.00 | 0.33 | 1.33 | 0.70 |
| **max** | 128.00 | 9.10 | 90.70 | 152.40 | 1.33 | 44.00 | 0.92 | 2.04 | 1.53 |

### Varianza

| Variable | Varianza |
|---|---|
| **alcalinidad** | 1459.51 |
| **ph** | 1.66 |
| **calcio** | 621.65 |
| **clorofila** | 949.65 |
| **conc_mercurio** | 0.12 |
| **num_peces** | 73.29 |
| **min_concentracion_peces** | 0.05 |
| **max_concentracion_peces** | 0.27 |
| **estimacion_concentracion** | 0.115 |

### Moda

| index | alcalinidad | ph | calcio | clorofila | media_merc_porc | num_peces | min_merc_porc | max_merc_porc | merc_estimado_porc |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 17.3 | 5.8 | 3 | 1.6 | 0.34 | 12 | 0.04 | 0.06 | 0.16 |
| 1 | 25.4 | 6.9 | 3.3 | 3.2 | | | | 0.26 | |
| 2 | | | 5.2 | 9.6 | | | | 0.4 | |
| 3 | | | 6.3 | | | | | 0.48 | |
| 4 | | | 20.5 | | | | | 0.69 | |
| 5 | | | | | | | | 0.84 | |
| 6 | | | | | | | | 1.4 | |
| 7 | | | | | | | | 1.5 | |
| 8 | | | | | | | | 1.9 | |

### Variables cualitativas

En este punto se obtuvo la tabla de distribución de frecuencia y la moda de cada una de las variables cualitativas y se encontraron las siguientes observaciones:

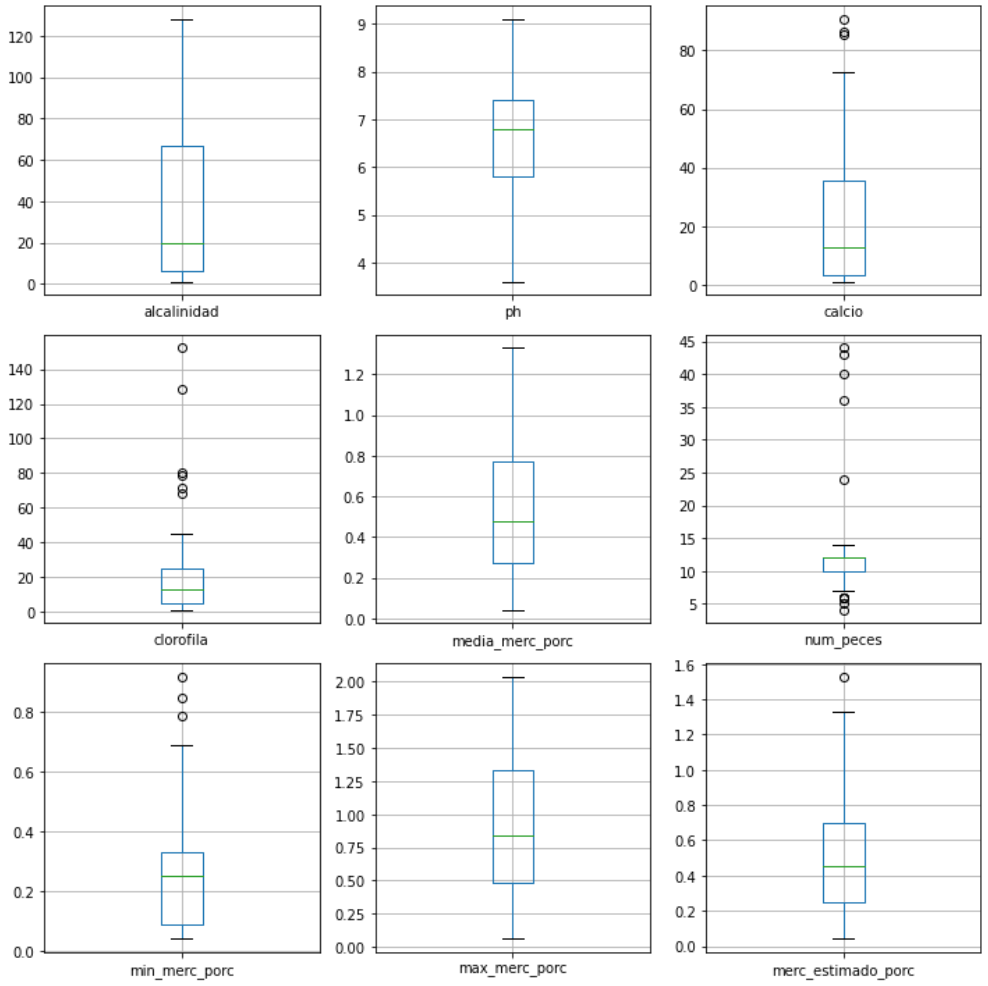
- La mayoría de los peces son maduros.

