Reporte Final – “Los peces y el mercurio”

Módulo 5: Estadística Avanzada para ciencia de datos
Inteligencia artificial avanzada para la ciencia de datos

Daniel Salvador Cázares García A01197517

2022-12-04

Resumen

En este trabajo, por medio del análisis de datos de un estudio realizado en 53 lagos de Florida, se buscó examinar los factores que influyen en la contaminación por mercurio en peces.

Para lo anterior, se emplearon métodos estadísticos como pruebas de normalidad, normalidad multivariada, sesgo, curtosis y componentes principales, los cuales permitieron explorar las diferentes variables y su relación.

Como resultado general, se encontró que el PH, la alcalinidad, el calcio y la clorofila son los factores que más influyen en la concentración de mercurio.

Introducción

La contaminación por mercurio en cuerpos de agua es un problema ambiental y de salud, que no solo afecta a los seres vivos que habitan estos ecosistemas, sino también a las personas que consumen peces o mariscos provenientes de estos lugares.

A continuación, se utilizarán diferentes métodos estadísticos para analizar los datos de 53 lagos de Florida y examinar ¿Qué factores influyen en el nivel de contaminación por mercurio?

Datos

Se tienen datos de 53 lagos de Florida obtenidos en un estudio reciente. Las variables que se midieron son las siguientes:

- X1 = número de indentificación
- X2 = nombre del lago
- X3 = alcalinidad (mg/l de carbonato de calcio)
- X4 = PH
- X5 = calcio (mg/l)
- X6 = clorofila (mg/l)

- X7 = concentración media de mercurio (parte por millón) en el tejido muscualar del grupo de peces * estudiados en cada lago
- X8 = número de peces estudiados en el lago
- X9 = mínimo de la concentración de mercurio en cada grupo de peces
- X10 = máximo de la concentración de mercurio en cada grupo de peces
- X11 = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de * mercurio cuando la edad no está disponible)
- X12 = indicador de la edad de los peces (0: jóvenes; 1: maduros)

Análisis

Análisis de normalidad

Para analizar el comportamiento de los datos, identificar la distribución de las variables y detectar una posible normalidad multivariada entre variables, se realizaran pruebas de normalidad.

Prueba de Anderson-Darling

De acuerdo con la prueba de Anderson Darling y viendo los valores de p, se observa que las variables X4 (PH) y X10 (máximo de la concentración de mercurio en cada grupo de peces) se comportan de forma normal, mientras que el resto de las variables no.

```
##               Test  Variable Statistic   p value Normality
## 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
```

Prueba de Mardia

En cuanto a la normalidad multivariada, la prueba de Mardia muestra que esta no existe una normalidad multivariada conjunta entre el grupo de variables.

```
##              Test        Statistic              p value Result
## 1 Mardia Skewness 434.339065916421 4.13584083502386e-26     NO
## 2 Mardia Kurtosis 5.76907272063337  7.9708906142173e-09     NO
## 3             MVN             <NA>                 <NA>     NO
```

