IMM con R. Lab7: GLM Multivariados (MANOVA un factor con 2 niveles)

Dra. Maite Mascaro y Dr. Edlin Guerra Castro

21/05/2021

Caso Morfometría de embriones del pulpo *Octupus maya*

Parte I: Aplicación de un MANOVA con funciones "enlatadas"

Un grupo de investigadores desea explorar los cambios en 4 variables morfométricas de los embriones del pulpo rojo *Octopus maya*, cuando los embriones son expuestos a un incremento gradual de la temperatura y compararlos con una condición control en la que los embriones se mantienen a una temperatura constante de 26°C. Las variables morfométricas registradas en los embriones de ambos tratamientos fueron longitud total, longitud del manto, longitud del brazo y diámetro del ojo (todas en mm). Esperaban encontrar diferencias en la forma de los embriones del tratamiento control y expuestos. Los datos se encuentran en el archivo `morfOmay.xlsx`.

1. Identifica las variables que forman parte de la respuesta, y las que son explicativas. ¿Cuál sería el modelo de ANOVA idóneo para poner a prueba la hipótesis de los investigadores (i.e. ANOVA de una vía, anidado, factorial, etc.)? Formula la hipótesis del modelo e identifica el componente de variación que debería resultar significativo si ésta llegara a ser corroborada. (PISTA: considera el problema una extensión natural de una prueba univariada con una sola respuesta).

2. Genera una matriz Y con los descriptores que constituyen la respuesta. Asegúrate que es una matriz con la función `as.matrix`, y explorala para responder a las siguientes preguntas. (PISTA: recuerda la función `pairs`, `boxplot`).

a) ¿Cuántas observaciones (nn) y descriptores (pp) tiene la matriz Y?
b) ¿Existen correlaciones entre las variables morfométricas? ¿Son lineales?
c) ¿Cómo son las escalas de magnitud en las varaibles de respuesta?
d) Si son muy diferentes entre si, estandariza la matriz Y mediante la función `decostand` de la librería `vegan`. Repite la exploración para asegurar que la estadarización surtió efecto.
e) ¿Cuántas observaciones y niveles tiene la variable explicativa? ¿Es una varaible categórica o continua? Si es categórica, asegúrate de convertirla a factor antes de comenzar.

3. Para explorar la posibilidad de que se corrobore la hipótesis, aplica un PCA de covarianza a la matriz Y.std, y elabora un gráfico con los primeros dos componentes.

a) La función `aggregate` permite aplicar ciertos algoritmos sencillos a juegos de datos de acuerdo a un codificador. Copia el siguiente código para calcular las medias de cada descriptor por cada nivel del factor `tratamiento`

```
aggregate(Y.std,by=list(dat$trat), mean)
```

b) Con lo observado hasta ahora ¿Crees que la hipótesis de los investigadores se cumple?

4. Aplica un ajuste lm a la matriz Y.std en función de tratamiento, y guárdalo como mod1. Aplica la función `summary` para ver la salida.

a) ¿Qué es lo que resulta? ¿Cuántos coeficientes estima en total?
b) Ahora aplica la función `manova` a la matriz de datos multivariados (Y.std) en función de trat, y pide un resumen. ¿Qué es lo que resulta cuando pides el resumen? ¿Puedes identificar los distintos elementos de la salida?

```
mod2 <- manova(Y.std~dat$trat)


# Columna 1 son los g.l. de X que en este caso en un factor;
# Columna 2 el estadístico de Pillai que es la traza de la matriz (H (E+H)^-1);
# Columna 3 es la aproximación a F de dicho estadístico; entre más grande
# más significativo

# Columna 4 son los g.l. del numerador: número de coeficientes estimados sin contar
# con los interceptos(gpos x pp)-pp=(2x4)-4

# Columna 5 son los g.l. del denominador nn-ss-pp, donde nn es el numero de observaciones
# y ss el número de variables X en el modelo reducido (en este caso 1, porque hay 2 grupos)

# Columna 6 es el valor de p asociado al estadistico F.
```

c) ¿Cómo interpretarías este resultado en términos estadístico? ¿Cómo lo interpretarías en los términos de la hipótesis del problema?
d) Cambia los argumentos de la función `summary` como se indica a continuación. Explica los cambios en la nueva salida.

