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 5. Estadística avanzada para ciencia de datos

Profesora: Blanca Ruiz

Grupo: TC3007C.502

03 de diciembre de 2022

Índice

Resumen

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.

Para abordar esta problemática, se realizó un análisis de normalidad para conocer el comportamiento de los datos, así como un análisis de componentes principales para identificar los factores principales que intervienen en el problema de la contaminación por mercurio de los peces en agua dulce.

Introducción

La exposición al mercurio, incluso si se trata de pequeñas cantidades, puede causar problemas graves en la salud de los seres vivos. Hay varias formas en las que un ser humano puede estar expuesto al mercurio y unas de las más comunes es consumir peces o mariscos que estuvieron viviendo en agua contaminada con este. (WHO, 2017). Debido a que se trata de un problema de salud para humanos y

animales, se realizó un estudio en los lagos de Florida, con el fin de examinar los factores influyentes en el nivel de la contaminación por mercurio en estos lagos. En este análisis se utilizan los datos de ese estudio con el fin de conocer esos factores influyentes, información importante para que ese problema pueda llegar a ser solucionado y se disminuya la amenaza del agua contaminada con mercurio.

Análisis de los resultados

Exploración de los datos

Análisis de normalidad

El conjunto de datos utilizados consta de las siguientes variables:

- X_1 = número de identificación
- X_2 = nombre del lago
- X_3 = alcalinidad (mg/l de carbonato de calcio)
- X_4 = PH
- X_5 = calcio (mg/l)
- X_6 = clorofila (mg/l)
- X_7 = concentración media de mercurio (parte por millón) en el tejido muscualar del grupo de peces estudiados en cada lago
- X_8 = número de peces estudiados en el lago
- X_9 = mínimo de la concentración de mercurio en cada grupo de peces
- X_{10} = máximo de la concentración de mercurio en cada grupo de peces
- X_{11} = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible)
- X_{12} = indicador de la edad de los peces (0: jóvenes; 1: maduros)

Para analizar el comportamiento de los datos identificando las variables normales y detectar posible normalidad entre grupos de variables se realizó la prueba de normalidad de Mardia y la prueba de normalidad de Anderson Darling.

Prueba de Anderson-Darling

	Test <S3: AsIs>	Variable <S3: AsIs>	Statistic <S3: AsIs>	p value <S3: AsIs>	Normality <S3: AsIs>
1	Anderson-Darling	X3	3.6725	<0.001	NO
2	Anderson-Darling	X4	0.3496	0.4611	YES
3	Anderson-Darling	X5	4.0510	<0.001	NO
4	Anderson-Darling	X6	5.4286	<0.001	NO
5	Anderson-Darling	X7	0.9253	0.0174	NO
6	Anderson-Darling	X8	8.6943	<0.001	NO
7	Anderson-Darling	X9	1.9770	<0.001	NO
8	Anderson-Darling	X10	0.6585	0.081	YES
9	Anderson-Darling	X11	1.0469	0.0086	NO
10	Anderson-Darling	X12	14.3350	<0.001	NO

1–10 of 10 rows

* Variables normales según la prueba de Anderson-Darling: X4 y X10

Según los resultados de la prueba de Anderson-Darling, se encontró que las variables PH y máximo de la concentración de mercurio en cada grupo de peces, son las únicas variables que presentan normalidad.

Prueba de normalidad de Mardia

Test <chr>	Statistic <fctr>	p value <fctr>	Result <chr>
Mardia Skewness	474.747945136974	8.64265750182826e-21	NO
Mardia Kurtosis	3.59794900484947	0.000320736483631068	NO
MVN	*NA*	*NA*	NO

3 rows

Tomando en cuenta los resultados de la prueba de Mardia, no se encontró normalidad multivariada entre el grupo de variables analizado.

Dados los resultados anteriores, posteriormente se realizaron las mismas pruebas de normalidad en las variables que sí tuvieron normalidad.