- Todos los registros son de lagos distintos, a excepción de 1, por lo que podría ser que esta variable no sea tan significativa en la concentración media de mercurio en esta muestra, o en caso de ser significativa, la mejor decisión sería no incluirla en el modelo, pues se tendrían que obtener los datos dummys y eso añadiría 51 variables.

## Explora los datos usando herramientas de visualización

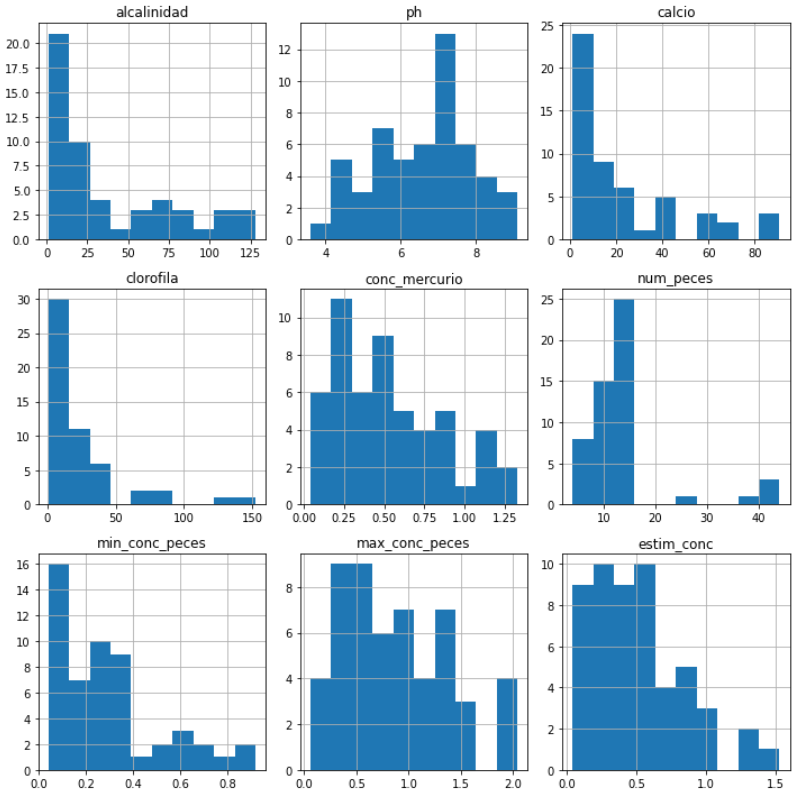
### Variables cuantitativas:

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



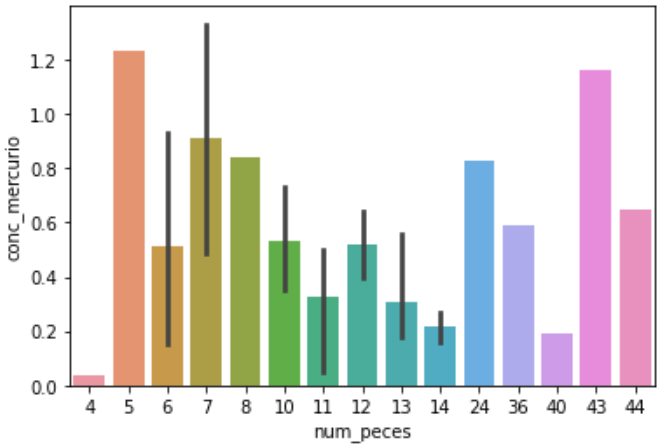
| Variable | Cantidad de outliers |
|---|---|
| alcalinidad | 0 |
| ph | 0 |
| calcio | 0 |
| clorofila | 2 |
| media_merc_porc | 0 |
| num_peces | 3 |
| min_merc_porc | 0 |
| max_merc_porc | 0 |
| merc_estimado_porc | 1 |

