Análisis Multivariado Aplicado a Ciencias Ambientales usando R

Actividad 3

Prof. Edlin Guerra Castro

June 20, 2024

Pruebas de Hipótesis Multivariadas

Parte I: Caso Ekofisk macrofauna y contaminantes en sedimento: MANOVA, ANOSIM, PERMANOVA y PERMDISP

Continuamos con el estudio publicado por Gray JS, Clarke KR, Warwick RM, Hobbs G (1990). Detection of initial effects of pollution on marine benthos: an example from the Ekofisk and Eldfisk oilfields, North Sea. Marine Ecology Progress Series 66, 285-299.

A diferencia de la actividad 3, en que se exploraron métodos de ordenación no restringidos para representar gráficamente el potencial efecto de las distancias a la plataforma sobre las características fisicoquímicas del sedimento y la macrofauna, en esta actividad vamos a formalmente evaluar la hipótesis de gradiente de contaminación y cambios en la macrofauna. En este sentido, evaluaremos dos hipótesis nulas específicas:

1. La posición de los centroides de cada grupo en el espacio multivariado es la misma.
2. La dispersión multivariada alrededor de los centroides de cada grupo es la misma.

Para someter a prueba la primera hipótesis aplicaremos las tres alternativas de pruebas disponibles en *R*: MANOVA (`manova{stats}`), ANOSIM (`anosim{vegan}`) y PERMANOVA (`adonis2{vegan}`). La segunda hipótesis la evaluaremos con PERMDISP (`betadisper`{vegan}).

MANOVA

a. Con fines *demostrativos*, tratemos de aplicar un *MANOVA* a los datos de macrofauna. Importe el archivo "macrofauna.cvs" y renómbrela `datos`. Al igual que en la actividad anterior, divida a datos en una matriz con la información de las especies (use la función `as.matrix` y nombre macrofauna) y con la información de las distancias a la plataforma en forma de factor (explícitamente conviértelo en variable explicativa con la función `factor`).

b. Para iniciar un MANOVA, se requiere estimar sumatorias cuadráticas de productos cruzados (**SSCP**) usando operaciones matriciales en que se multiplica la matriz de residuos (Z) con su transpuesta, y esto es viable siempre que n sea mayor a p. Intente multiplicar la matriz macrofauna con su transpuesta. ¿qué resultó? intente ahora reduciendo las columnas a 40, 39 y 38 (use las primeras) ¿qué resultó en cada intento?.

c. Acabamos de comprobar que no es posible calcular la **SSCP** a un juego de datos que tenga $p > n$. Nuevamente, con fines demostrativos seleccionemos las 38 especies más abundantes para poder aplicar un MANOVA. Use para ello estas líneas de comandos.

```
library(tidyr)
library(dplyr)

# Identifiquemos a las especies más abundantes y
# ordenemos de mayor a menor
N <- datos|>
  pivot_longer(cols = 3:175, names_to = "especies", values_to = "abundancia")|>
  group_by(especies)|>
  summarise(abundancia = mean(abundancia))|>
  arrange(-abundancia)

# Seleccionemos las 20 especies más abundantes
spp <- N$especies[1:20]

# Prepare la matriz reducida para el análisis MANOVA

macrofauna20 <- as.matrix(macrofauna[,spp])

# Ahora reduzcamos las diferencias en abundancias con raíz
# cuadrada y corra la funcion manova. Llame a al resultado mod1,
# y aplique la función summary a mod1

r.c.20 <- sqrt(macrofauna20)
```

```r
# Modelo lineal multivariado
mod1 <- manova(r.c.20 ~ distancia)

# Tabla MANOVA
summary(mod1)
```

d. Ahora desarrolle una ordenación de PCA para ser coherentes con el uso de MANOVA. Use las funciones `prcomp` y los paquetes `ggplot2` y `ggfortify`. Analice y describa qué ocurren en cada línea de código. ¿cómo interpreta el resultado? ¿Considera que es una buena ordenación? Considerando el resultado del MANOVA y este PCA ¿qué puede concluir sobre la hipótesis nula? ¿Es este análisis representativo?

```r
library(ggplot2)
library(ggfortify)

pca <- prcomp(r.c.20,
                  retx = T,
                  center = TRUE,
                  scale. = FALSE)

macrofauna20 <- macrofauna20 |>
  as.data.frame()|>
  mutate(distancia = factor(distancia,
                            levels = c("1", "2", "3", "4"),
                            labels = c("<250m", "<1.0 km", "< 3.5 km","> 3.5 km")))

autoplot(pca, data = macrofauna20, colour = 'distancia',
         loadings = TRUE, loadings.colour = 'blue',
         loadings.label = TRUE, loadings.label.size = 3)
```

