Los peces y el mercurio

Javier de Golferichs García (A01139500)

30 de octubre de 2022

Abstract

La contaminación por mercurio en lagos puede tener efectos negativos en la salud humana en caso de ingerir peces provenientes de aquellos lagos en donde la concentración sea lo suficientemente alta. Los métodos y tecnicas estadísticas usadas en el reporte incluyen un análisis de normalidad mediante la prueba de Anderson Darling, para analizar normalidad multivariada entre variables y evaluar el sesgo y curtosis de estas. En la segunda parte se usa un análisis de componentes principales, Se encontró que si existe normalidad multivariada entre X4 y X10.

Contents

1 Introducción

Se realizó un estudio sobre la concentración de mercurio en varios lagos de Florida y otras variables relevantes a esta.

Este problema es importante pues contribuye a explicar los factores que influyen en la contaminación por mercurio que puede afectar la salud de los consumidores.

Este documento presenta un análisis de normalidad de las variables para encontrar aquellas que tienen normalidad multivariada, así como de PCA para hacer agrupación de variables y reducir su complejidad al analizarlo.

Al final se presentan las conclusiones del estudio, así como el repositorio en donde se encuentra el código y las referencias utilizadas.

2 Análisis de resultados

En esta sección se presenta un análisis de normalidad multivariada y el análisis de componentes principales.

2.1 Análisis de normalidad

Observese a continuación la prueba Anderson Darling aplicada a las variables X3 a X11, para encontrar variables con normalidad univariada.

```
##               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
```

Con base en lo anterior, se explora si las variables X4 y X10 tienen normalidad multivariada, para lo cual se aplica la prueba de Mardia, y detectar normalidad multivariada en este grupo.

```
## $multivariateNormality
##              Test         Statistic           p value Result
## 1 Mardia Skewness  6.17538668676458 0.186427564928852    YES
## 2 Mardia Kurtosis -1.12820795824432  0.25923210375991    YES
## 3             MVN              <NA>              <NA>    YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling   M.X4       0.3496    0.4611    YES
## 2 Anderson-Darling   M.X10      0.6585    0.0810    YES
##
## $Descriptives
##        n      Mean   Std.Dev Median  Min  Max 25th 75th       Skew   Kurtosis
## M.X4  53 6.5905660 1.2884493   6.80 3.60 9.10 5.80 7.40 -0.2458771 -0.6239638
## M.X10 53 0.8745283 0.5220469   0.84 0.06 2.04 0.48 1.33  0.4645925 -0.6692490
##
## $multivariateOutliers
## NULL
```

En la columna de resultado, se observa que efectivamente si se cuenta con normalidad multivariada.

2.1.1 Gráfica de contorno de normal multivariada previa.

Notese el sesgo del diagrama de contorno presente en la figura 1, el cual al estar cargado a la derecha, se interpreta como la presencia de sesgo a la izquierda, lo cual también se observa en la figura 2. Las elipsoides también confirman la presencia de curtosis como observado en la prueba de Mardia, realizada en la sección anterior.

2.1.2 Detección de datos atípicos en la normal multivariada

Para la detectar los datos atípicos se uso un gráfico de QQplot multivariado y la distancia de Mahalanobis.

Se observa en la figura 2 que el QQ-plot no tiene una pendiente de 1, por lo que se asume que, como fue visto en la prueba de Mardia, los datos padecen de sesgo y curtosis.

Si, bien podría hacerse una transformación para lograr que se asemejen más a una recta de pendiente 1, la siguiente sección no debe ser elaborada con datos transformados, sino los originales.

