Los peces y el mercurio

Módulo 1: Estadística para ciencia de datos
Inteligencia Artificial Avanzada para la Ciencia de Datos

Angel Corrales Sotelo
A01562052
a01562052@tec.mx
Grupo 102

Repositorio con todos los procedimientos:
https://github.com/AngelCorso/Los-peces-y-el-mercurio

Índice

INTRODUCCIÓN

El siguiente análisis se realizó con el propósito de identificar los principales factores que influyen en el nivel de contaminación por mercurio en los 53 lagos de Florida.

EXPLORACIÓN DE LA BASE DE DATOS

Explora las variables y familiarízate con su significado.

Variable	Tipo de atributo	Significado
nombre	String	Nombre del lago
alcalinidad	Float	mg/l de carbonato de calcio
ph	Float	grado de acidez o basicidad del agua
calcio	Float	Calcio mg/l
clorofila	Float	Clorofila mg/l
media_merc_porc	Float	Concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago
num_peces	Int	Número de peces estudiados en el lago
min_merc_porc	Float	Mínimo de la concentración de mercurio en cada grupo de peces
max_merc_porc	Float	Máximo de la concentración de mercurio en cada grupo de peces
merc_estimado_porc	Float	Estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible)
edad_peces	Bool	Indicador de la edad de los

		peces (0: jóvenes; 1: maduros)

Identifica la cantidad de datos y variables presentes.

El conjunto de datos contiene 53 registros y 12 variables incluyendo el identificador.

Clasifica las variables de acuerdo a su tipo y escala de medición.

Variable	Tipo de dato	Escala de medición
nombre	Categórica	Nominal
alcalinidad	Numérica	De razón
ph	Numérica	De razón
calcio	Numérica	De razón
clorofila	Numérica	De razón
media_merc_porc	Numérica	De razón
num_peces	Numérica	De razón
min_merc_porc	Numérica	De razón
max_merc_porc	Numérica	De razón
merc_estimado_porc	Numérica	De razón
edad_peces	Categórica	Nominal

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	media_merc_porc	num_peces	min_merc_porc	max_merc_porc	merc_estimado_porc
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

Variable	Varianza
alcalinidad	1459.51
ph	1.66
calcio	621.65
clorofila	949.65
media_merc_porc	0.12
num_peces	73.29
min_merc_porc	0.05
max_merc_porc	0.27
merc_estimado_porc	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	NaN	NaN	NaN	0.26	NaN
2	NaN	NaN	5.2	9.6	NaN	NaN	NaN	0.4	NaN
3	NaN	NaN	6.3	NaN	NaN	NaN	NaN	0.48	NaN
4	NaN	NaN	20.5	NaN	NaN	NaN	NaN	0.69	NaN
5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.84	NaN
6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.4	NaN
7	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.5	NaN
8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.9	NaN

Variables cualitativas

Tabla de distribución de frecuencia

Variable nombre: Los registro tienen nombres distintos, por lo que la frecuencia de cada valor es 1

edad_peces	Freq
0	10
1	43

Moda

Variable nombre: Ya que todos los nombres se repiten una vez, todos son la moda.

