Reporte “Los peces y el mercurio”

Materia

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)
Módulo 1: Estadística para ciencia de datos

Autor

Facundo Vecchi - A01283666

18 de septiembre del 2022, Monterrey N.L.

Resumen

La contaminación por mercurio de peces en el agua dulce comestibles es una amenaza directa contra nuestra salud. Se llevó a cabo un estudio reciente en 53 lagos de Florida con el fin de examinar los factores que influían en el nivel de contaminación por mercurio. Se realizo un análisis estadístico para poder determinar si se puede predecir el nivel de mercurio en los peces de los lagos de Florida en el cual se utilizó un modelo de regresión lineal simple. Se determinó que la concentración media de mercurio de los lagos de Florida se puede predecir con un 95% de confianza utilizando el promedio de mercurio en cada grupo de peces. Esto se comprobó utilizando varias pruebas de hipótesis y verificación de supuestos.

Introducción

A través del análisis realizo se intenta contestar la siguiente pregunta:

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

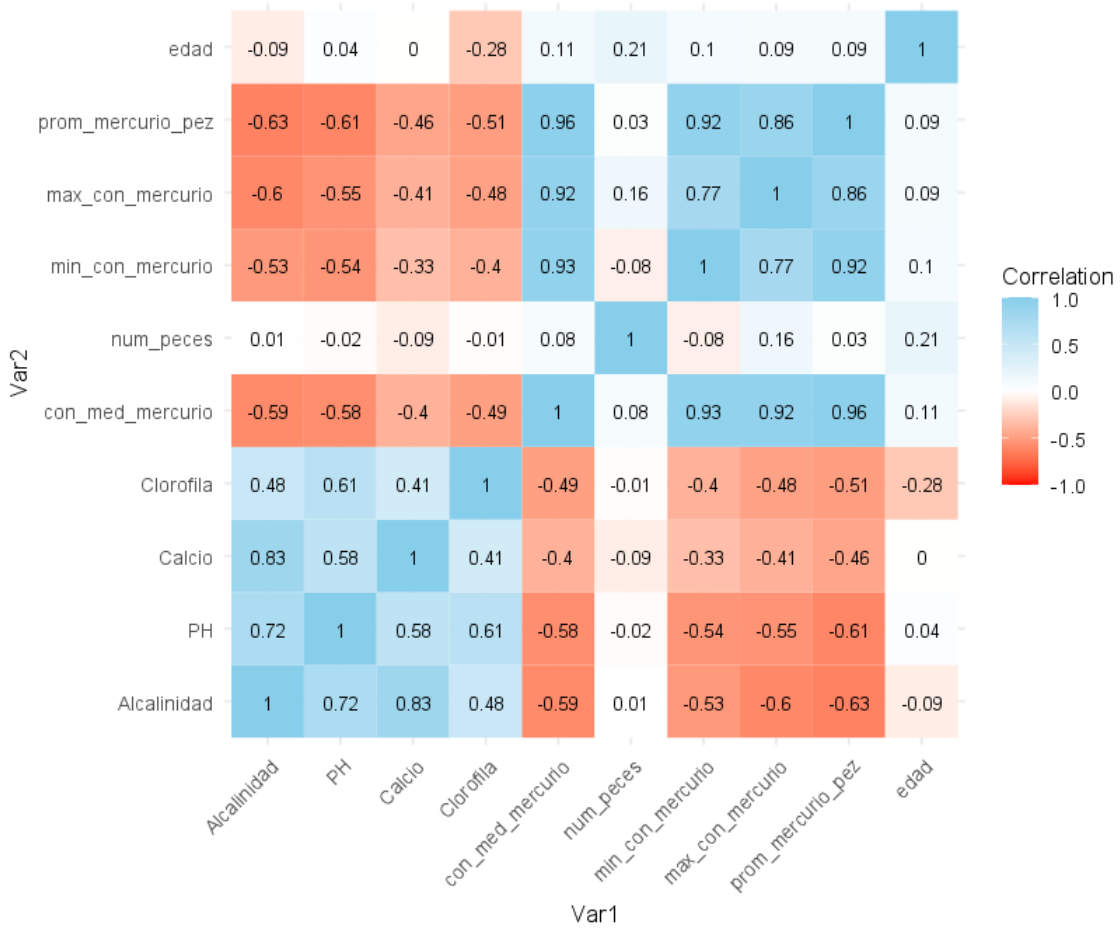
El mercurio es un metal pesado que se encuentra en la naturaleza y que se puede acumular en los organismos acuáticos. Es un metal tóxico que puede causar daños irreparables en los humanos. Es por esto por lo que el nivel de contaminación de mercurio en el lago de los peces es muy importante poder predecir ya que esto nos ayuda a saber si los peces que se consumen son seguros para la salud de las personas o no.