5. Para explorar algunos de los resultados de mod2 sigue las instrucciones a continuación. Con tu conocimiento sobre los Modelos Generales Lineales responde a las preguntas en cada inciso.

a) Aplica la función `coef` sobre mod2. ¿Cuántos valores son y qué representan? ¿Con qué elementos de la salida de un modelo univariado se corresponden?
b) Obtén las matrices SSCP de la X y de los residuales copiando el siguiente código. ¿Qué son los elementos de la diagonal en cada caso?

```
summary(mod2)$SS
```

c) Llama a mod2 y examina lo que se obtiene. ¿Reconoces algún valor de esta tabla? ¿Qué representa la suma de todos los valores de la tabla (excluyendo los g.l.)?
d) Llama `mod2$residuals` e identifica los elementos de esta matriz. ¿Cómo deberían de proyectarse estos valores en una ordenación por PCA? Verifica tu respuesta usando el código de la librería `pca3d`.
e) Llama mod2$fitted e identifca los elementos de esta matriz. ¿Cómo deberían proyectarse estos valores en una ordenación por PCA? Verifica tu respuesta.
f) Llama mod2$model e identifica los elementos de esta matriz. ¿En qué difiere la matriz "model" de la "fitted"?

Parte II: Aplicación del MANOVA desde el algebra matricial

1. El siguiente código aplica un MANOVA a mano, es decir, siguiendo las fórmulas de álgebra matricial que vimos en clase. Sigue las instrucciones para ir haciendo la correspondencia entre la salida de la función `manova` y las distintas matrices de los resultados del análisis de GLM.

```
# Redefine la matriz XX para que sea binaria y tenga un vector de `1`.
dat$ntrat <- as.numeric(dat$trat)
dat[dat$ntrat == 1, 6] <- 0
dat[dat$ntrat == 2, 6] <- 1
XX <- cbind(rep(1, 119), dat[, 6])
```

```r
# Redefinir matrices YY y XX iguales a Y.std y X
YY <- Y.std
# Resolver para Beta: obtener los coeficientes Beta0 y Beta1 para cada variable respuesta.
BETA <- solve(t(XX) %*% XX) %*% t(XX) %*% YY
BETA

# Compara con los valores obtenidos con el procedimiento enlatado.
coef(mod2)

# Obtener valores de Y predichos por el modelo: Y.hat = XB
YY.gor <- XX %*% BETA
head(YY.gor)

# Compara con los valores obtenidos con el procedimiento enlatado.
head(as.matrix(predict(manova(mod2))))

# Obtener los SSresidual: elevar al cuadrado la diferrencia de YY-YY.hat
SSCP.res <- t(YY - YY.gor) %*% (YY - YY.gor)
head(SSCP.res)

# Compara con los valores obtenidos con el procedimiento enlatado.
head(summary(mod2)$SS$Residuals)

# Obtener SSCPtotal: obtener el vector fila de medias
colMeans(YY)

# vector columnas de unos
ones <- matrix(1, nrow = nrow(YY), ncol = 1)
dim(ones)

# Obtener matriz de medias YYbarra
YY.bar <- ones %*% colMeans(YY)

# Obtener SStotal: elevar al cuadrado la diferencia YY-YY.bar
SSCP.tot <- t(YY - YY.bar) %*% (YY - YY.bar)
head(SSCP.tot)

# Obtener SSCP modelo
SSCP.mod <- SSCP.tot - SSCP.res
head(SSCP.mod)

# Compara con los valores obtenidos con el procedimiento enlatado.
head(summary(mod2)$SS$`dat$trat`)

# Cálculo de lambda de Wilks
lambda <-
  det(SSCP.res) / det(SSCP.mod + SSCP.res)
lambda

# Compara con los valores obtenidos con el procedimiento enlatado.
summary(mod2, intercept = T, test = "Wilks")

# número de observaciones (filas)
```

```r
nn <- 119

# número de variables X en el modelo reducido
# (aquí no hay variables X, solo intercepto es 1)
ss <- 1

# número de variables Y
pp <- 4

# Calcula F a partir de lambda
F.lamb <- ((1 - lambda) / lambda) * ((nn - ss - pp) / pp)

# Compara con el valor obtenido con el procedimiento enlatado.
F.lamb
summary(mod2, intercept = T, test = "Wilks")

# Obtiene la probabilidad asociada al valor de F bajo la Ho.
pf(F.lamb, pp, (nn - ss - pp), lower.tail = FALSE)
```

