Mercurio_2

Luis Cano Irigoyen A00827178

2022-10-26

Problema

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. Las variables que se midieron se encuentran en mercurio.csv Descargar mercurio.csv y su descripción es la siguiente:

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)

Datos

```
D=read.csv("mercurio.csv")
N=nrow(D)
```

Cambiamos el nombre de las columnas para comprender mejor los análisis

```
colnames(D) <- c("ID","Nombre","Alcalinidad", "PH", "Calcio",
"Clorofila", "MediaMercurio",
                    "NumPez", "MinMercurio", "MaxMercurio",
"TresMercurio","Edad")
head(D, 5)
```

```
##   ID       Nombre Alcalinidad  PH Calcio Clorofila MediaMercurio
NumPez
## 1  1    Alligator         5.9 6.1    3.0       0.7          1.23
5
## 2  2        Annie         3.5 5.1    1.9       3.2          1.33
```

```
7
## 3  3       Apopka       116.0 9.1   44.1     128.3          0.04
6
## 4  4 Blue Cypress        39.4 6.9   16.4       3.5          0.44
12
## 5  5        Brick         2.5 4.6    2.9       1.8          1.20
12
##   MinMercurio MaxMercurio TresMercurio Edad
## 1        0.85        1.43         1.53    1
## 2        0.92        1.90         1.33    0
## 3        0.04        0.06         0.04    0
## 4        0.13        0.84         0.44    0
## 5        0.69        1.50         1.33    1
```

ID y Nombre no son variables númericas y Edad no es variable continua No son explicativas, así que las eliminamos

```
D$ID <- NULL
D$Nombre <- NULL
D$Edad <- NULL
```

1. Análisis de normalidad de las variables continuas para identificar variables normales.

A.

Prueba de normalidad de Mardia y prueba de Anderson Darling para identificar las variables que son normales y detectar posible normalidad multivariada de grupos de variables.

Hipótesis: H_0: Si hay normalidad multivariada

H_a: No hay normalidad multivariada

```
library(MVN)
mvn(D, subset = NULL, mvn = "mardia")
```

```
## $multivariateNormality
##              Test        Statistic               p value Result
## 1 Mardia Skewness  434.33906591642 4.13584083502475e-26     NO
## 2 Mardia Kurtosis 5.76907272063334  7.9708906142173e-09     NO
## 3             MVN             <NA>                 <NA>     NO
##
## $univariateNormality
##               Test      Variable Statistic   p value Normality
## 1 Anderson-Darling  Alcalinidad     3.6725  <0.001      NO
## 2 Anderson-Darling      PH          0.3496  0.4611      YES
## 3 Anderson-Darling    Calcio        4.0510  <0.001      NO
## 4 Anderson-Darling   Clorofila      5.4286  <0.001      NO
```

```
## 5 Anderson-Darling MediaMercurio    0.9253  0.0174      NO
## 6 Anderson-Darling    NumPez        8.6943  <0.001      NO
## 7 Anderson-Darling  MinMercurio     1.9770  <0.001      NO
## 8 Anderson-Darling  MaxMercurio     0.6585   0.081      YES
## 9 Anderson-Darling TresMercurio     1.0469  0.0086      NO
##
## $Descriptives
##                 n       Mean    Std.Dev Median  Min    Max  25th  75th
## Alcalinidad    53 37.5301887 38.2035267  19.60 1.20 128.00  6.60 66.50
## PH             53  6.5905660  1.2884493   6.80 3.60   9.10  5.80  7.40
## Calcio         53 22.2018868 24.9325744  12.60 1.10  90.70  3.30 35.60
## Clorofila      53 23.1169811 30.8163214  12.80 0.70 152.40  4.60 24.70
## MediaMercurio  53  0.5271698  0.3410356   0.48 0.04   1.33  0.27  0.77
## NumPez         53 13.0566038  8.5606773  12.00 4.00  44.00 10.00 12.00
## MinMercurio    53  0.2798113  0.2264058   0.25 0.04   0.92  0.09  0.33
## MaxMercurio    53  0.8745283  0.5220469   0.84 0.06   2.04  0.48  1.33
## TresMercurio   53  0.5132075  0.3387294   0.45 0.04   1.53  0.25  0.70
##                      Skew   Kurtosis
## Alcalinidad     0.9679170 -0.4705349
## PH             -0.2458771 -0.6239638
## Calcio          1.3045868  0.6130359
## Clorofila       2.4130571  6.1042185
## MediaMercurio   0.5986343 -0.6312607
## NumPez          2.5808773  6.0089455
## MinMercurio     1.0729099  0.4060828
## MaxMercurio     0.4645925 -0.6692490
## TresMercurio    0.9449951  0.5733500
```