edad_peces
1

Observaciones

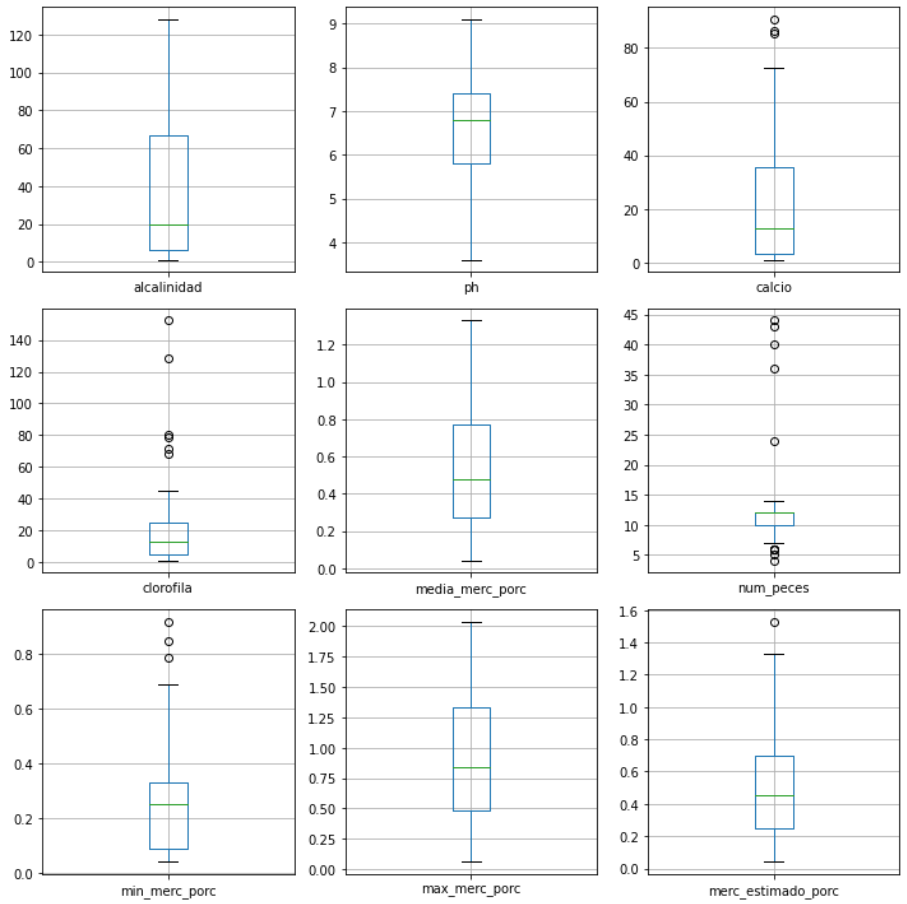
- El valor estimado por la regresión aparenta estar bastante cerca del valor real

- La mayoría de los peces son maduros

Explora los datos usando herramientas de visualización

Variables cuantitativas:

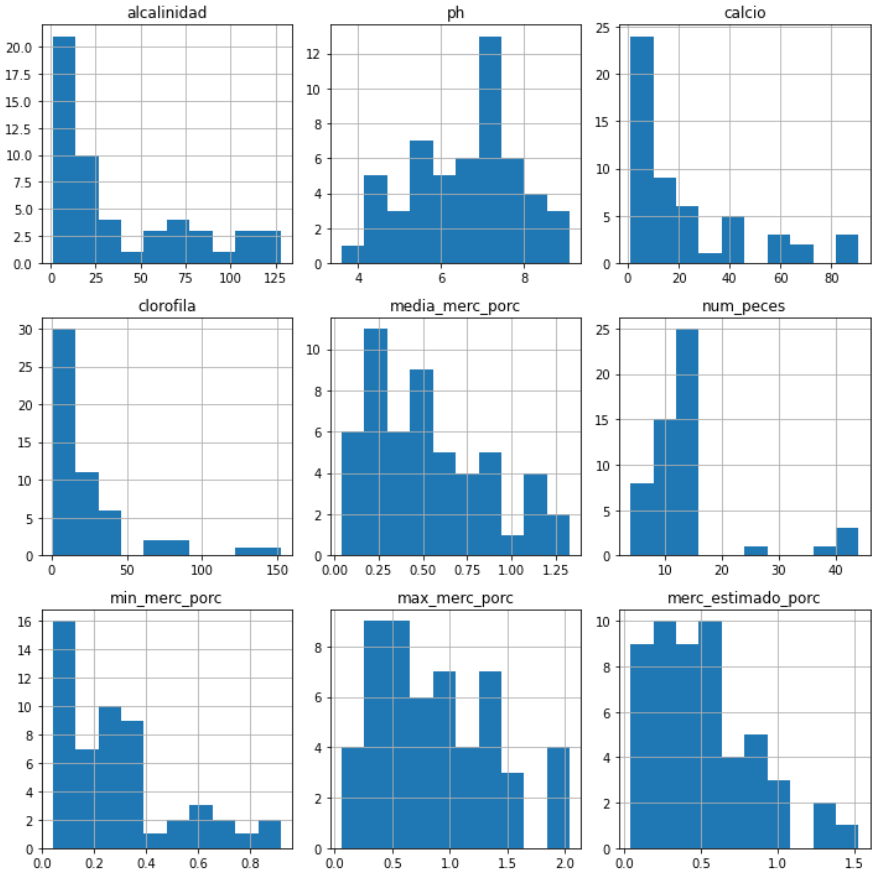
Medidas de posición: cuartiles, outlier (valores atípicos), boxplots