Análisis de los resultados

A continuación se demostrará el proceso que se llevó a cabo para llegar a los resultados previamente establecidos.

Matriz de correlaciones

Para poder contestar la pregunta establecida anteriormente, primero se comenzó analizando las variables para determinar cuáles son las más significativas y que posiblemente se podrían utilizar para un modelo de regresión lineal.



Se genero una matriz de correlaciones entre todas las variables del set datos de mercurio, y como se puede observar en la misma, existe un nivel de correlación entre muchas de estas. La variable que dependiente, la que se intenta predecir, es la concentración media de mercurio en el lago (con_med_mercurio). Si observamos sus correlación con las demás variables, notaremos que tiene una correlación alta con todas excepto con el número de peces y la edad, así que automáticamente estas quedan excluidas. Nos queda Alcalinidad, pH, Calcio, Clorofila, min_con_mercurio, max_con_mercurio, y prom_mercurio_pez, De todas estas la últimas tres mencionadas son las que mayor correlación tienen con la variable dependiente, debido a que nos importa el caso promedio para este problema, utilizaremos prom_mercurio_pez que representa el valor promedio de mercurio por grupo de peces. De las variables restantes, se evaluará la posibilidad de utilizar Alcalinidad en vez del promedio de mercurio por grupo de peces.

Normalidad de las variables

Antes de proceder a la creación del modelo, se verificaron las normalidades de las dos variables que se escogieron. Para verificar la normalidad de estas variables, se realizaron pruebas de normalidad Shapiro-Wilk y se establecieron las siguientes hipótesis:

Hipótesis:
H0: los datos provienen de una distribución normal
H1: los datos no provienen de una distribución normal

Reglas de decisión:

- 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

En cuanto a la variable de prom_mercurio_pez, tras realizar una prueba de normalidad Shapiro-Wilk, se obtuvo un p-value de 0.002782 y en cuanto a la variable de Alcalinidad se obtuvo un p-value de 1.537e-06. Ambas variables fallan la prueba de Shapiro-Wilk, indicando que no se distribuyen como una normal.

Utilizando Box-Cox se transformaron las variables para intentar conseguir que se distribuyan como una normal. Tras la transformación, se realizaron nuevamente pruebas de Shapiro-Wilk y se obtuvo un p-value de 0. 0.4879 para la prom_mercurio_pez y 0.01461 para Alcalinidad.

En este caso, Alcalinidad sigue sin poder distribuirse como una normal y por esta razón se termina descartando la posibilidad de ser utilizada para un modelo de regresión lineal ya que al no distribuirse como una normal bajaría la precisión del modelo.

Modelo de regresión lineal

Utilizando la variable de prom_mercurio pez, se creó un modelo de regresión lineal el cual resulto con las siguiente características:

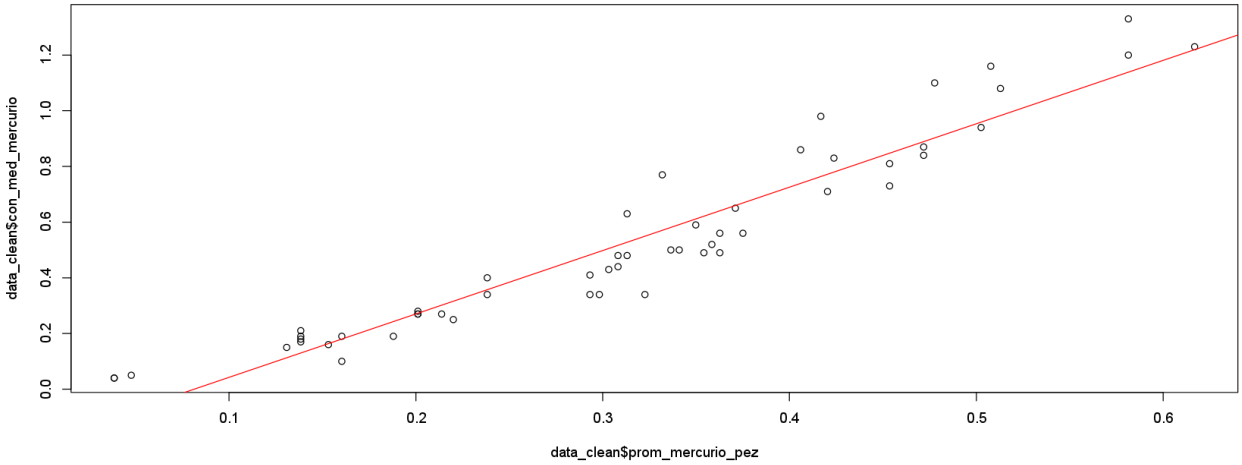
```
Call:
lm(formula = con_med_mercurio ~ prom_mercurio_pez, data = data_clean)

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
```