ANOSIM

a. Apliquemos el equivalente no paramétrico de análisis multivariado a la hipótesis nula con la función `anosim`. Para ello, calcule el índice de similitud Bray-Curtis a macrofauna (considerando todas las 173 especies) y someta a aprueba la hipótesis de cambio en macrofauna según la distancia a la plataforma ¿cómo se puede interpretar el resultado? Qué implica el valor de R global.

b. Identifique las especies que mejor reflejen las diferencias entre los grupos de distancia 1 y 4 con `simper`. ¿cuáles especies contribuyen consistentemente en las diferencias? ¿son las mismas elegidas para aplicar el MANOVA?

c. Genere el nMDS (MDS no métrico) con `metaMDS` y represente los grupos de distancia usando el paquete `ggplot2`. ¿cómo interpreta el resultado? ¿Considera que es una buena ordenación? Considerando el resultado del ANOSIM y este nMDS ¿qué puede concluir sobre la hipótesis nula? ¿Es este análisis representativo?

PERMANOVA y PERMDISP

a. Ahora someta a aprueba la hipótesis nula de diferencias en la posición de los centroides según la distancia usando `adonis2`. ¿cómo se puede interpretar el resultado? Interprete el valor de *pseudo-F* y la probabilidad de pertenecer a una hipótesis nula de igual centroide para todos los grupos.

```
adonis2(bray~distancia)
```

b. Genere nuevamente un PCO para las muestras. Este gráfico es el conceptualmente más adecuado para representar resultados de un PERMANOVA, no obstante, refleja poca variación. ¿Qué gráfico alternativo al PCO se pueden construir y que preserve las relaciones métricas entre los grupos? ¿qué ventajas ofrece su elección? Demuestre que el gráfico propuesto tendrá mejor calidad que un PCO.

c. Someta a prueba la hipótesis nula de igual dispersión multivariada según el grupo de distancias usando `betadisper` y `permutest`. ¿cómo puede interpretar ecológicamente el resultado? ¿qué implicaciones tiene este resultado en la interpretación del MANOVA, ANOSIM y PERMANOVA que ejecutó previamente?

d. Reconociemdo el efecto que tiene la dispersión multivariada sobre las pruebas estadísticas, aplique la versión modificada a PERMANOVA con los códigos indicados abajo.

```
## Funciones preliminares

# Sum of Squares using Theorem of Huygen
SS <- function (d) {
  n <- dim(as.matrix(d))[1]
  ss <- sum(d ^ 2) / n
  return(ss)
}

# Multivariate Dispersion
v = function (d) {
  n <- dim(as.matrix(d))[1]
  ss <- sum(d ^ 2) / n
  v <- ss / (n - 1)
  return(v)
```

```r
}

# Modified Pseudo-F (Anderson et al., 2017)
pseudo.F <- function (x, factor, distancia) {
  d <- vegdist(x, method = distancia)
  TSS <- SS(d)
  group <- as.data.frame(factor)
  x$grp <- factor
  factor <- as.factor(factor)
  lab <- names(table(group))
  lev <- table(group)
  CR <- c(1:nlevels(factor))
  for (i in 1:nlevels(factor)) {
    CR[i] <-
      SS(vegdist(x[x$grp == lab[i], 1:length(x) - 1], method = distancia))
  }
  RSS <- sum(CR)
  Var <- c(1:nlevels(factor))
  d.res <-
    as.data.frame(matrix(nrow = length(levels(factor)), ncol = 3))
  for (i in 1:nlevels(factor)) {
    Var[i] <-
      v(vegdist(x[x$grp == lab[i], 1:length(x) - 1], method = distancia))
    d.res[i, ] <- c(lev[i],
                    Var[i],
                    (1 - (lev[i] / sum(lev))) * Var[i])
  }
  den <- sum(d.res$V3)
  ASS <- TSS - RSS
  Fobs <- ASS / den
  return(Fobs)
}

## PERMANOVA2

PERMANOVA2 <- function(x, factor, distancia, nperm = 999) {
  control <- how(nperm = nperm, within = Within(type = "free"))
  Fobs <- pseudo.F(x, factor, distancia = distancia)
  Nobs <- nobs(x)
  F.permu <- numeric(length = control$nperm) + 1
  F.permu[1] <- Fobs
  ## Generation of pseudo.F values for HO using permutations without replacement
```

```
  for (i in seq_along(F.permu)) {
    ## return a permutation
    want <- permute(i, Nobs, control)
    ## calculate permuted F
    F.permu[i + 1] <-
      pseudo.F(x[want, ], factor, distancia = distancia)
  }
  ## probability for Fobs
  pval <- sum(abs(F.permu) >= abs(F.permu[1])) / (control$nperm + 1)
  ## Results
  return(data.frame("Pseudo-F" = F.permu[1], "p(perm)" = pval))
}## Permutation test based on Anderson et al. (2017)


# Aplicación de PERMANOVA modificado
PERMANOVA2(x = sqrt(macrofauna), factor = distancia, distancia = "bray")
```

e. ¿Qué puede concluir de todos estos análisis?¿obtuvo resultados contradictorios entre las pruebas respecto a la hipótesis nula? ¿pudo detectar diferencias en los tamaños de efecto evaluados?