Cantidad de outliers por variable

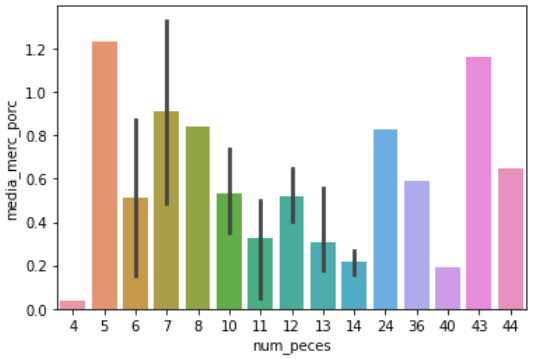
alcalinidad 0
ph 0
calcio 0
clorofila 2
media_merc_porc 0
num_peces 3
min_merc_porc 0
max_merc_porc 0
merc_estimado_porc 1

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



- Las variables no se comportan de manera simétrica.

Comparaciones de número de peces con concentración de mercurio

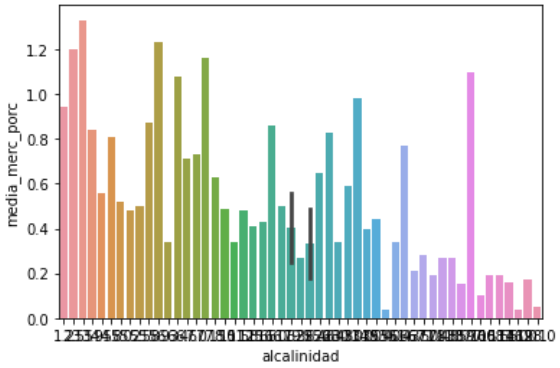


Observaciones

- La concentración media de mercurio cambia con la cantidad de peces.
- Parece que 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.

Angel Corrales Sotelo A01562052

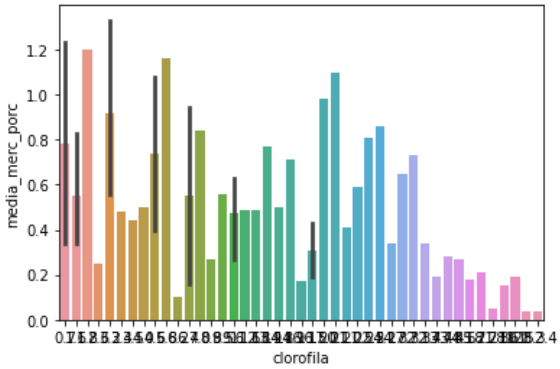
Comparación de alcalinidad con concentración de mercurio



Observaciones

- Parece haber una tendencia negativa, disminuyendo la concentración de mercurio cuando aumenta la alcalinidad. (Posible variable a tomar en cuenta)

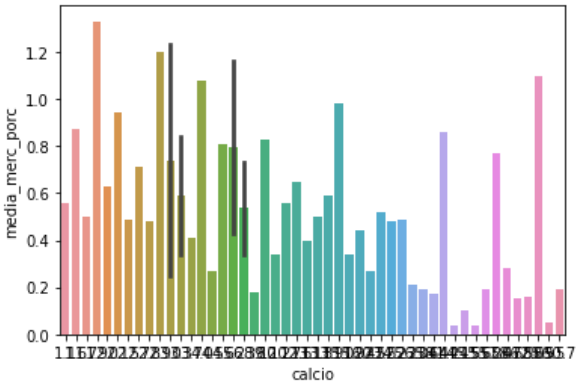
Comparación de clorofila con concentración de mercurio



Observaciones

- Aunque aparentemente se muestre una tendencia negativa igual que la alcalinidad, es menos notable y con valores menos consistentes.