Prueba de normalidad en variables con normalidad

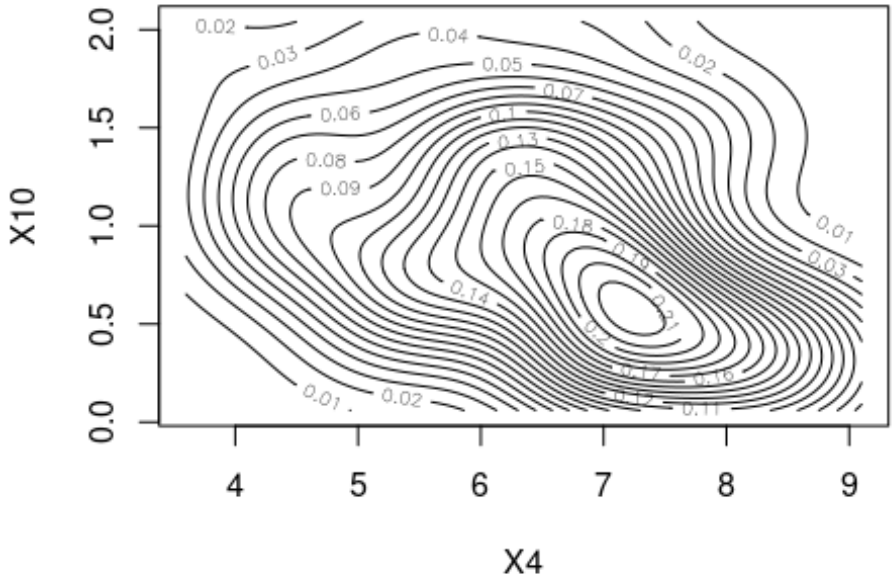
Dado el resultado anterior, se volvieron a realizar ambas pruebas, pero ahora para las variables que mostraron normalidad.

```
##                Test  Variable Statistic   p value Normality
## 1 Anderson-Darling    X4        0.3496    0.4611    YES
## 2 Anderson-Darling    X10       0.6585    0.0810    YES

##              Test          Statistic           p value Result
## 1 Mardia Skewness   6.53855430534145 0.162377302354508    YES
## 2 Mardia Kurtosis -0.889321233851276 0.373830462900113    YES
## 3             MVN               <NA>              <NA>    YES
```

De nuevo se obtiene que X4 y X10 tienen normalidad. Sin embargo, ahora también se puede ver que existe normalidad multivariada conjunta entre ambas variables. Adicionalmente, se observa un bajo sesgo y curtosis, lo cual indica igualmente una cercanía a la normalidad.

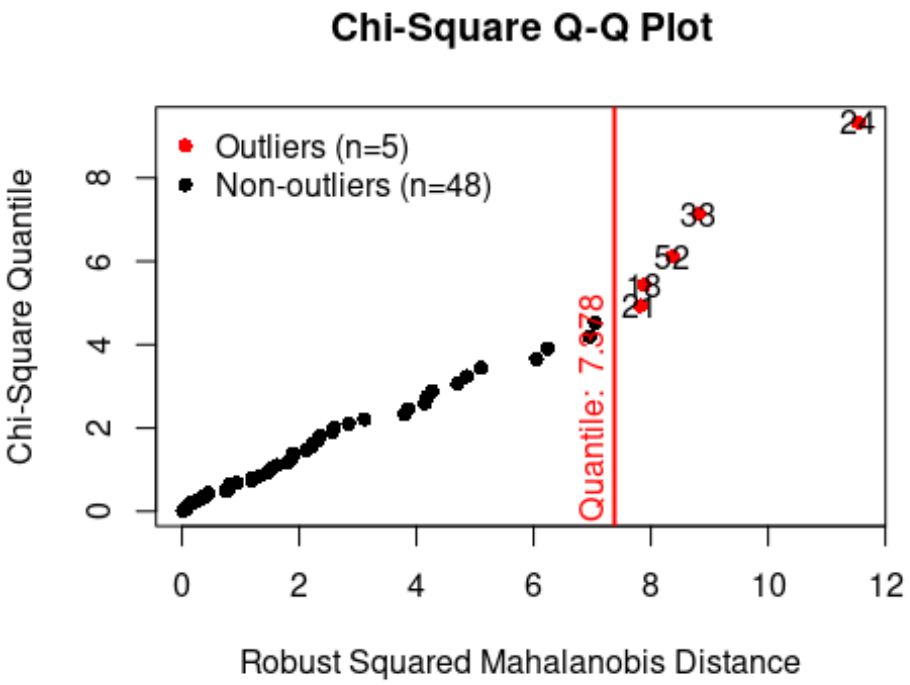
Este comportamiento de normalidad entre ambas variables también se puede ver de forma visual si se realiza una gráfica de contorno.



Detección de datos atípicos

La existencia de normalidad multivariada supone la ausencia de valores atípicos. Por lo tanto, es necesario verificar si existen valores atípicos en los datos. Para detectar datos atípicos en la normal multivariada entre variables se pueden utilizar la distancia de Mahalanobis y el gráfico QQplot (Prueba Chi Cuadrada), los cuales permiten medir la distancia entre 2 puntos en un espacio multivariado.