Parte II: Desempeño de pruebas estadísticas en presencia de diferencias de dispersión.

Para evaluar el desempeño de las pruebas estadísticas multivariadas más populares usaremos simulación de datos garantizando igualdad en la posición de sus centroides grupales pero generando diferencias en dispersión. Luego someteremos a prueba la hipótesis nula (cierta) de igual posición de los centroides usando las pruebas ANOSIM (`anosim` de `vegan`), PERMANOVA (`adonis2` de `vegan`) y una versión modificada de PERMANOVA. Lo primero que haremos será simular dos juegos de datos (30 variables cada uno) usando los siguientes comandos:

```
# Parámetros para simular las 30 variables
set.seed(42)
ave <- runif(30, min = 10, max = 40)
var <- sqrt(ave)

# Matriz para juego de datos 1, 20 observaciones
sim1 <- matrix(data = NA, nrow = 20, ncol = 30)
```

```r
# Matriz para juego de datos 2, 10 observaciones
sim2 <- matrix(data = NA, nrow = 10, ncol = 30)

# Bucle para llenar matrices, noten como se duplica la varianza
# en el juego de datos 2
for (j in 1:30){
    sim1[,j] <- rnorm(20, mean = ave[j], sd=var[j])
    sim2[,j] <- rnorm(10, mean = ave[j], sd=2*var[j])
    }

# Convertimos ambas matrices en data.frames, y se combinan
sim1 <- as.data.frame(sim1)
sim2 <- as.data.frame(sim2)
sim <- rbind(sim1, sim2)

# Generamos un vector que reconoce ambos grupos
grupos <- c(rep("a",20), rep("b",10))
```

El objeto de nombre **sim** alberga los dos juegos de datos. Podemos verificar las propiedades multivariadas de este juego de datos proyectando los 30 objetos en un MDS no métrico, y luego podemos medir la dispersíón. Esto se logra con los siguientes códigos:

```r
# centramos todas las variables con una normalización
sim.n <- decostand(sim, method = "standardize")

#Estimamos las distancias Euclidianas
euc <- dist(sim.n)

#Generamos un MDS con vegan y graficamos con ggplot2
mds1<-metaMDS(euc)
MDS1 <- as.data.frame(mds1$points)
MDS1$grp <- grupos

plot.MDS1 <- ggplot(data=MDS1 ,aes(x=MDS1, y=MDS2))+
  geom_point(aes(colour=grupos), size=3.5)+
  theme_bw(base_size=16)

plot.MDS1
```

Con base en este gráfico, responda las siguientes preguntas:

a. ¿Puede apreciar las diferencias multivariadas en dispersión entre ambos grupos?

b. ¿Puede apreciar las diferencias (de haber) en la posición de los centroides de ambos grupos?

c. ¿Cómo puede confirmar que ambas diferencias en dispersión son poco probables de ocurrir por simple azar?

Para responder la útima pregunta se sugiere aplicar una prueba estadística para dispersión multivariada. Una de las más exitosas es **PERMDISP**, que en `vegan` está disponible como la función `betadisper` y `permutest`. Le recomiendo aplicar estas funciones para responder la pregunta. Verificada las diferencias en dispersión, lo que sigue es someter a prueba la hipótesis nula de que ambos grupos tienen el mismo centroide. Recuerde que la hipótesis nula es cierta, por lo que cualquier resultado significativo será un error inferenical del tipo I.

Para esto use `anosim`, `adonis` y `PERMANOVA2`. Para efectos de este ejercicio, las pruebas correrán así:

```
anosim(x = euc, grouping = grupos)

adonis2(euc ~ grupos)

PERMANOVA2(x = sim.n,
  factor = grupos,
  distancia = "euclidean",
  nperm = 999
)
```

d. Con base en los resultados, interprete los resultados de **ANOSIM**.

e. ¿Qué sugiere el resultado de **PERMANOVA**?

f. En ambos casos ¿cuál fue la probabilidad del estadístico en pertenecer a la hipótesis nula?

g. ¿qué indicó PERMANOVA2, modificado por Anderson et al 2017?

h. ¿Cuál de las tres pruebas le permitió respaldar lo que se aprecia en el gráfico MDS arriba generado?