Prueba de Anderson-Darling

	Test <S3: AsIs>	Variable <S3: AsIs>	Statistic <S3: AsIs>	p value <S3: AsIs>	Normality <S3: AsIs>
1	Anderson-Darling	X4	0.3496	0.4611	YES
2	Anderson-Darling	X10	0.6585	0.0810	YES

2 rows

Prueba de normalidad de Mardia

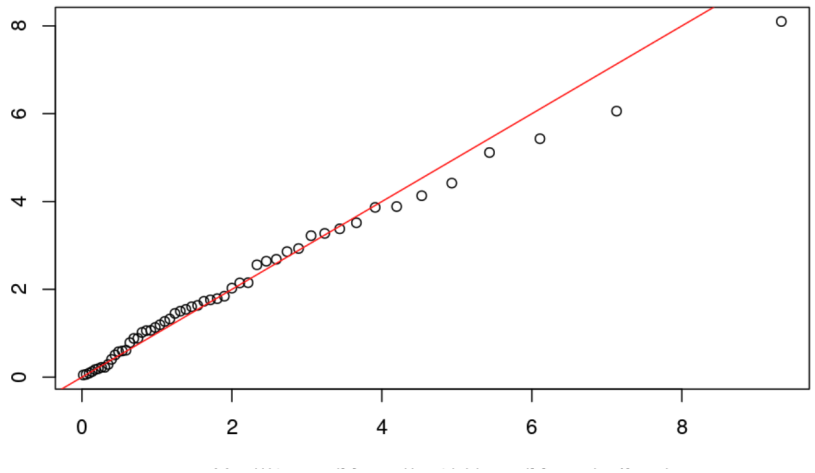
Test <chr>	Statistic <fctr>	p value <fctr>	Result <chr>
Mardia Skewness	6.17538668676458	0.186427564928852	YES
Mardia Kurtosis	-1.12820795824432	0.25923210375991	YES
MVN	*NA*	*NA*	YES

3 rows

Prueba de sesgo y curtosis

Skew <dbl>	Kurtosis <dbl>
-0.2458771	-0.6239638
0.4645925	-0.6692490

Qqplot



Se encontró que las variables PH y máximo de la concentración de mercurio en cada grupo de peces y que este grupo de variables tiene normalidad multivariada, de acuerdo con la prueba de normalidad de Mardia y el gráfico qqplot. De igual forma se encontró que ambas tienen un bajo sesgo y una baja curtosis.

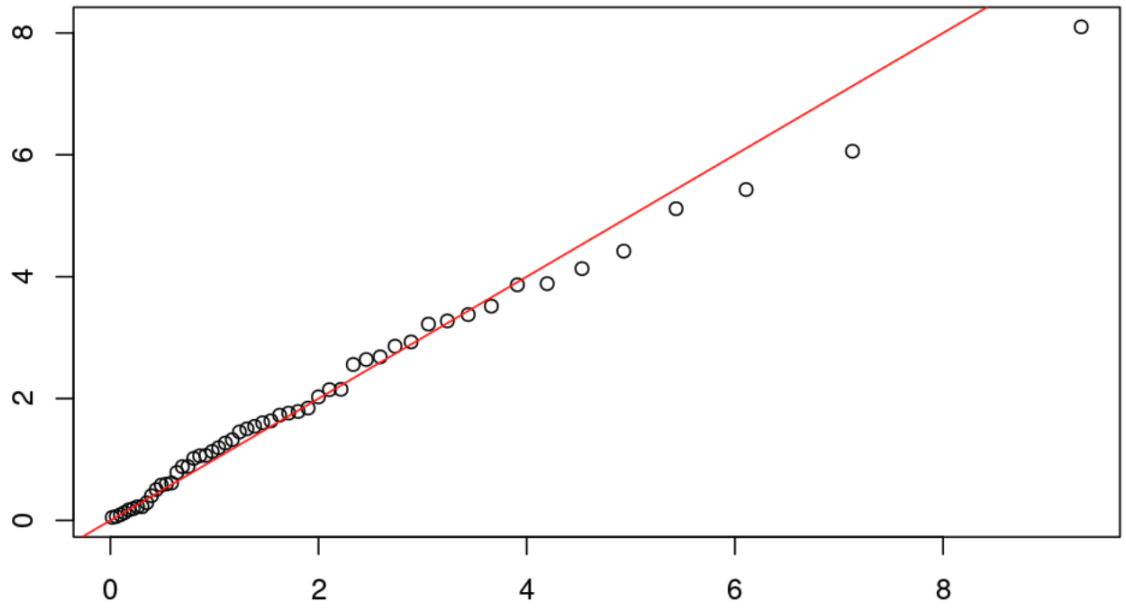
Detección de datos atípicos o influyentes