Como se puede observar en los coeficientes del modelo, este detecta la variable de prom_mercurio_pez como significativa, indicando que paso un prueba de hipótesis indicando su significancia para explicar la variable dependiente, con_med_mercurio.

La ecuación de la regresión lineal es la siguiente:

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

Y al graficarla se ve de la siguiente manera:



A simple vista podemos observar que la línea de regresión lineal parece ajustarse bastante bien a los datos reales. De igual manera a continuación se hará la validación de supuestos del modelo para verificar que si sea bueno y realmente explique la variable con_med_mercurio.

Verificación de supuestos

Para comprobar que el modelo es bueno, se verificaran los supuestos.

Normalidad

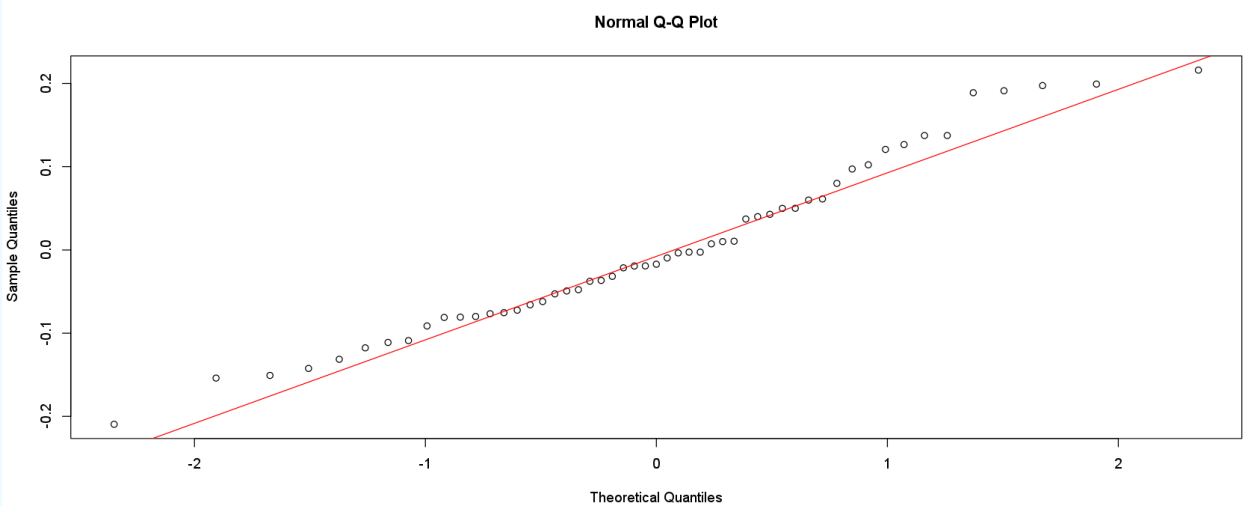
Primero se empezó verificando la normalidad de los residuos, se graficó un qqplot y se realizaron dos pruebas, T test y Shapiro-Wilk.