**Análisis de distribución de los datos (Histogramas).**



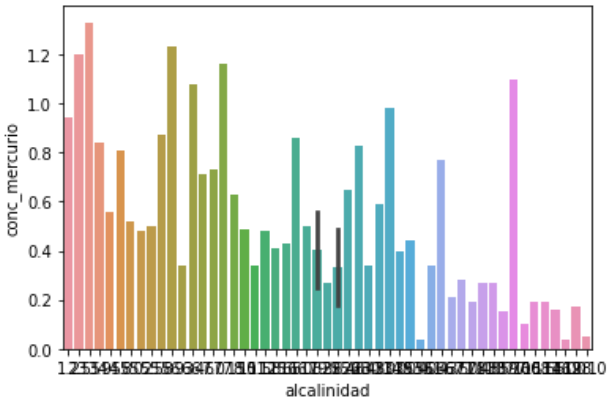
- En la mayoría de las variables los datos no se comportan de manera simétrica y no se observa una distribución normal.
- Se observan outliers en las variables calcio, clorofila, número de peces, concentración mínima de mercurio, concentración máxima de mercurio y estimación de la concentración.

**Comparación de número de peces con concentración de mercurio**



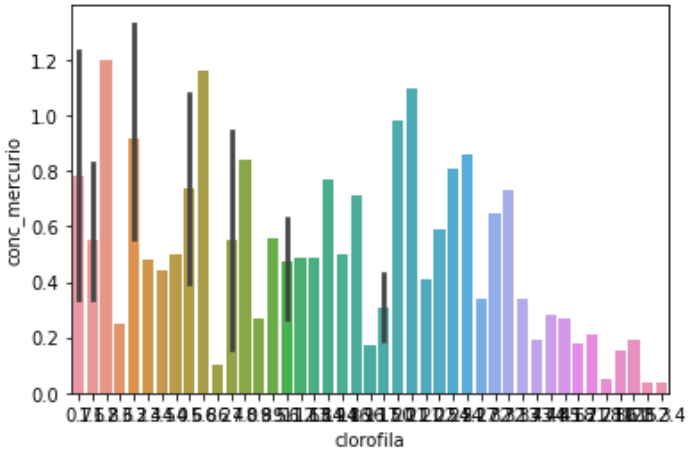

- La concentración media de mercurio cambia con la cantidad de peces.
- La concentración del mercurio comparada con la cantidad de peces cambia de forma aleatoria, es decir, no se observa algún patrón o tendencia que se siga.

**Comparación de alcalinidad con concentración de mercurio**



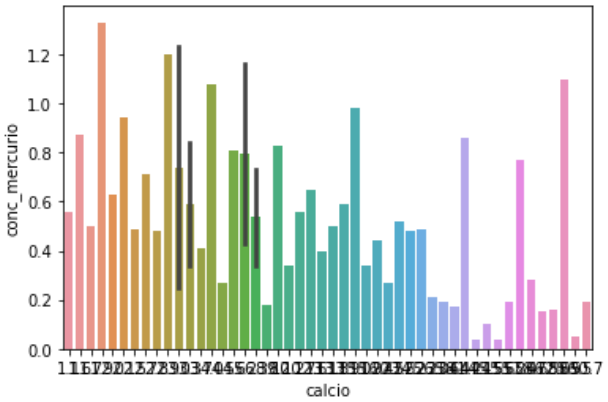

- Parece que mientras aumenta el valor de la alcalinidad, la concentración media de mercurio disminuye.

**Comparación de clorofila con concentración de mercurio**




- La clorofila se comporta de forma similar a la alcalinidad al relacionarse con la concentración de mercurio, este tipo de relación podría indicar una correlación negativa y que la clorofila y la alcalinidad están correlacionadas.

**Comparación de calcio con concentración de mercurio**

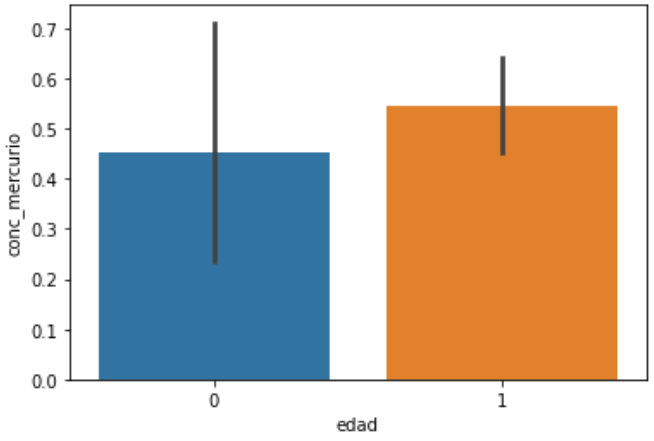



- En la relación del calcio con la concentración de mercurio también se observa una tendencia a que la concentración disminuya cuando el valor de la clorofila aumenta, sin embargo esta no está tendencia no es tan notoria como en las variables anteriores.