Para la detección de datos atípicos o influyentes en la normal multivariada de las variables PH y máximo de la concentración de mercurio en cada grupo de peces, se utilizó la distancia de Mahalanobis y la prueba chi-cuadrada.

```
chi-cuadrada 11.82901
 [1] 1.19340732 3.86623312 4.13223576 0.06401297 2.55854258 0.59679099 1.78765285 1.50256004 1.01802464 1.63445643 1.26957081 0.22408796 0.61327299
[14] 5.43003539 2.14558317 2.15128332 2.68426029 5.11541618 0.78631278 1.84332204 3.88561874 0.04988963 0.88374242 8.09988683 2.02671417 0.58001802
[27] 0.50566324 2.64029211 2.92897033 1.06510339 0.88216107 0.40467236 6.05908470 3.51592769 1.45168260 1.32485980 1.13164718 2.85893264 0.22445752
[40] 3.37787450 1.54397421 1.72672287 0.28960341 1.76103444 1.60277372 0.09262808 0.19257450 3.22197239 3.27393627 0.17321589 0.12786732 4.41942776
[53] 1.06000856
```

Teniendo el valor de chi-cuadrada para una probabilidad del 99.73% y con 2 grados de libertad, el cual fue de 11.829, se sabe que las observaciones que tengan un valor de distancia menor o igual a 11.829 son aquellas que están dentro del

contorno de probabilidad estimado del 99.73%, es decir que no son datos atípicos. En este caso ninguna de las distancias fue mayor a 11.829, por lo que no se encontraron datos atípicos.



La gráfica qqplot reafirma nuestro resultado anterior al mostrar que las distancias se ajustan casi idealmente a la línea y sólo unas pocas se desvían.

Análisis de componentes principales

Ya una vez conocido el comportamiento de los datos, se realizó un análisis de componentes principales con el objetivo de identificar los factores principales que intervienen en el problema de la contaminación por mercurio de los peces en agua dulce.

Como primera parte del análisis, se obtuvo la matriz de correlaciones y una alta correlación entre varias de las variables fue observada, por lo que se encontró que es necesario reducir la dimensionalidad utilizando componentes principales, ya que se quiere conservar la mayor cantidad de información posible.

Matriz de correlaciones

```
             X3          X4           X5          X6          X7          X8          X9         X10         X11          X12
X3   1.00000000  0.71916568  0.832604192  0.47753085 -0.59389671  0.01029074 -0.52535654 -0.60479558 -0.62795845 -0.094938825
X4   0.71916568  1.00000000  0.577132721  0.60848276 -0.57540012 -0.01860607 -0.54196524 -0.55181523 -0.61284905  0.038000214
X5   0.83260419  0.57713272  1.000000000  0.40991385 -0.40067958 -0.08937901 -0.33247623 -0.40791663 -0.46440947 -0.002111124
X6   0.47753085  0.60848276  0.409913846  1.00000000 -0.49137481 -0.01182027 -0.40045856 -0.48497215 -0.50644193 -0.283002338
X7  -0.59389671 -0.57540012 -0.400679584 -0.49137481  1.00000000  0.07903426  0.92720506  0.91586397  0.95921481  0.108738958
X8   0.01029074 -0.01860607 -0.089379013 -0.01182027  0.07903426  1.00000000 -0.08165278  0.16109174  0.02580046  0.207956171
X9  -0.52535654 -0.54196524 -0.332476229 -0.40045856  0.92720506 -0.08165278  1.00000000  0.76535319  0.91908939  0.100661967
X10 -0.60479558 -0.55181523 -0.407916635 -0.48497215  0.91586397  0.16109174  0.76535319  1.00000000  0.85975810  0.093752072
X11 -0.62795845 -0.61284905 -0.464409465 -0.50644193  0.95921481  0.02580046  0.91908939  0.85975810  1.00000000  0.089411267
X12 -0.09493882  0.03800021 -0.002111124 -0.28300234  0.10873896  0.20795617  0.10066197  0.09375207  0.08941127  1.000000000
```