Hipótesis t test:
H0: miu = 0
H1: miu != 0

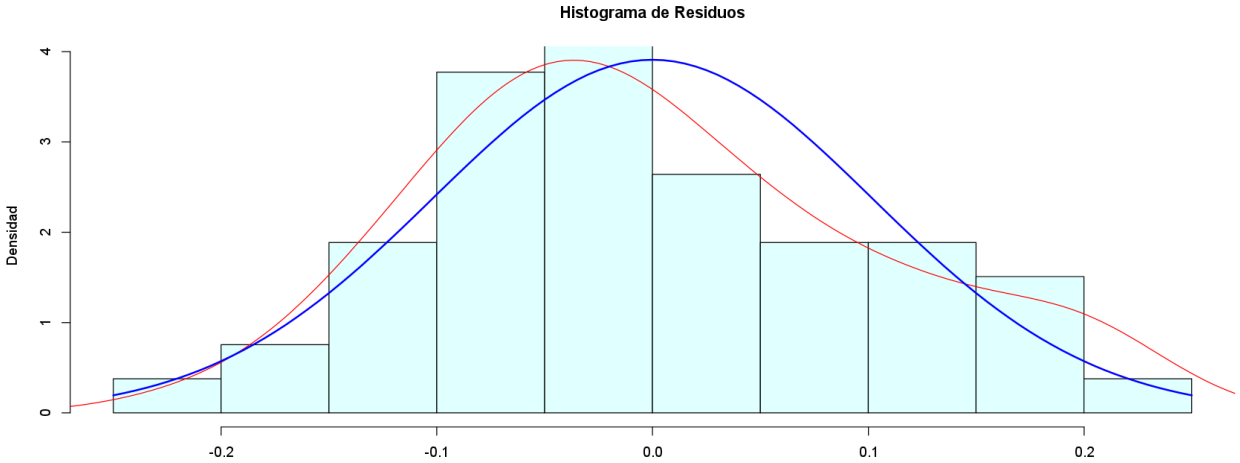
Reglas de decisión:
- Alpha = 0.05
- Si p-value < Alpha, se rechaza H0 y se acepta H1
- Si p-value > Alpha, se rechaza H1 y se acepta H0

Hipótesis Shapiro-Wilk:
H0: los datos provienen de una distribución normal
H1: los datos no provienen de una distribución normal

Reglas de decisión:
- 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



Como se puede observar en el qq plot, en su mayoria, los residuos se distribuyen como normal, hay un poco de sesgo en las colas asi que se verificara la normalidad con las pruebas anteriormente mencionadas.



En el histograma podemos observar lo mismo que se observa en el qqplot, los residuos parecen tener un muy ligero sesgo a la derecha.

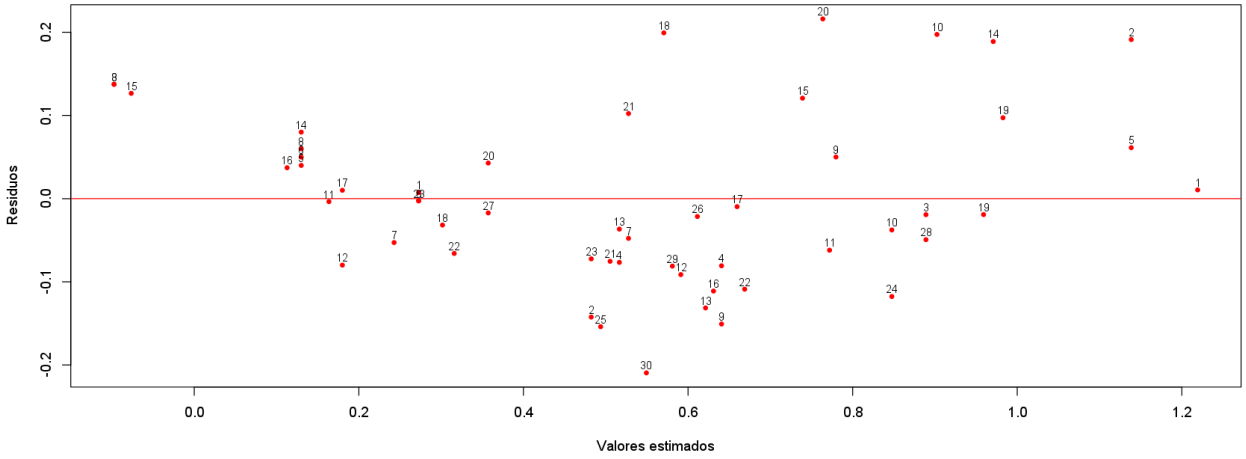
En cuanto a las pruebas realizadas, el T test dio que los residuos tienen una media de -1.700822e-18 y de la prueba de normalidad de Shapiro-Wilk se obtuvo un p-value de 0.1886**.** En ambas pruebas se acepta la hipótesis nula, indicando que los residuos se distribuyen de manera normal.

Homocedasticidad

Para comprobar la homocedasticidad del modelo, se graficaron los residuos y se realizó una prueba de Breusch-Pagan para verificar los resultados.