Comparación de calcio con concentración de mercurio



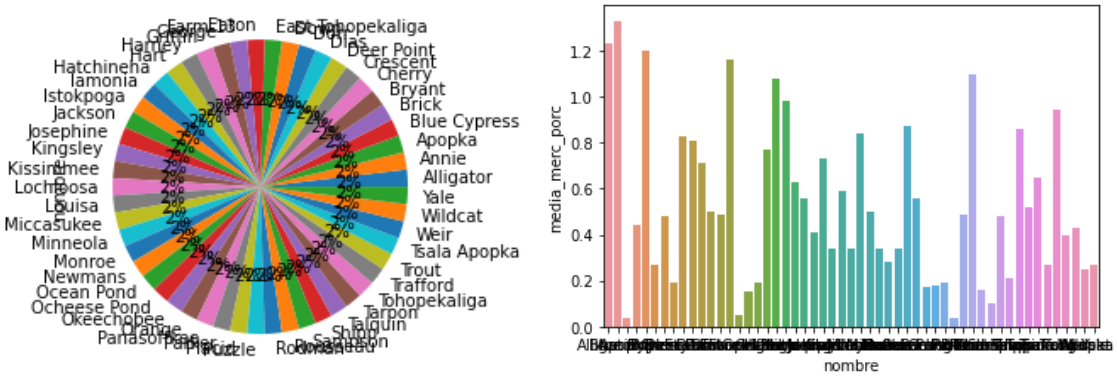
Observaciones

- Parecen haber 3 intervalos consecutivos que muestran una tendencia negativa

Variables categóricas

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

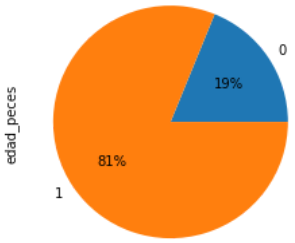
Nombre

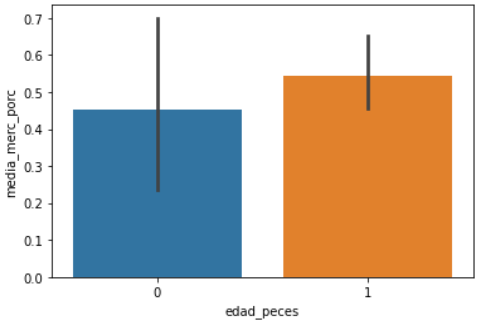


Observaciones

- No parece haber algún patrón en la concentración de mercurio con respecto al nombre del lago