Contamos con una Mardia Skewness de 434.34 y una Mardia Kurtosis de 5.77 Esta función realiza pruebas de normalidad multivariada de sesgo y curtosis y nos da como resultado que No hay normalidad multivariada. De igual manera, utilizando un nivel de significancia de 0.05, podemos ver como los p-values de sesgo (4.1e-26) y curtosis (7.9e-09) son menores al nivel de significancia, por lo que rechazamos a H_0 y determinamos que No hay nomralidad multivariada en las variables.

Con el Test de Anderson-Darling encontramos que las variables que son normales son PH y MaxMercurio

B.

Prueba de Mardia y Anderson Darling de las variables que sí tuvieron normalidad en los incisos anteriores.

```
mvn(D[, c("PH", "MaxMercurio") ], mvn = "mardia")
```

```
## $multivariateNormality
##              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
##
```

```
## $univariateNormality
##                Test    Variable Statistic   p value Normality
## 1 Anderson-Darling     PH         0.3496    0.4611    YES
## 2 Anderson-Darling MaxMercurio    0.6585    0.0810    YES
##
## $Descriptives
##               n      Mean   Std.Dev Median  Min  Max 25th 75th
Skew
## PH           53 6.5905660 1.2884493   6.80 3.60 9.10 5.80 7.40 -
0.2458771
## MaxMercurio 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33
0.4645925
##                Kurtosis
## PH           -0.6239638
## MaxMercurio -0.6692490
```

A diferencia de la prueba de Mardia con todas las variables, al usar solo las normales (PH y MaxMercurio) obtenemos una curtosis entre -1 y 1, lo cual nos dice que hay normalidad. Asímismo, la prueba de normalidad multivariada de sesgo y curtosis realizada nos da como resultado que Si hay normalidad multivariada, y contamos con p-values que son mayores al nivel de significancia.

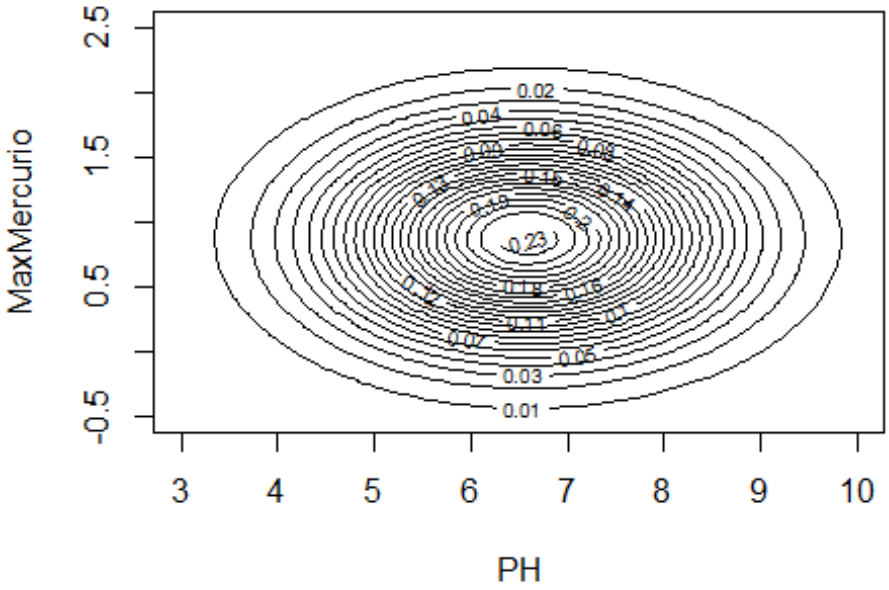
C.