Hipótesis:
H0: existe homocedasticidad
H1: existe heterocedasticidad

Reglas de decisión:
- 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



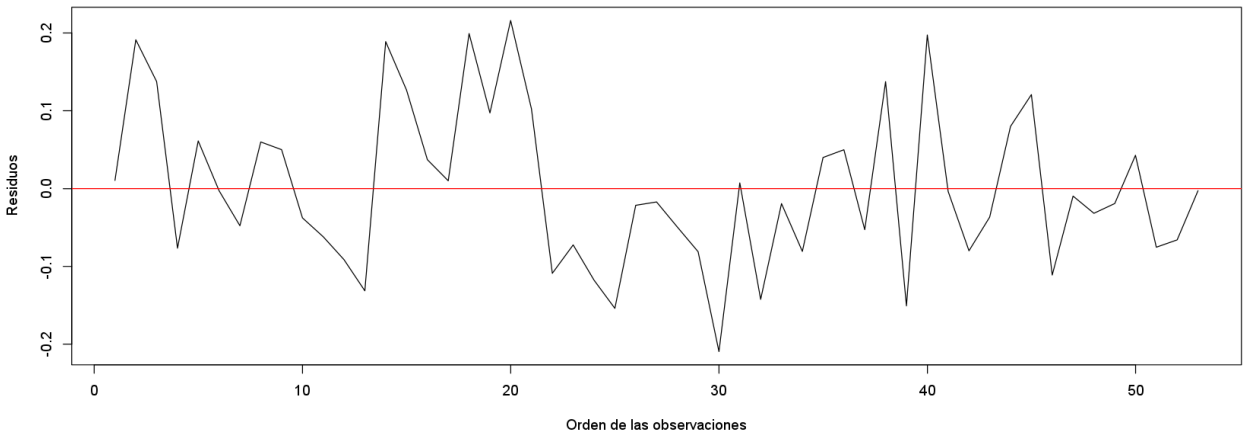
Como se puede observar en la gráfica, los residuos no parecen seguir algún tipo de patrón especifico o tendencia, aunque como tal no se distribuyen de manera simétrica. Es por esto que se acudió a la prueba de Breusch-Pagan de la cual se obtuvo un p-value de 0.113 lo que indica que la hipótesis nula se acepta así afirmando que hay homocedasticidad.

Independencia

Para verificar la independencia, se graficaron los residuos en orden y se realizó una prueba de Durbin Watson.

Hipótesis:
H0: rho = 0
H1: rho != 0

Reglas de decisión:
- 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



Como se puede observar en la gráfica, los residuos no siguen algún tipo de patrón indicando independencia, de igual manera en la prueba de Durbin Watson, se obtuvo un p-value de 0.226 lo que significa que se acepta la hipótesis nula así afirmando independencia.

Conclusiones

Ya que se comprobó el modelo de regresión lineal generado es adecuado para explicar la variable dependiente, concentración media de mercurio en los lagos, se puede responder la pregunta establecida al inicio de este reporte.

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

Los principal factores que influyen en el nivel de contaminación por mercurio es prom_mercurio_pez. Aunque, esta no seria la respuesta completa. Debido al contexto del problema se puede aproximar este de diferentes maneras. Dependiendo de lo que se quiera obtener, el máximo de mercurio en cada grupo de peces (max_con_mercurio) podría ser significativa también. Si se quiere obtener cual seria el peor caso, entonces utilizar el promedio no seria optimo, debido a que estas dos variables tienen una correlación muy alta entre sí, se pudiera esperar un modelo con significancia similar al obtenido en este análisis.

Anexo

Link de GitHub conteniendo todas las entregas del bloque:
https://github.com/facund015/ai_avanzada_personal

Todos los archivos utilizados para realizar este análisis se encuentran bajo el siguiente directorio del repositorio de GitHub:
ai_avanzada_personal/Periodo_1/Portafolios/Implementacion/Modulo_1/Peces_Mercurio/

Link directo al directorio:
https://github.com/facund015/ai_avanzada_personal/tree/main/Periodo_1/Portafolios/Implementacion/Modulo_1/Peces_Mercurio

Link de Google drive conteniendo todos los archivos utilizados para este análisis:
https://drive.google.com/drive/folders/19xVZH6oFDA2PEXob1xy6_K0tlVOWTtJm?usp=sharing