edad_peces

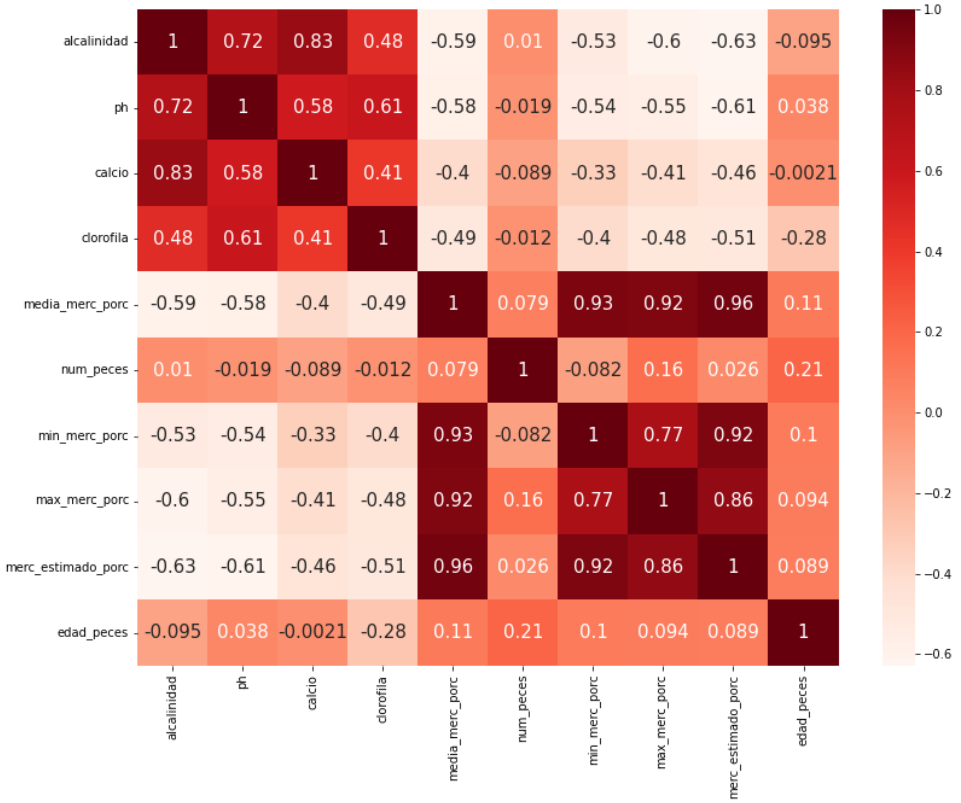




Observaciones

- Aunque la concentración de mercurio es aparentemente mayor en los peces maduros, la desviación estándar de los peces jóvenes es muy alta y engloba de muy cerca la media de los maduros. (Podría no ser significante para el modelo).
- Ya que no se repite ningún nombre, se decidió no tomarlo en cuenta para el modelo.

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



Observaciones

- La correlación del mínimo, máximo y estimado de concentración de mercurio con la media de concentración de mercurio 0.93, 0.92, 0.96 respectivamente.
- La alcalinidad y el ph muestran una correlación alta de 0.72 debido a que el ph muestra el nivel de acidez del agua. (Posible razón para realizar un cambio en la preparación de datos)

- La alcalinidad y el calcio muestran una correlación de 0.83. Ya que la alcalinidad es una medida con respecto al calcio. (Posible razón para realizar un cambio en la preparación de datos)

ANALIZA LOS DATOS Y MODELO

Buscando el mejor modelo

Variables eliminadas por el contexto del problema

Las variables **min_merc_porc**, **max_merc_porc** y **merc_estimado_porc**, el mínimo y máximo son obtenidos tras la captura de datos y como intentamos predecir la concentración del mercurio no tendría sentido tomarlas en cuenta, pues tenerlas significa que ya conocemos dicho valor. El valor estimado ya es una regresión, por lo que tampoco se tomará en cuenta.

Normalización

Tras la exploración de modelos, se optó por normalizar los datos para que se cumplieran los supuestos de los residuos. Esto se realizó utilizando la transformación de Box-Cox, proceso que se hizo en cada una de las variables.

Escalamiento

Se decidió escalar los datos mediante el escalamiento estándar, de manera que cada dato se centra en la media y se divide entre su desviación estándar.

Análisis de correlación

Hipótesis

Para analizar la correlación se establecen las siguientes hipótesis:

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

Regla de decisión:

$\alpha = 0.05$

Se rechaza H_0 si valor p < α

Angel Corrales Sotelo A01562052

Análisis de resultado

alcalinidad

valor p = 0.0000

Debido a que el valor p de la alcalinidad es menor de 0.05, la hipótesis nula se rechaza.