El análisis de componentes principales se realizó con la matriz de correlación, ya que los datos no se encontraban escalados, lo cual de haber utilizado la varianza, habría ocasionado que ciertas variables tuvieran más peso debido únicamente a que tenían mayores valores y no por el comportamiento de los datos.

Descomposición en valores y vectores propios

```
eigen() decomposition
$values
 [1] 5.36122641 1.25426109 1.21668138 0.90943267 0.59141736 0.30314741 0.20673634 0.08682133 0.05163902 0.01863699

$vectors
             [,1]        [,2]       [,3]         [,4]        [,5]        [,6]        [,7]        [,8]        [,9]       [,10]
 [1,] -0.35065869 -0.21691594 -0.3472906  0.009131194  0.34050534 -0.07547497 -0.33823501  0.68622998  0.04284021  0.02239801
 [2,] -0.33700381 -0.21940887 -0.2360975 -0.017242162 -0.39396038 -0.73121012 -0.08629646 -0.28769221  0.01363551 -0.04445261
 [3,] -0.28168286 -0.26250672 -0.5113780  0.146950070  0.36205937  0.31342329  0.34312185 -0.45568753 -0.11508339 -0.02634676
 [4,] -0.28334182  0.10195058 -0.2639612 -0.432676049 -0.63093376  0.44112169  0.13435159  0.19006976 -0.06333133  0.03982419
 [5,]  0.39830786 -0.12104244 -0.2996635 -0.080630070 -0.03046869 -0.07436922 -0.01377825 -0.01674789  0.06243320  0.84827636
 [6,]  0.02667579 -0.57556151  0.3050633 -0.692854505  0.19646415  0.05926732 -0.14693148 -0.16809481  0.02532023 -0.04805976
 [7,]  0.36839224 -0.04432459 -0.3876861  0.044658983 -0.13236038  0.19602465 -0.45674057 -0.18260535  0.53803577 -0.35020485
 [8,]  0.37893835 -0.14237181 -0.2024901 -0.167921215  0.02678086 -0.26671839  0.67376588  0.33602914  0.18844932 -0.30445219
 [9,]  0.40206100 -0.05279514 -0.2562319 -0.042242268 -0.05607416 -0.03863899 -0.23387764  0.02613406 -0.80648296 -0.24018040
[10,]  0.05931430 -0.67421026  0.2294446  0.521815581 -0.37253140  0.21612970  0.05759514  0.16451240 -0.02782678  0.01839703
```