### Variables categóricas

Se realizaron diagramas de barras, de pastel y se comparó la media de las variables respecto a la concentración de mercurio y se obtuvieron las siguientes observaciones:

- La concentración media de mercurio cambia dependiendo del lago.




- La concentración media de mercurio no varía mucho dependiendo de la edad, aun así, parece que hay menos concentración en los peces jóvenes.

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

Después de haber hecho la exploración de los datos y analizar los histogramas de cada una de las variables, se decidió realizar una normalización de los datos, así como un escalamiento estándar. Esto debido a que en ninguna de las variables se observa un comportamiento normal y este es necesario, pues más adelante se va a realizar una regresión lineal para obtener la significancia de cada una de las variables, así como predecir la concentración media de mercurio en grupos de peces. El escalamiento estándar se realiza para que las variables a utilizar tengan una misma escala y media cero.

Una vez realizada la normalización se obtuvo la matriz de correlación entre las variables, así como los valores p de la correlación.

**Correlaciones más fuertes**

- Hay una correlación mayor a 0.92 entre las variables mínimo de concentración de mercurio, máximo de concentración de mercurio y estimación de la concentración de mercurio en el pez de 3 años, con la variable de concentración media de mercurio, lo cual tiene sentido pues se trata de datos provenientes del mismo grupo de peces.

- Las variables calcio y alcalindad tienen una correlación de 0.83, lo cual tiene sentido pues la alcalinidad está medida en mg por litros de carbonato de calcio.

- Las variables ph y alcalinidad tienen una correlación de 0.72, ya que se trata de medidas similares, pues el ph es una medidad de acidez o alcalinidad en el agua.

**Hipótesis de correlación entre variables independientes y variable objetivo**

$\alpha$=0.05

A continuación se muestran las hipótesis de correlación, mostrando el valor p encontrado a un lado del nombre de la variable (ejemplo: alcalinidad 0.0000).

**alcalinidad** 0.0000

$H_0 : \rho = 0$

$H_0 : \rho \neq 0$

Se rechaza la hipótesis nula si:

- valor p < 0.05

Ya que el valor p de la alcalinidad en relación a la concentración media de mercurio es de 0.0000, se rechaza la hipótesis nula y se concluye que esta variable tiene correlación con la con concentración.

**ph** 0.0000

$H_0 : \rho = 0$

$H_0 : \rho \neq 0$

Se rechaza la hipótesis nula si:

- valor p < 0.05

Ya que el valor p del ph en relación a la concentración media de mercurio es de 0.0000, se rechaza la hipótesis nula y se concluye que esta variable tiene correlación con la con concentración.

**calcio** 0.0029

$H_0 : \rho = 0$

$H_0 : \rho \neq 0$

Se rechaza la hipótesis nula si:

- valor p < 0.05

Ya que el valor p del calcio en relación a la concentración media de mercurio es de **0.0029**, se rechaza la hipótesis nula y se concluye que esta variable tiene correlación con la con concentración.

**clorofila** 0.0002

$H_0 : \rho = 0$

$H_0 : \rho \neq 0$

Se rechaza la hipótesis nula si:

- valor p < 0.05

Ya que el valor p de la clorofila en relación a la concentración media de mercurio es de **0.0002**, se rechaza la hipótesis nula y se concluye que esta variable tiene correlación con la con concentración.

**num_peces** 0.5737

$H_0 : \rho = 0$

$H_0 : \rho \neq 0$

Se rechaza la hipótesis nula si:

- valor p < 0.05

Ya que el valor p del número de peces en relación a la concentración de mercurio es de **0.5737**, no se rechaza la hipótesis nula y se concluye que esta variable no tiene correlación con la con concentración.

