UNIVERSITAT DE BARCELONA

Universitat Oberta de Catalunya

Prueba de evaluación continua 1
Estadística Multivariante

Francesc Carmona

6 de noviembre de 2019

Mercurio en los lagos de Florida

La contaminación por mercurio de los peces comestibles de agua dulce representa una amenaza directa para la salud humana. Esto hace que sea importante conocer los factores que influyen en el nivel de contaminación. Por ejemplo, ¿cómo afecta la química del agua en un lago a la concentración de mercurio en los peces que viven allí? Los investigadores Lange, Royals y Connor (1993) estudiaron este problema para la lubina negra (largemouth bass - *Micropterus salmoides*) que vive en 53 lagos de Florida diferentes.

Figura 1: Largemouth bass *Micropterus salmoides*

Los datos se hallan en el archivo `MERCBASS.TXT` que se encuentra en la web[1]
Electronic Encyclopedia of Statistical Examples & Exercises

En la base de datos hallaremos las siguientes variables:

ID number = número de idendificación de la observación

Lake = nombre del lago

Alkalinity (mg/l) = Carbonato cálcico

pH

Calcium (mg/l) = Calcio

Chlorophyll (mg/l) = Clorofila

Avg Mercury = Concentración media de mercurio (partes por millón) en el tejido muscular de los peces de la muestra de ese lago

samples = el número de peces de la muestra de ese lago

Min = Concentración media de mercurio cuando se hacen 0 los valores censurados

Max = Concentración media de mercurio cuando se hacen 0.04 los valores censurados

3 yr Standard Mercury = Estimación por regresión de la concentración de mercurio en peces de tres años de edad en el lago (o = Avg Mercury cuando no se sabe la edad)

Age data = 1 si se sabe la edad de los peces, 0 en otro caso

El aparato de medida de la concentración de mercurio puede detectar un nivel mínimo de 0.04 ppm. Un dato con valor real por debajo de este nivel se dice "censurado". Para calcular la media se fijó un valor de 0.02 ppm. Para tener en cuenta esta imprecisión, los autores proporcionaron también las variables `Min` y `Max` con distintos valores asignados a los datos censurados.

[1] `http://bcs.whfreeman.com/WebPub/Statistics/shared_resources/EESEE/MercuryinFloridasBass/index.html`

Ejercicio 1

(a) Preparar la base de datos `mercbass` a partir del archivo `MERCBASS.TXT`. Habrá que tener cuidado con los nombres de las columnas. Tal vez se deba utilizar un `read.delim()`. También habrá que nombrar correctamente las observaciones.

(b) El Estado de Florida ha fijado un nivel de concentración de mercurio en alimentos comestibles de 0.5 ppm por encima del cual se consideran no saludables. ¿Qué porcentaje de lagos en este estudio tienen un nivel superior de concentración de mercurio para considerarse no saludables? ¿Podemos considerar esta estimación como un buen estimador del porcentaje de lagos de Florida que superan el nivel declarado? ¿Porqué o porqué no?

(c) Como estamos interesados en la relación entre las 4 variables químicas y la concentración de mercurio, realizar diagramas de dispersión dos a dos. ¿Qué dificultades observamos?

(d) Consideremos transformar las variables originales mediante alguna de las potencias de la *escalera de Tukey* (Tukey ladder of powers). Nos centraremos en las variables químicas y la concentración media de mercurio. La función `transformTukey()` del paquete `rcompanion` nos puede servir.

Mostrar gráficamente que con las variables transformadas (si es necesario), mejora la relación lineal entre las variables químicas y la concentración de mercurio.

(e) Sean $X_1 = f(\texttt{Alkalinity})$, $X_2 = \texttt{pH}$, $X_3 = f(\texttt{Calcium})$, $X_4 = f(\texttt{Chlorophyll})$, $Y_1 = g(\texttt{Min})$, $Y_2 = g(\texttt{Avg.Mercury})$ y $Y_3 = g(\texttt{Max})$, donde f y g son las transformaciones propuestas en el apartado anterior.

Calcular la matriz de varianzas-covarianzas de las variables $(X_1, X_2, X_3, X_4, Y_1, Y_2, Y_3)$ con la matriz de centrado $\mathbf{H}$. Comprobar que da el mismo resultado que con la función `cov()`.