Por lo tanto $\rho \neq 0$, hay correlación.

ph

valor p = 0.0000

Debido a que el valor p del ph es menor de 0.05, la hipótesis nula se rechaza.

Por lo tanto $\rho \neq 0$, hay correlación.

calcio

valor p = 0.0029

Debido a que el valor p del calcio es menor de 0.05, la hipótesis nula se rechaza.

Por lo tanto $\rho \neq 0$, hay correlación.

clorofila

valor p = 0.0002

Debido a que el valor p de la clorofila es menor de 0.05, la hipótesis nula se rechaza.

Por lo tanto $\rho \neq 0$, hay correlación.

num_peces

valor p = 0.5737

Debido a que el valor p del número de peces es mayor de 0.05, la hipótesis nula no se rechaza

Por lo tanto $\rho = 0$, no hay correlación.

edad_peces

valor p = 0.4383

Debido a que el valor p de la edad de los peces es mayor de 0.05, la hipótesis nula no se rechaza.

Por lo tanto $\rho = 0$, no hay correlación.

Variables significativas por correlación

- Alcalinidad
- Ph
- Calcio
- Clorofila

Variables no significativas por correlación

- Número de peces
- Edad de los peces

Elección del mejor modelo

Se decidió el mejor modelo utilizando el método mixto de eliminación y agregación de atributos así como la utilización del criterio de información de Akaike.

```
Start:  AIC=-30.34
media_merc_porc ~ alcalinidad + ph + calcio + clorofila

             Df Sum of Sq    RSS     AIC
- ph          1    0.0415 24.800 -32.251
- calcio      1    0.9431 25.702 -30.358
<none>                    24.759 -30.340
- clorofila   1    1.0836 25.842 -30.069
- alcalinidad 1    6.1090 30.867 -20.651

Step:  AIC=-32.25
media_merc_porc ~ alcalinidad + calcio + clorofila

             Df Sum of Sq    RSS     AIC
- calcio      1    0.9016 25.702 -32.358
<none>                    24.800 -32.251
- clorofila   1    1.5601 26.360 -31.017
+ ph          1    0.0415 24.758 -30.340
- alcalinidad 1    8.2080 33.008 -19.098

Step:  AIC=-32.36
media_merc_porc ~ alcalinidad + clorofila

             Df Sum of Sq    RSS     AIC
<none>                    25.702 -32.358
+ calcio      1    0.9016 24.800 -32.251
- clorofila   1    1.3443 27.046 -31.656
+ ph          1    0.0000 25.702 -30.358
- alcalinidad 1   13.0297 38.731 -12.623

Call:
lm(formula = media_merc_porc ~ alcalinidad + clorofila)
```

Se encontró que el mejor modelo para predecir la concentración media de mercurio en un lago es mediante los atributos alcalinidad y clorofila.

Análisis de los residuos

Normalidad de los residuos

Hipótesis:

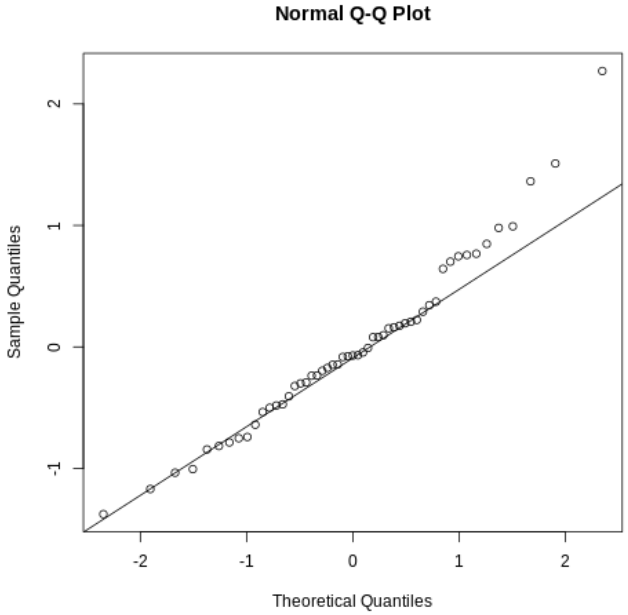
H_0 : Los datos provienen de una población normal