**min_conc_peces** 0.0000

$H_0 : \rho = 0$

$H_0 : \rho \neq 0$

Se rechaza la hipótesis nula si:

- valor p < 0.05

Ya que el valor p del mínimo de la concentración de mercurio en relación a la concentración media de mercurio es de **0.0000**, se rechaza la hipótesis nula y se concluye que esta variable tiene correlación con la con concentración.

**max_conc_peces** 0.0000

$H_0 : \rho = 0$

$H_0 : \rho \neq 0$

Se rechaza la hipótesis nula si:

- valor p < 0.05

Ya que el valor p del máximo de la concentración de mercurio en relación a la concentración media de mercurio es de **0.0000**, se rechaza la hipótesis nula y se concluye que esta variable tiene correlación con la con concentración.

**estim_conc** 0.0000

$H_0 : \rho = 0$

$H_0 : \rho \neq 0$

Se rechaza la hipótesis nula si:

- valor p < 0.05

Ya que el valor p del mínimo de la concentración de mercurio en relación a la estimación de la concentración media de mercurio es de **0.0000**, se rechaza la hipótesis nula y se concluye que esta variable tiene correlación con la con concentración.

**edad** 0.4383

$H_0 : \rho = 0$

$H_0 : \rho \neq 0$

Se rechaza la hipótesis nula si:

- valor p < 0.05

Ya que el valor p de la edad en relación a la concentración media de mercurio es de **0.4383**, no se rechaza la hipótesis nula y se concluye que esta variable no tiene correlación con la con concentración.

# ANALIZA LOS DATOS Y PREGUNTA BASE

## Modelo de regresión lineal múltiple

**Variables que tienen correlación con la concentración media encontradas:**
Alcalinidad, ph, calcio, clorofila, mínimo de la concentración de mercurio, máximo de la concentración de mercurio, estimación de la concentración de mercurio

**Variables que no tienen correlación con la concentración media encontradas:**
num_peces, edad

Debido a que las tres últimas variables enlistadas arriba son medidas de concentración, estas no se toman en cuenta para contestar la pregunta de investigación, dado que no son factores que influyen en el nivel de contaminación por mercurio en los peces, sino que estos solo miden la concentración del mercurio. Sin embargo, encontramos que la alcalinidad, el ph, el calcio y la clorofila tienen una correlación significativa con la concentración media de mercurio, por lo que son esas las variables que se van a utilizar para el modelo.

Se buscó un modelo de regresión lineal múltiple utilizando el método mixto, para así conocer cuales son las variables que generan el mejor modelo para estimar la concentración media de mercurio utilizando como criterio el criterio de información de Akaike.

Como resultado se obtuvo que las variables clorofila y alcalinidad son con las que se generaría el mejor modelo.

Posteriormente se obtuvo el mejor modelo analizando el coeficiente de determinación y realizando las siguientes pruebas a los residuos: normalidad, verificación de media cero, homocedasticidad.

# Verificación del modelo

## Coeficiente de determinación

**Coeficiente de determinación =** 0.5057

**Coeficiente de determinación ajustado =** 0.486

El coeficiente de determinación tiene un valor de 0.5057 lo cual indica que hay algo de variación explicada por el modelo, sin embargo, no es el valor más óptimo.

## Normalidad

**Valor p en la prueba de Shapiro:** 0.1523

- En la qqplot se observó que la mayoría de los residuos se ajustan idealmente a la linea normal.
- La forma del histograma se asemeja a una distribución normal.
- El valor-p en la prueba de Shapiro tiene un valor de 0.1523, este es mayor a 0.05 por lo que no se rechaza la hipótesis nula.

Tomando en cuenta las observaciones, se llegó a la conclusión que los residuos son normales.

## Verificación de media cero

**Hipótesis**

$H_0 = 0$

$H_1 \neq 0$

**Regla de decisión**

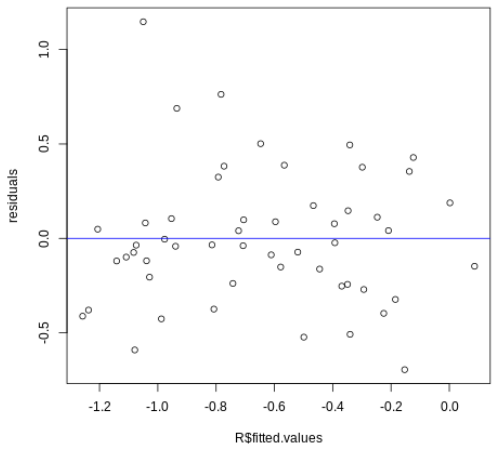
Se rechaza H_0 si:

- Valor p < $\alpha$

**Valor p de la prueba t de student: 1**

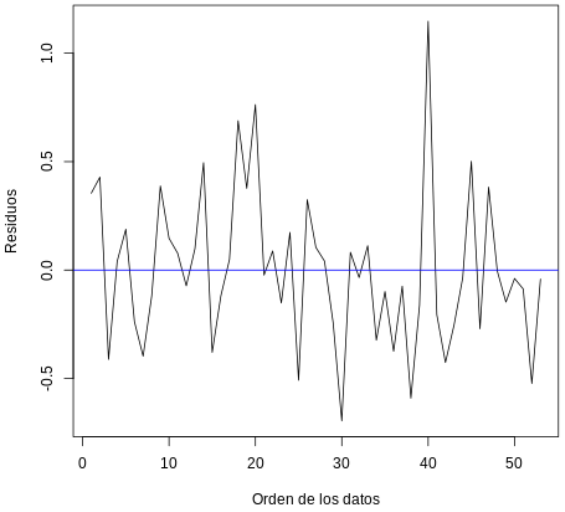
El valor p es mayor a 0.05, por esta razón no se rechaza la hipótesis nula y se concluye que la media de los residuos es 0.

## Homocedasticidad




Los puntos no siguen algún patrón visible y se encuentran esparcidos alrededor de la línea, por lo que parecen ser homogéneos.

## Independencia




Los datos no siguen ningún patrón y parecen acomodarse de forma aleatoria, por lo que se puede decir que son independientes.

## Conclusión de verificación de modelo

Ya que los datos son normales, hay homocedasticidad e independencia, se concluye que este modelo es apto para la base de datos y puede ser útil para predecir la concentración media de mercurio en un grupo de peces con la alcalinidad y clorofila.

# Conclusión

**¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida?**

Para el realizamiento de este análisis hubo preguntas clave que ayudaron en el proceso:

- **¿Hay evidencia para suponer que la concentración promedio de mercurio en los lagos es dañina 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.

- **¿Habrá diferencia significativa entre la concentración de mercurio por la edad de los peces?**

  Con base en la gráfica de comparación encontrada en la sección de exploración de datos en donde se compara la edad de los peces con la concentración de mercurio, así como en la correlación de ambas variables. Se encontró que la concentración media de mercurio no varía mucho dependiendo de la edad y que no hay correlación entre estas dos variables. Por lo que puede decirse que no hay diferencias significativas entre la concentración de mercurio por la edad de los peces.

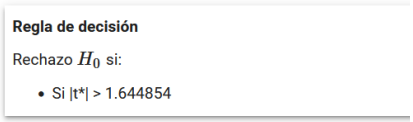
- **Si el muestreo se realizó lanzando una red y analizando los peces que la red encontraba ¿Habrá influencia del número de peces encontrados en la concentración de mercurio en los peces?**

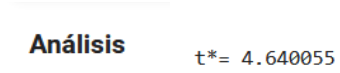
  Con base en la gráfica de comparación encontrada en la sección de exploración de datos en donde se compara la edad de los peces con la concentración de mercurio, así como en la correlación de ambas variables. Se encontró que la concentración media de mercurio cambia con la cantidad de peces y que esta relación cambia de forma aleatoria, es decir, no se observa algún patrón o tendencia que se siga; mientras que la correlación de ambas variables es de 0.079. Por lo que según lo deducido del análisis no hay influencia del número de peces encontrados con la concentración media de mercurio en los peces.

**Hipótesis**

$H_0: \mu = 0.5$
$H_1: \mu > 0.5$

```
t0=  2.674689
```




Debido a que t* es mayor a 1.64 se rechaza la hipótesis nula, lo que significa que la media de concentración de mercurio es mayor a 0.5.

Se concluye que la concentración promedio de mercurio en los peces, según la muestra evaluada, supera los 0.5 mg de Hg/kg. Por lo que hay evidencia para suponer que la concentración promedio de mercurio en los lagos de Florida es dañina para la salud humana.

De acuerdo al análisis anterior de los datos, tomando en cuenta los resultados de las preguntas guía, las pruebas de hipótesis sobre la correlación de las variables con la concentración media de mercurio, así como las variables que resultaron ser mejores en la regresión lineal, se concluye que los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida son los valores del calcio, clorofila y alcalinidad.