(f) Consideremos las variables $U = 0.4X_1 + 0.2X_2 + 0.2X_3 + 0.2X_4$ y $V = (Y_1 + Y_2 + Y_3)/3$. Calcular las medias y las varianzas de U y V en función de las medias, varianzas y covarianzas de las variables X_i y Y_j. También la correlación entre las variables U y V.

(g) Hallar las combinaciones lineales $U = a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4$ y $V = b_1Y_1 + b_2Y_2 + b_3Y_3$ de forma que la correlación sea máxima.

La solución pasa por hacer un Análisis de la correlación canónica[2] (ACC) que se puede resolver paso a paso con **R**. Para ello necesitamos los valores y vectores propios de las matrices:

$$\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}\mathbf{S}_{YY}^{-1}\mathbf{S}_{YX}$$

$$\mathbf{S}_{YY}^{-1}\mathbf{S}_{YX}\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}$$

donde $\mathbf{S}$ es la matriz de varianzas-covarianzas de las variables.

El primer valor propio de ambas matrices coincide y la correlación máxima es la raíz cuadrada de ese valor. Las combinaciones lineales que buscamos son los vectores propios de cada matriz asociados al primer valor propio común. Sin embargo, como exigimos que la combinación lineal $\mathbf{a}'\mathbf{x}$ verifique $\mathbf{a}'\mathbf{S}_{XX}\mathbf{a} = 1$ y que la combinación lineal $\mathbf{b}'\mathbf{y}$ verifique $\mathbf{b}'\mathbf{S}_{YY}\mathbf{b} = 1$, habrá que dividir cada vector propio por la raíz cuadrada de $\mathbf{u}'\mathbf{S}_{XX}\mathbf{u}$ y la raíz cuadrada de $\mathbf{v}'\mathbf{S}_{YY}\mathbf{v}$ respectivamente, donde $\mathbf{u}$ y $\mathbf{v}$ son los vectores propios asociados al primer valor propio de cada una de las matrices.

Dado que las variables químicas del agua de los lagos están en las mismas unidades, salvo el pH, ¿cual es la variable mejor correlacionada según el ACC con la concentración de mercurio? ¿Cual de las tres variables de concentración de mercurio presenta mayor correlación con las variables químicas según el ACC?

[2] `http://pensamientoestadistico.com/analisis-correlacion-canonica/`

(h) Comprobar con alguna función de **R** que el resultado del apartado anterior es correcto.

Nota: La función `cc()` del paquete `CCA` proporciona las correlaciones y los vectores canónicos correctos. La función `cancor()` da unos vectores canónicos que hay que multiplicar por $\sqrt{n-1}$, donde n es el tamaño de la muestra.

Ejercicio 2

En este ejercicio vamos a comparar conjuntamente las tres variables Y_j respecto al pH de los lagos según sean ácidos (`pH < 7`) o alcalinos (`pH > 7`). Los lagos con valor de pH neutro (`= 7`) no pertenecen a ninguno de los dos grupos y no deben intervenir en la comparación.

(a) Dibujar un único boxplot múltiple con las tres variables (Y_1, Y_2, Y_3) en la misma escala que comparen los valores en los dos grupos de lagos (ácidos y alcalinos).

(b) Estudiar gráficamente y con algún test la normalidad univariante y multivariante (si es posible) en cada uno de los grupos.

La normalidad multivariante se puede estudiar con la asimetría y la kurtosis multivariante de Mardia gracias a la función `mvn` del paquete `MVN`. La misma función permite el estudio univariante y también hace los gráficos.

Acompañar el test con un gráfico qq-plot de ajuste a la distribución ji-cuadrado de las distancias de Mahalanobis (clásicas y robustas) de los datos a la media en cada grupo.

Comentar el resultado en cuanto a los tests y en cuanto a la presencia de datos atípicos.

(c) Comparar las matrices de covarianzas de los grupos con el test M de Box. También podemos comparar la variabilidad con un test de Levene múltiple. ¿Ambos métodos contrastan lo mismo?

En caso de duda sobre la normalidad multivariante de los datos, proponer y calcular un método multivariante para contrastar si hay diferencias de variabilidad entre los grupos.

(d) Comparar de forma multivariante las medias de las variables (Y_1, Y_2, Y_3) en los dos grupos de lagos según su pH. Realizar un test suponiendo normalidad multivariante y otro de permutaciones.

Ejercicio 3

Con la base de datos y las variables del ejercicio 1(e).