2. El siguiente código es para obtener otras visualización del mod2 usando PCA.

a) Sobre los objetos obtenidos con el código a mano; sin vectores; sin texto.

```r
pca.YY <-
  prcomp(
    YY,
    retx = TRUE,
    center = TRUE,
    scale. = TRUE,
    tol = NULL
  )
pca.YY.scores <- pca.YY$x
plot(
  pca.YY.scores,
  asp = 1,
  pch = 21,
  col = "black",
  bg = ifelse(XX[, 2] == "0", "blue", "red")
)
```

b) El biplot enlatado; tiene el problema de no aceptar colores distintos para cada punto.

```r
biplot(pca.YY, scale=1, cex=.7)
```

c) El biplot "fancy" de la librería `pca3d`;

```r
library(pca3d)
pca2d(pca.YY, group=dat$trat, fancy=T, biplot=T)
```

d) Un biplot hecho a mano, donde puedes cambiar cualquier rasgo.

```r
lambda <- pca.YY$sdev * sqrt(nrow(pca.YY$x))
plot (
  t(t(pca.YY$x) / lambda),
  col = "black",
  asp = 1,
  pch = 21,
```

```
  bg = ifelse(XX[, 2] == "0", "blue", "red")
)
text (
  t(t(pca.YY$x) / lambda),
  rownames(dat),
  col = "grey40",
  pos = 4,
  cex = .5
)
par (new = T)
Rot <- t(t(pca.YY$rotation) * lambda)
XLIM <- c(-max(abs(Rot[, 1])), max(abs(Rot[, 1])))
XLIM <- XLIM + (XLIM * 0.1)
plot(
  Rot,
  col = 4,
  axes = FALSE,
  xlim = XLIM,
  ylim = XLIM,
  pch = ""
)
#flechas de los vectores
arrows (rep(0, nrow(pca.YY$rotation)),
        rep(0, nrow(pca.YY$rotation)),
        Rot[, 1], Rot[, 2], col = 4)
#texto de los vectores
text (Rot[, 1:2], rownames(Rot), col = 4, pos = 4)
# ejes para los vectores
axis (3)
axis (4)
#linea vertical que pasa en el zero
abline(v = 0, lty = 3, col = "grey")
# linea horizontal que pasa en el zero
abline(h = 0, lty = 3, col = "grey")
```

3. ¿Puedes detectar las diferencias estadísticas (significancia de la prueba de MANOVA) entre los embriones el tratamiento control y rampa en las distintas representaciones gráficas? Usando toda la información obtenida (eigenvalores, eigenvectores, MANOVA, residuales, etc.) ¿cómo respondes a la pregunta de los investigadores?

Parte III: Modelo de dos factores en MANOVA

Imagina que los investigadores tenían la hipótesis de que las diferencias en la morfometría de los embriones control y expuestos se incrementaran a lo largo del tiempo, justamente como parte del efecto de la temperatura. Con los mismos datos ajusta un GLM multivariado considerando un arreglo factorial con tratamiento (2 niveles) y días (7 niveles). Recuerda que en la estructura fija del modelo, el "full factorial" puede escribirse A*B, o bien, A+B+A:B. Antes de ajustar el modelo asegúrate de que todas las variables explicativas sean factores.

1. ¿Se corrobora la hipótesis de los investigadores? Explica tu respuesta.
2. Obtén los residuales del modelo nuevo, y su proyección en una ordenación por PCA usando la librería `pca3d`. ¿Desapareció el patrón observado anteriormente? ¿Cuáles serían los mejores estimadores de la variación residual para cada descriptor o variable morfométrica medida?
3. Obtén los valores gorro o ajustados, y copia el siguiente código para obtener su proyección en una

ordenación por PCA usando la libreria `pca3d`.

```
mod3$fitted -> Fmat3
pcaFmat3 <- prcomp(Fmat3, scale. = TRUE)
pca2d(
  pcaFmat3,
  group = paste(dat$trat, dat$dia),
  fancy = T,
  biplot = T
)
```