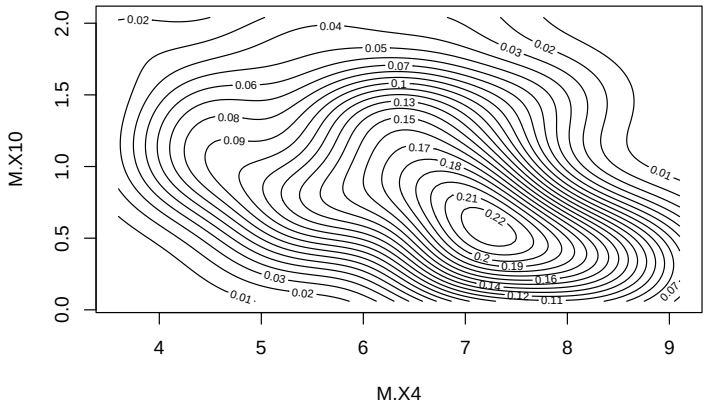

Figure 1: Gráfica de contorno de X4 y X10

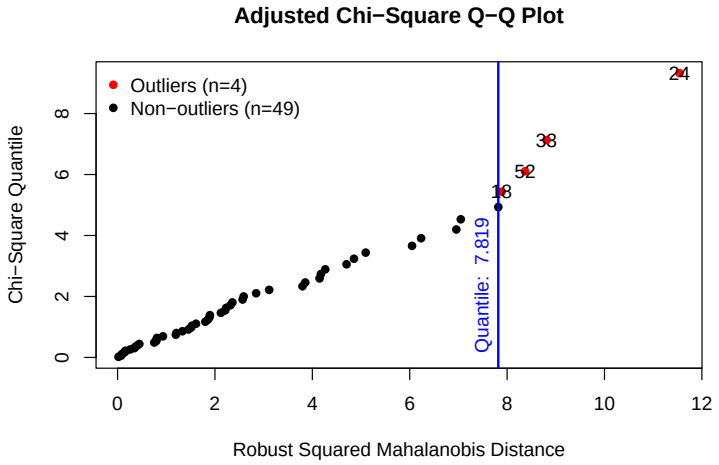

Figure 2: Datos atípicos en QQplot

2.2 Análisis de componentes principales

El método de componentes principales (PCA) busca reducir la dimensionalidad de un problema al agrupar variables que provocan cierto comportamiento en el modelo, asegurando la ortogonalidad de las componentes, tal que estas son independientes entre si. (Johnson and Wichern 2016)

2.2.1 Justificación para usar PCA

El análisis de PCA será aplicado al grupo de las variables X3 a X11 (variables numéricas no categóricas), con la intención de agrupar características que influyen en el modelo y reducir la dimensionalidad y por ende complejidad de este.

Recuerdese que para aplicar PCA no se requiere que las variables en cuestión sean normales, sin embargo la existencia de normalidad si enriquece el análisis, pues como tal PCA no asegura normalidad en las componentes pero si ortogonalidad.

2.2.2 Gráfico de vectores asociados a variables y puntuaciones de las observaciones.

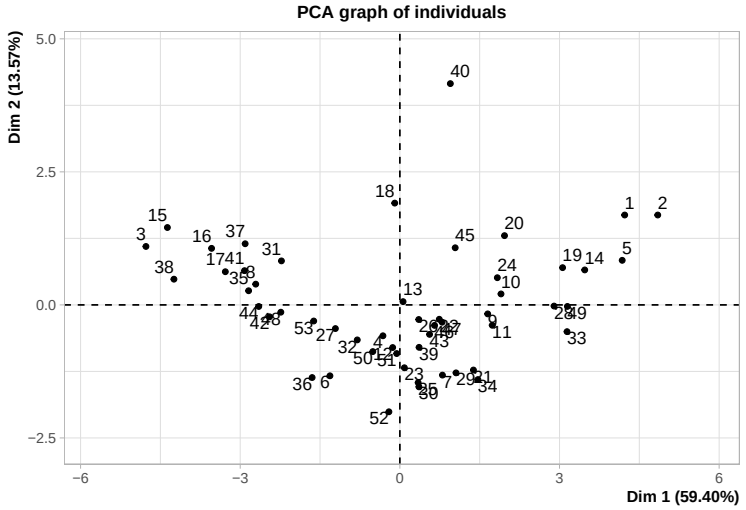

Figure 3: Aplicación de PCA a variables X3 a X11.s

En la figura 4 se observa el PCA por variables, en donde se observa que se hay 2 grupos principales ubicados en los cuadrantes 1 y 2, y que la variable X8 no tiene mucha contribución.

En este caso se observa que Dim 1, es la que separa estos dos grupos, donde el de la izquierda es por variables relacionadas con características del agua, y de la derecha con las que tienen que ver con la concentración del mercurio.

En la figura 5 se observan las puntuaciones de las observaciones.

2.2.3 PCA y justificación de número de componentes.

Observese de la figura 6 que el cambio de curvatura se da en la componente 2, por lo que se considera que se deben de usar las 2 primeras componentes.

2.2.4 Interpretación y significancia de PCA

Se encontró que el PCA aplicado a este problema ayuda a reducir la dimensionalidad del modelo y por ende su complejidad, al también descartar variables de baja contribución como X8, y dividir en dos componentes de las que la primera esta relacionada con las variables relacionadas con mercurio y la segunda con aquellas características encontradas en el lago.