Por medio de la gráfica realizada se puede observar que existen 5 valores atípicos. Sin embargo, +90% de las observaciones se encuentran dentro del estimado.



Análisis de componentes principales

Después de explorar el comportamiento de los datos y su normalidad, se llevará a cabo un análisis de componentes principales de la base de datos. El uso de esta herramienta es adecuado, pues permitirá identificar los factores que influyen principalmente (las que aportan mayor variabilidad) en la contaminación por mercurio en los peces y reducir la dimensionalidad del conjunto de datos a variables que realmente nos importan.

Matriz de correlaciones

Como primera parte del análisis y puesto que se busca reducir la dimensionalidad del conjunto de datos, se utiliza una matriz de correlación para explorar la correlación entre las variables. Se puede observar que existe una correlación considerable entre varias de las variables, por lo que si es posible el uso de componentes principales.

```
##               X3          X4           X5          X6          X7
X8
## X3   1.00000000  0.71916568  0.832604192  0.47753085 -0.59389671
0.01029074
## X4   0.71916568  1.00000000  0.577132721  0.60848276 -0.57540012 -
0.01860607
## X5   0.83260419  0.57713272  1.000000000  0.40991385 -0.40067958 -
0.08937901
## X6   0.47753085  0.60848276  0.409913846  1.00000000 -0.49137481 -
0.01182027
## X7  -0.59389671 -0.57540012 -0.400679584 -0.49137481  1.00000000
```

```
0.07903426
## X8   0.01029074 -0.01860607 -0.089379013 -0.01182027  0.07903426
1.00000000
## X9  -0.52535654 -0.54196524 -0.332476229 -0.40045856  0.92720506 -
0.08165278
## X10 -0.60479558 -0.55181523 -0.407916635 -0.48497215  0.91586397
0.16109174
## X11 -0.62795845 -0.61284905 -0.464409465 -0.50644193  0.95921481
0.02580046
## X12 -0.09493882  0.03800021 -0.002111124 -0.28300234  0.10873896
0.20795617
##              X9         X10         X11          X12
## X3  -0.52535654 -0.60479558 -0.62795845 -0.094938825
## X4  -0.54196524 -0.55181523 -0.61284905  0.038000214
## X5  -0.33247623 -0.40791663 -0.46440947 -0.002111124
## X6  -0.40045856 -0.48497215 -0.50644193 -0.283002338
## X7   0.92720506  0.91586397  0.95921481  0.108738958
## X8  -0.08165278  0.16109174  0.02580046  0.207956171
## X9   1.00000000  0.76535319  0.91908939  0.100661967
## X10  0.76535319  1.00000000  0.85975810  0.093752072
## X11  0.91908939  0.85975810  1.00000000  0.089411267
## X12  0.10066197  0.09375207  0.08941127  1.000000000
```