Gráfica de contorno de la normal multivariada obtenida en el inciso B.

```
library(mnormt)

# create bivariate normal distribution
x = seq(3, 10, length.out = 100)
y = seq(-0.5, 2.5, length.out = 100)
mu = c(mean(D$PH), mean(D$MaxMercurio))
sigma <- matrix(c(sd(D$PH)^2, 0, 0,sd(D$MaxMercurio)^2),2,2)
z = outer(x, y, function(x, y) dmnorm(cbind(x, y), mu, sigma))

# create contour plot
contour(x, y, z, nlevels = 20, xlab = "PH", ylab = "MaxMercurio")
```



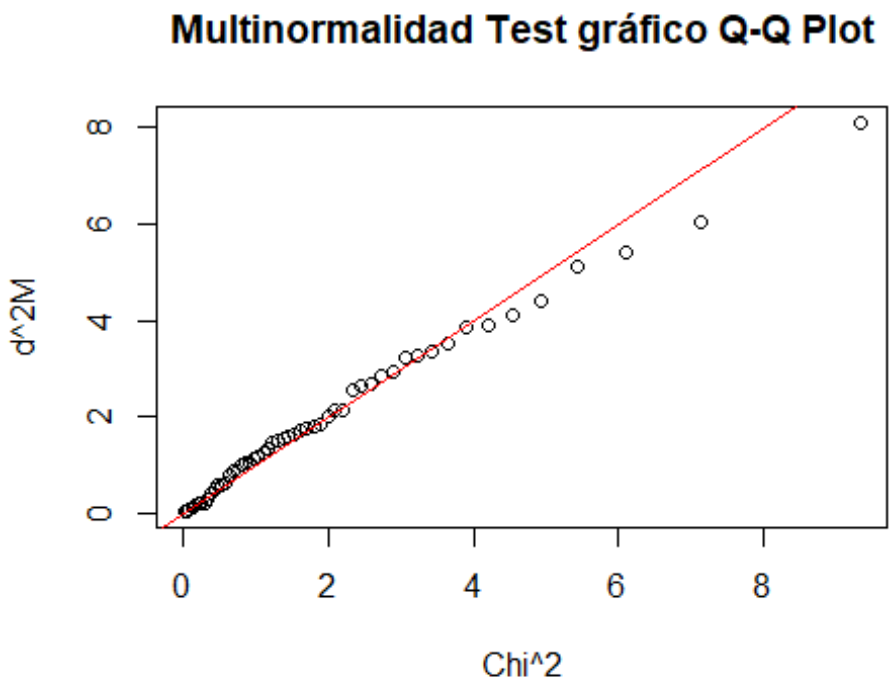
D.

Detecta datos atípicos o influyentes en la normal multivariada encontrada en el inciso B

```
d = D[, c("PH", "MaxMercurio") ]

p = 2 # usando 2 variables
# Vector de medias
X = colMeans(d)
# Matriz de covarianza
S = cov(d)
# Distancia de Mahalanobis
d2M =  mahalanobis(d,X,S)

# Multinormalidad Test gráfico Q-Q Plot
plot(qchisq(((1:nrow(d)) - 1/2)/nrow(d), df=p), sort( d2M ), main =
"Multinormalidad Test gráfico Q-Q Plot", xlab = "Chi^2", ylab = "d^2M")
abline(a=0, b=1,col="red")
```



En la gráfica de distancias de Mahalanobis podemos observar que al inicio los puntos se encuentran cerca de la normalidad multivariada, pero conforme avanza la gráfica en las últimas instancias a la derecha se observa una ligera curva hacia abajo. Los últimos 3 puntos podríamos considerarlos como datos atípicos ya que la distancia se aleja bastante.

2. Análisis de componentes principales con la base de datos completa para identificar los factores principales que intervienen en el problema de la contaminación por mercurio de los peces en agua dulce.

A.

Justifique por qué es adecuado el uso de componentes principales para analizar la base (haz uso de la matriz de correlaciones)

B.

Realiza el análisis de componentes principales y justifica el número de componentes principales apropiados para reducir la dimensión de la base

C.

Representa en un gráfico los vectores asociados a las variables y las puntuaciones de las observaciones de las dos primeras componentes

D.

Interprete los resultados. Explique brevemente a qué conclusiones llega con su análisis y qué significado tienen los componentes seleccionados en el contexto del problema

3. Conclusión general

- ¿de qué forma te ayuda este nuevo análisis a contestar la pregunta principal del estudio: ¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida?
- ¿en qué puede facilitar el estudio la normalidad encontrada en un grupo de variables detectadas?
- ¿cómo te ayudan los componentes principales a abordar este problema?