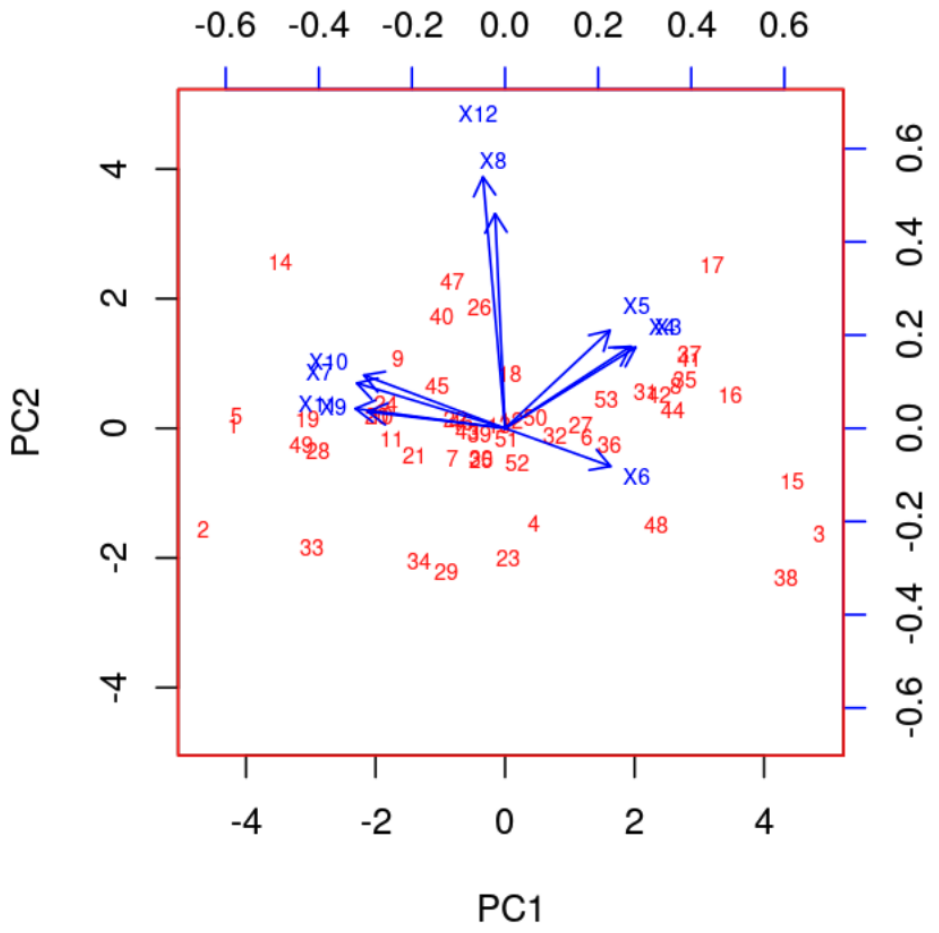
Proporción de varianza explicada acumulada

values <dbl>
0.5361226
0.6615488
0.7832169
0.8741602
0.9333019
0.9636166
0.9842903
0.9929724
0.9981363
1.0000000

Luego de haber hecho la descomposición de los datos en valores y vectores propios se obtuvo la proporción de varianza explicada y posteriormente la proporción de varianza explicada acumulada.

Tomando en cuenta la suma acumulada de la proporción de varianza explicada por los componentes, se encontró que los componentes recomendados para reducir la dimensionalidad son los primeros 5, ya que con estos se explica el 93% de los datos y se reduce la dimensionalidad a la mitad.

Utilizando los vectores propios se realizó un gráfico para conocer las puntuaciones de las observaciones de los dos primeros componentes principales.



Se observó que no varía mucho el peso entre las variables, sin embargo las variables 8 y 10 tienen un peso más bajo que en las otras.También se observó que en el componente dos, las variables 8 y 10 son las que ganan más importancias, mientras que las demás variables mantienen un valor por debajo de |0.2|.

Conclusión

Los componentes que tienen una mayor varianza explicada podrían indicar cuales son las variables que tienen más influencia, dentro de los primeros 5 componentes principales (con los que se llega al 93% de varianza explicada), se identificaron las siguientes variables con más peso en estos: número de peces estudiados en el lago (X8), indicador de la edad de los peces (X12), calcio (X5) y clorofila (X6).

Referencias

World Health Organization. (2017, March 31). *Mercury and Health*. World Health Organization. Retrieved December 2, 2022, from https://www.who.int/news-room/fact-sheets/detail/mercury-and-health

Anexo

Link a notebook