(a) Calcular las componentes principales con alguna función específica de **R** a partir de la matriz de correlaciones para que las variables no tengan pesos distintos. ¿Qué relación tienen estas componentes con los vectores propios de la matriz de correlaciones $\mathbf{R}$? ¿Cuantas componentes parecen necesarias para tener una buena representación de los datos?

Aunque existe una plétora de métodos para contestar la última pregunta, los cuatro métodos más utilizados son:

1. El criterio de Kaiser: se eligen las primeras componentes cuyos valores propios son superiores a 1.
2. Varianza explicada: se eligen las primeras componentes que explican como mínimo el 70 o 80 % de la varianza total.
3. Scree plot de Cattel (1966): Éste es un método gráfico por el que se eligen las componentes hasta el codo del gráfico.

4. El criterio de la interpretabilidad: se trata de elegir todas las componentes que mejor recojan la esencia del significado de las variables y verificar que esa interpretación tiene sentido en los términos conocidos del problema bajo investigación.

Seguramente, la solución es una combinación de los cuatro criterios.

(b) Interpretar las dos primeras componentes mediante las variables mejor relacionadas con cada una de ellas. Los gráficos de correlaciones con las variables originales pueden ayudar.

(c) Los investigadores Canfield and Hoyer (1988) hallaron que el pH y la alcalinidad en los lagos de Florida generalmente crecen si se va del Noroeste al Sudeste y de las tierras altas a la costa del estado.

Representar los grupos en un gráfico de dos dimensiones con la función `PCA` del paquete `FactoMineR`[3].

¿Donde están situados en el gráfico los lagos del Noroeste? ¿Se ajusta este resultado con la hipótesis de los investigadores?

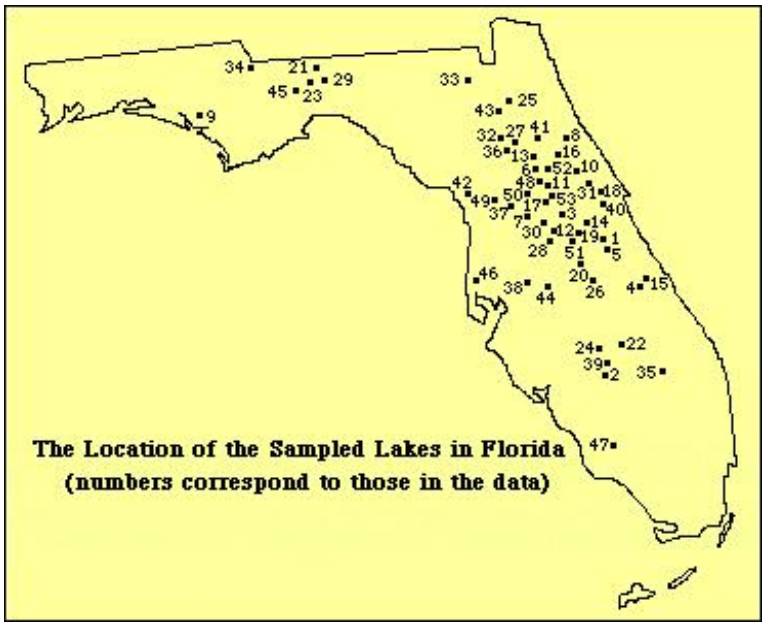

Figura 2: The Location of the Sampled Lakes in Florida

(d) Dado que las variables estudiadas en los apartados anteriores de este ejercicio están claramente en dos grupos, podemos realizar un Análisis factorial múltiple[4].

¿Difiere substancialmente el resultado del MFA del obtenido con las componentes principales?

[3] `http://factominer.free.fr/factomethods/principal-components-analysis.html`
[4] `http://factominer.free.fr/factomethods/multiple-factor-analysis.html`

Referencias

[1] D.E. Canfield and M.V. Hoyer (1988), Regional geology and the chemical and trophic state characteristics of Florida lakes, Lake Reservoir Mgmt. 4: 21-31.

[2] T.R. Lange, H.E. Royals and L.L. Connor (1993), Influence of Water Chemistry on Mercury Concentration in Largemouth Bass from Florida Lakes. *Transactions of the American Fisheries Society*, Vol. 122-1, pp. 74-84.

[3] Bryan F.J. Manly and Jorge A. Navarro Alberto (2017), *Multivariate Statistical Methods: A Primer*, Fourth Edition. Chapman and Hall/CRC. Springer-Verlag.