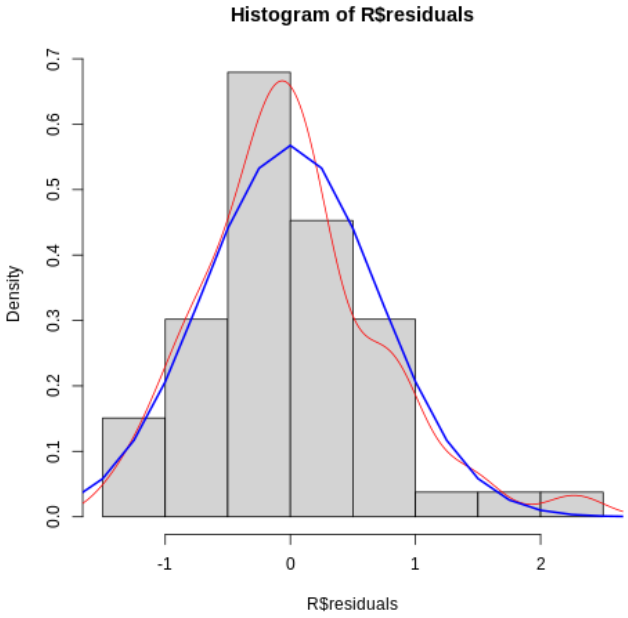
H_1 : Los datos no provienen de una población normal

Regla de decisión:

$\alpha = 0.05$

Se rechaza H_0 si valor p < α






```
        Shapiro-Wilk normality test

data:  R$residuals
W = 0.96719, p-value = 0.1523
```

Conclusión

Debido a que el valor p es mayor a 0.05, no se rechaza la hipótesis nula, por lo tanto los datos provienen de una población normal

Angel Corrales Sotelo A01562052

Verificación de media cero

Hipótesis:

$H_0 : \mu = 0$

$H_1 : \mu \neq 0$

Regla de decisión:

$\alpha = 0.05$

Se rechaza H_0 si valor p < α

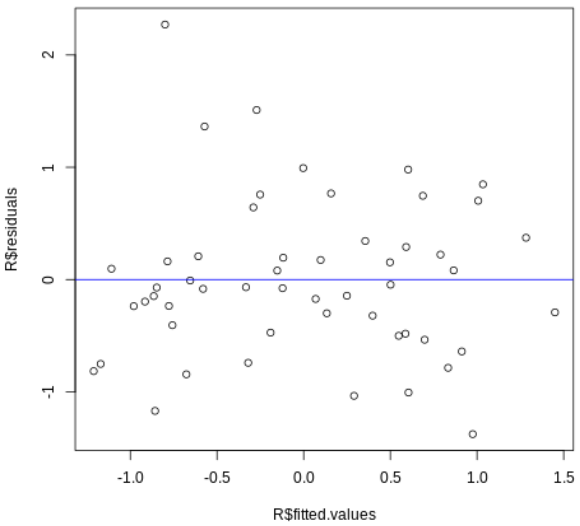
```
        One Sample t-test

data:  R$residuals
t = 2.2182e-16, df = 52, p-value = 1
```

Conclusión

Ya que el valor p es 1, mayor a α=0.05 , la hipótesis nula no se rechaza, por lo tanto la media de los residuos es 0