Obtención de componentes principales

Para continuar con el análisis de componentes, es necesario descomponer los datos en valores y vectores propios.

```
## eigen() decomposition
## $values
##  [1] 5.36122641 1.25426109 1.21668138 0.90943267 0.59141736 0.30314741
##  [7] 0.20673634 0.08682133 0.05163902 0.01863699
##
## $vectors
##              [,1]        [,2]       [,3]         [,4]        [,5]
[,6]
##  [1,] -0.35065869 -0.21691594 -0.3472906  0.009131194  0.34050534 -
0.07547497
##  [2,] -0.33700381 -0.21940887 -0.2360975 -0.017242162 -0.39396038 -
0.73121012
##  [3,] -0.28168286 -0.26250672 -0.5113780  0.146950070  0.36205937
0.31342329
##  [4,] -0.28334182  0.10195058 -0.2639612 -0.432676049 -0.63093376
0.44112169
##  [5,]  0.39830786 -0.12104244 -0.2996635 -0.080630070 -0.03046869 -
0.07436922
##  [6,]  0.02667579 -0.57556151  0.3050633 -0.692854505  0.19646415
0.05926732
##  [7,]  0.36839224 -0.04432459 -0.3876861  0.044658983 -0.13236038
0.19602465
```

```
##  [8,]  0.37893835 -0.14237181 -0.2024901 -0.167921215  0.02678086 -
0.26671839
##  [9,]  0.40206100 -0.05279514 -0.2562319 -0.042242268 -0.05607416 -
0.03863899
## [10,]  0.05931430 -0.67421026  0.2294446  0.521815581 -0.37253140
0.21612970
##              [,7]        [,8]        [,9]       [,10]
##  [1,] -0.33823501  0.68622998  0.04284021  0.02239801
##  [2,] -0.08629646 -0.28769221  0.01363551 -0.04445261
##  [3,]  0.34312185 -0.45568753 -0.11508339 -0.02634676
##  [4,]  0.13435159  0.19006976 -0.06333133  0.03982419
##  [5,] -0.01377825 -0.01674789  0.06243320  0.84827636
##  [6,] -0.14693148 -0.16809481  0.02532023 -0.04805976
##  [7,] -0.45674057 -0.18260535  0.53803577 -0.35020485
##  [8,]  0.67376588  0.33602914  0.18844932 -0.30445219
##  [9,] -0.23387764  0.02613406 -0.80648296 -0.24018040
## [10,]  0.05759514  0.16451240 -0.02782678  0.01839703
```

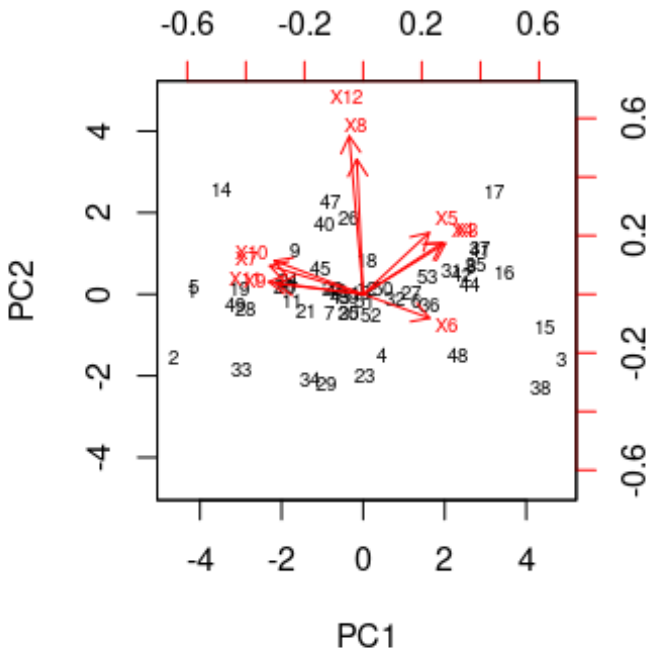
Posteriormente, puesto que se busca encontrar el número de componentes apropiado para reducir la dimensionalidad, es necesario calcular la proporción de varianza explicada acumulada por lo componentes.

Con esto, se puede observar que 4 o 5 ya empieza a ser una cantidad adecuada de componentes principales, pues con los primeros 4 componentes se explica el 87% de la varianza, mientras que 5 componentes ya son suficientes para explicar el 93% de los datos y reducir la dimensionalidad de los datos a la mitad.