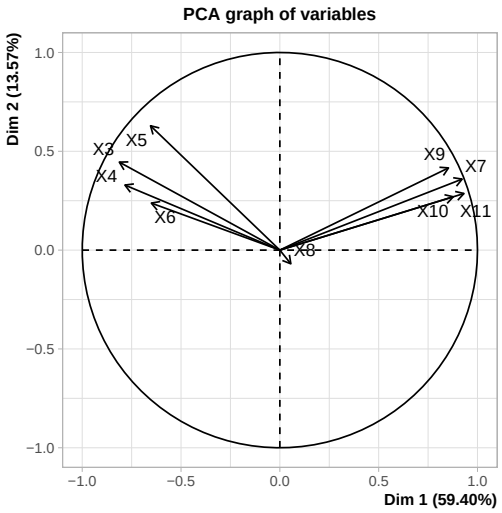

Figure 4: Aplicación de PCA a variables X3 a X11.s

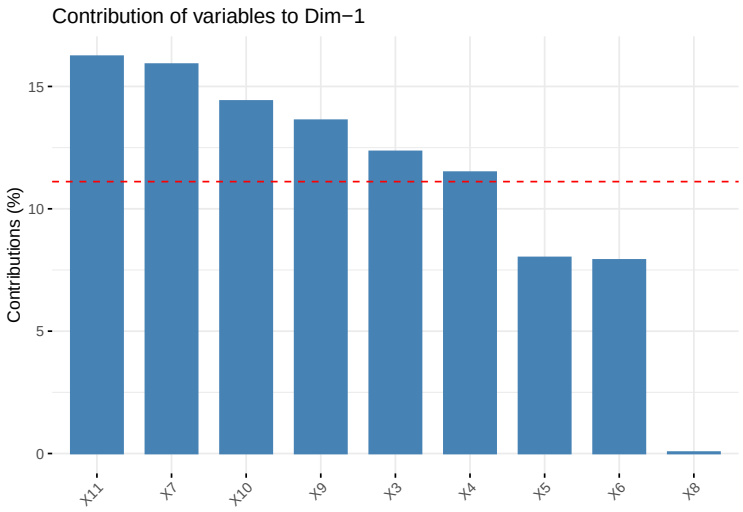

Figure 5: Puntuaciones de las observaciones.

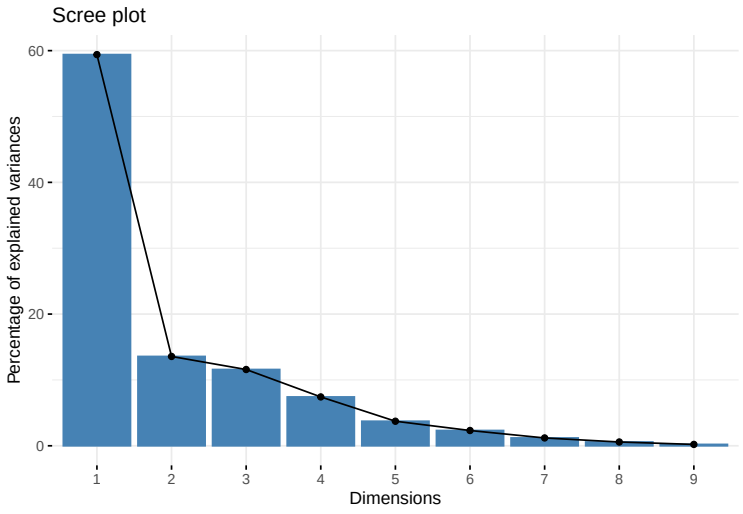

Figure 6: Aplicación de PCA a variables X3 a X11.

El PCA ayuda también a simplificar el análisis en comparación al anterior, para el cual se tuvo que buscar correlaciones entre variables y hacer pruebas para encontrar cuales son las que tienen correlación para despues buscar nivel de influencia, mientras que en este basto con hacer la separación en componentes.

3 Conclusión

Este análisis contribuye a responder a la pregunta del estudio al encontrar 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 el PH, obtenidos de la 5.

La normalidad encontrada en las variables de PH y concentración máxima de mercurio facilitan responder la pregunta sobre la variable con mayor influencia en la concentración por mercurio, puesto que al tener una distribución normal multivariada involucra que estas dos estan correlacionadas.

Los componentes principales ayudan a abordar este problema al agrupar diferentes variables en dos dimenciones para facilitar el análisis e interpretación del modelo.

Se logró reducir el modelo a dos componentes agrupadas por categoría divididas por la dimensión 1, así como descartar la variable X8 por su baja contribución, con lo cual, en comparación de la entrega anterior, agiliza el análisis al descartar variables que no necesariamente deben ser tomadas en cuenta.

Se considera óptimo que como próximas investigaciones se haga un análisis de causalidad de cada una de las variables de la componente 1 (alcalinidad, PH, etc.)

4 Anexos

Liga al código en GitHub https://github.com/1dgog/tc3007_portafoliodeimplementacion_m5/blob/main/peces_mercurio.Rmd

Referencias bibliográficas

Johnson, Richard A., and Dean W. Wichern. 2016. *Applied Multivariate Statistical Analysis.* Pearson. http://0-search.ebscohost.com.biblioteca-ils.tec.mx/login.aspx%3fdirect%3dtrue%26db%3dcat03431a%26AN%3dbdis.b1913723%26lang%3des%26site%3deds-live%26scope%3dsite.