Homocedasticidad



Conclusión

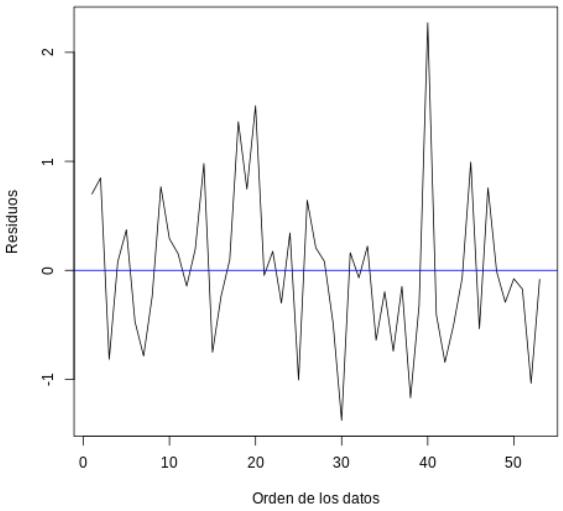
Se puede observar que los datos no siguen ninguna tendencia distribuyéndose de manera homogénea a excepción de algunos datos aparentemente atípicos.

Independencia

Prueba de autocorrelación para verificar independencia:

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$




```
lag Autocorrelation D-W Statistic p-value
  1      0.02690187      1.926779   0.766
Alternative hypothesis: rho != 0
```

Conclusión

No parecen encontrarse patrones ni tendencias entre los residuos, además de tener un valor p mayor a 0.05, por lo que no se rechaza la hipótesis nula, no hay autocorrelación y por lo tanto hay independencia.

Conclusión del análisis de los residuos

Debido a los resultados de los análisis anteriores de normalidad, homocedasticidad, media cero e independencia, podemos confiar en la fiabilidad del modelo.

Análisis Modelo

Coeficiente de determinación

Coeficiente de correlación: 0.5057. La proporción de varianza explicada por el modelo, aun cuando esta no es óptima, no es un valor cercano a 0.

F-statistic

Hipótesis

H_0 : El modelo no tiene capacidad predictiva

H_1 : El modelo tiene capacidad predictiva

Regla de decisión

$\alpha = 0.05$

Se rechaza la hipótesis nula si:

valor p < α

Conclusión

Para la F-statistic se presenta un valor p menor a 0.05, por lo que se rechaza su hipótesis nula, por lo que el modelo sí tiene capacidad predictiva.

PREGUNTA BASE Y COMPLEMENTARIAS

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

   Tras el análisis realizado, podemos concluir que los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida son la alcalinidad y la clorofila.

   Aun cuando otras variables tienen correlación con la concentración de mercurio, como el ph o calcio, se concluyó que la mejor forma de distinguir el nivel de contaminación es mediante la utilización de las variables mencionadas anteriormente.

2. ¿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.

   Para determinar si hay evidencia suficiente para suponer lo anterior es necesario hacer una prueba de hipótesis en cuanto a la media supuesta y la encontrada en la muestra de datos proporcionada:

Hipótesis

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

Regla de decisión

Rechazo H_0 si:

- Si el valor p es menor a alfa (0.05)
- Si |t*| > $|t_0|$ (2.674)

t*= 4.640055

valor p = 1.196306e-05

Conclusión

- Como la $|t*| = 4.640$ es mayor a $|t_0| = 2.674$, se rechaza H_0
- Como valor p = 1.196306e-05 es menor a 0.05 (alpha), se rechaza H_0

El verdadero promedio de concentración de mercurio en productos de la pesca supera los 0.5 mg de Hg/kg. Por lo que podemos concluir que es dañino para salud humana con un 95% de confianza.

3. ¿Habrá diferencia significativa entre la concentración de mercurio por la edad de los peces?

   Según el análisis de correlación realizado podemos concluir que la edad de los peces no es significativa para determinar la concentración de mercurio.

4. 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?

   El número de peces no es significativo para determinar la concentración de mercurio, conclusión a la que se llegó tras analizar la correlación entre ambas variables.

5. ¿Las concentraciones de alcalinidad, clorofila, calcio en el agua del lago influyen en la concentración de mercurio de los peces?

   Dichas variables resultaron significativas para determinar la concentración de mercurio de los peces, sin embargo, tras el análisis y los métodos utilizados para encontrar el mejor modelo se concluyó que la mejor forma de identificar la concentración de mercurio es tan solo con el nivel de alcalinidad y clorofila.