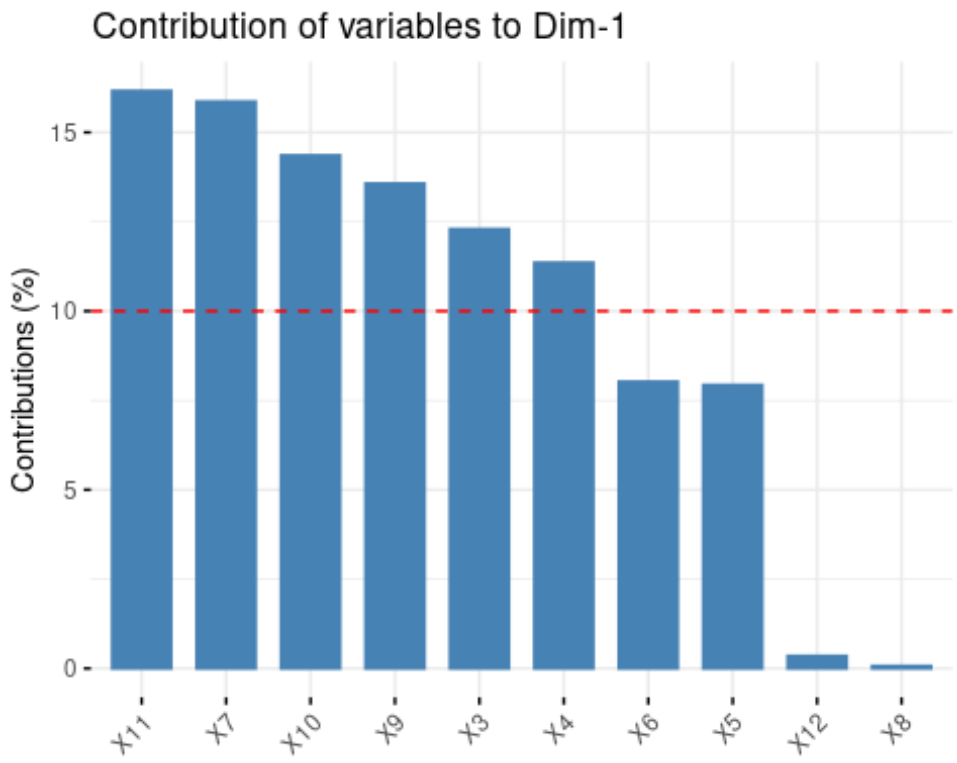
```
##       values
## 1  0.5361226
## 2  0.6615488
## 3  0.7832169
## 4  0.8741602
## 5  0.9333019
## 6  0.9636166
## 7  0.9842903
## 8  0.9929724
## 9  0.9981363
## 10 1.0000000
```

Por último, se realizó un gráfico con los vectores propios asociados a las variables y las puntuaciones de las observaciones de las dos primeras componentes.

Cuanto más largos son los vectores o flechas rojas, mayor es el valor la observación en ese componente. Del mismo modo que se observó con la varianza acumulada, se puede observar que el componente 1 tiene una mayor variabilidad que el componente 2 por la forma en la que distribuyen los datos. Según la dirección, también se puede observar el tipo de relación (positiva o negativa) que tiene cada variable con los componentes. Cabe destacar que X8 es la variable con menos peso y se comporta de forma poco relacionada con el resto de las variables, así como con ambos componentes.



Por último, y puesto que es el objetivo principal de este análisis, se puede observar aquellas variables que contribuyen mayormente a los componentes. Los primeros resultados son aquellas relacionadas con el nivel de mercurio (X11, X7, X10, X9), la cual es la variable objetivo, por lo que no se toman en cuenta. Seguida de estas, se puede observar que aquellas variables que principalmente contribuyen de forma significativa son X3 (alcalinidad) y X4 (PH). En menor medida, X5 (calcio) y X6 (clorofila) también influyen de forma notable.



Conclusión

La utilización de herramientas estadísticas como las pruebas de normalidad y el análisis de componentes principales nos pueden ayudar a realizar mejores modelaciones por medio de la examinación de las diferentes variables.

Tras el análisis realizado, se pudo identificar que las variables que mayormente contribuyen a la contaminación de mercurio en los lagos son:

- X3: Alcalinidad
- X4: PH
- X5: Calcio
- X6: clorofila

Si seleccionamos estas variables, podríamos reducir la dimensionalidad del conjunto de datos y seguir obteniendo un buen análisis de los datos.

Anexos

Repositorio de Github: https://github.com/dscazares/Portafolio-TC3007C