Análisis Canónico

Karina Itzel Rodríguez Conde

26/5/2022

ANÁLISIS CANÓNICO

Introducción

El Análisis Canónico es una técnica multivariante que cuantifica la validez de la relación entre dos conjuntos de variables, sea dependiente o independiente y tiene como objetivo principal el determinar si estos dos conjuntos de variables son independientes uno de otro.

Matriz de datos

Se trabajó con la matriz **penguins** la cual está precargada en R y contiene la información de tres diferentes especies de pingüinos a partir de información anatómica: largo del pico, grosor del pico, largo de la aleta y masa corporal,

Cada pingüino es de una de las tres especies siguientes: Adelie, Gentoo y Chinstrap.

Importar la matriz de datos

```
library(readxl)
```

```
library(readxl)
penguins <- read_excel("C:/Users/TOSHIBA/Downloads/penguins.xlsx")
View(penguins)
```

Exploración de la matriz

1.- Dimensión de la matriz:

```
dim(penguins)
```

```
## [1] 344   9
```

2.- Nombres de las columnas:

```r
colnames(penguins)
```

```
## [1] "ID"              "especie"         "isla"            "largo_pico_mm"
## [5] "grosor_pico_mm"  "largo_aleta_mm"  "masa_corporal_g" "genero"
## [9] "año"
```

3.- Tipo de variables:

```r
str(penguins)
```

```
## tibble [344 x 9] (S3: tbl_df/tbl/data.frame)
##  $ ID             : chr [1:344] "i1" "i2" "i3" "i4" ...
##  $ especie        : chr [1:344] "Adelie" "Adelie" "Adelie" "Adelie" ...
##  $ isla           : chr [1:344] "Torgersen" "Torgersen" "Torgersen" "Torgersen" ...
##  $ largo_pico_mm  : num [1:344] 39.1 39.5 40.3 37.8 36.7 39.3 38.9 39.2 34.1 42 ...
##  $ grosor_pico_mm : num [1:344] 18.7 17.4 18 18.1 19.3 20.6 17.8 19.6 18.1 20.2 ...
##  $ largo_aleta_mm : num [1:344] 181 186 195 190 193 190 181 195 193 190 ...
##  $ masa_corporal_g: num [1:344] 3750 3800 3250 3700 3450 ...
##  $ genero         : chr [1:344] "male" "female" "female" "female" ...
##  $ año            : num [1:344] 2007 2007 2007 2007 2007 ...
```

4.- Saber si existen datos nulos:

```r
anyNA(penguins)
```

```
## [1] FALSE
```

Esta base de datos no contiene datos nulos.

Escalamiento de la matriz

Paqueterías a utilizar

```r
library(tidyverse)
```

1.- Generación de variables X

```r
X <- penguins %>%
  select(grosor_pico_mm, largo_pico_mm) %>%
  scale()
head(X)
```

```
##      grosor_pico_mm largo_pico_mm
## [1,]      0.7863145    -0.8825216
## [2,]      0.1267012    -0.8093460
## [3,]      0.4311381    -0.6629947
## [4,]      0.4818776    -1.1203424
## [5,]      1.0907514    -1.3215754
## [6,]      1.7503647    -0.8459338
```

2.- Generación de variables Y

```
Y <- penguins %>%
  select(largo_aleta_mm,masa_corporal_g) %>%
  scale()
head(Y)
```

```
##      largo_aleta_mm masa_corporal_g
## [1,]     -1.4166210      -0.5646829
## [2,]     -1.0614850      -0.5022529
## [3,]     -0.4222402      -1.1889828
## [4,]     -0.7773762      -0.6271129
## [5,]     -0.5642946      -0.9392628
## [6,]     -0.7773762      -0.6895429
```

Análisis canónico con un par de variables

Librería y paquetería a utilizar

```
library(CCA)
```

1.- Análisis

```
ac<-cancor(X,Y)
ac
```

```
## $cor
## [1] 0.79268475 0.09867305
##
## $xcoef
##                       [,1]       [,2]
## grosor_pico_mm  0.03098538 0.04615243
## largo_pico_mm  -0.03746177 0.04107014
##
## $ycoef
##                         [,1]       [,2]
## largo_aleta_mm  -0.055220261 -0.0951545
## masa_corporal_g  0.001411466  0.1100076
##
## $xcenter
## grosor_pico_mm  largo_pico_mm
##   4.412451e-16   2.713430e-16
##
## $ycenter
##  largo_aleta_mm masa_corporal_g
##   -8.958914e-16   -1.045272e-16
```

2.- Visualización de la matriz X

```
ac$xcoef
```

```
##                       [,1]       [,2]
## grosor_pico_mm  0.03098538 0.04615243
## largo_pico_mm  -0.03746177 0.04107014
```

3.- Visualización de la matriz Y

```
ac$ycoef
```

```
##                         [,1]       [,2]
## largo_aleta_mm  -0.055220261 -0.0951545
## masa_corporal_g  0.001411466  0.1100076
```

4.- Visualización de la correlación canónica

```
ac$cor
```

```
## [1] 0.79268475 0.09867305
```

5.- Obtención de la matriz de variables canónicas

Se obtiene multiplicando los coeficientes por cada una de las variables (X1 y Y1)

```
ac1_X <- as.matrix(X) %*% ac$xcoef[, 1]
ac1_Y <- as.matrix(Y) %*% ac$ycoef[, 1]
```

6.- Visualización de los primeros 20 datos

```
ac1_X[1:20,]
```

```
##  [1] 0.05742508 0.03424542 0.03819593 0.05690117 0.08330590 0.08592589
##  [7] 0.04464608 0.07088939 0.08225809 0.06113346 0.04117935 0.04432371
## [13] 0.02642463 0.10015624 0.12599695 0.06040849 0.06488291 0.06556776
## [19] 0.08491867 0.05415894
```

```
ac1_Y[1:20,]
```

```
##  [1] 0.07742915 0.05790657 0.02163800 0.04204177 0.02983476 0.04195365
##  [7] 0.07720886 0.02414936 0.02987882 0.04301106 0.05702539 0.08126317
## [13] 0.07253771 0.03829586 0.01189829 0.06165247 0.02199048 0.01599667
## [19] 0.06491373 0.02723438
```

7.- Correlación canónica entre variable X1 y Y1

```
cor(ac1_X,ac1_Y)
```

```
##           [,1]
## [1,] 0.7926848
```

8.- Verificación de la correlación canónica

```
assertthat::are_equal(ac$cor[1],
                      cor(ac1_X,ac1_Y)[1])
```

```
## [1] TRUE
```

Análisis canónico con dos pares de variables

1.- Cálculo de las variables X2 y Y2

```
ac2_X <- as.matrix(X) %*% ac$xcoef[, 2]
ac2_Y <- as.matrix(Y) %*% ac$ycoef[, 2]
```

2.- Agregamos las variables generadas a la matriz original de penguins

```
ac_df <- penguins %>%
  mutate(ac1_X=ac1_X,
         ac1_Y=ac1_Y,
         ac2_X=ac2_X,
         ac2_Y=ac2_Y)
```

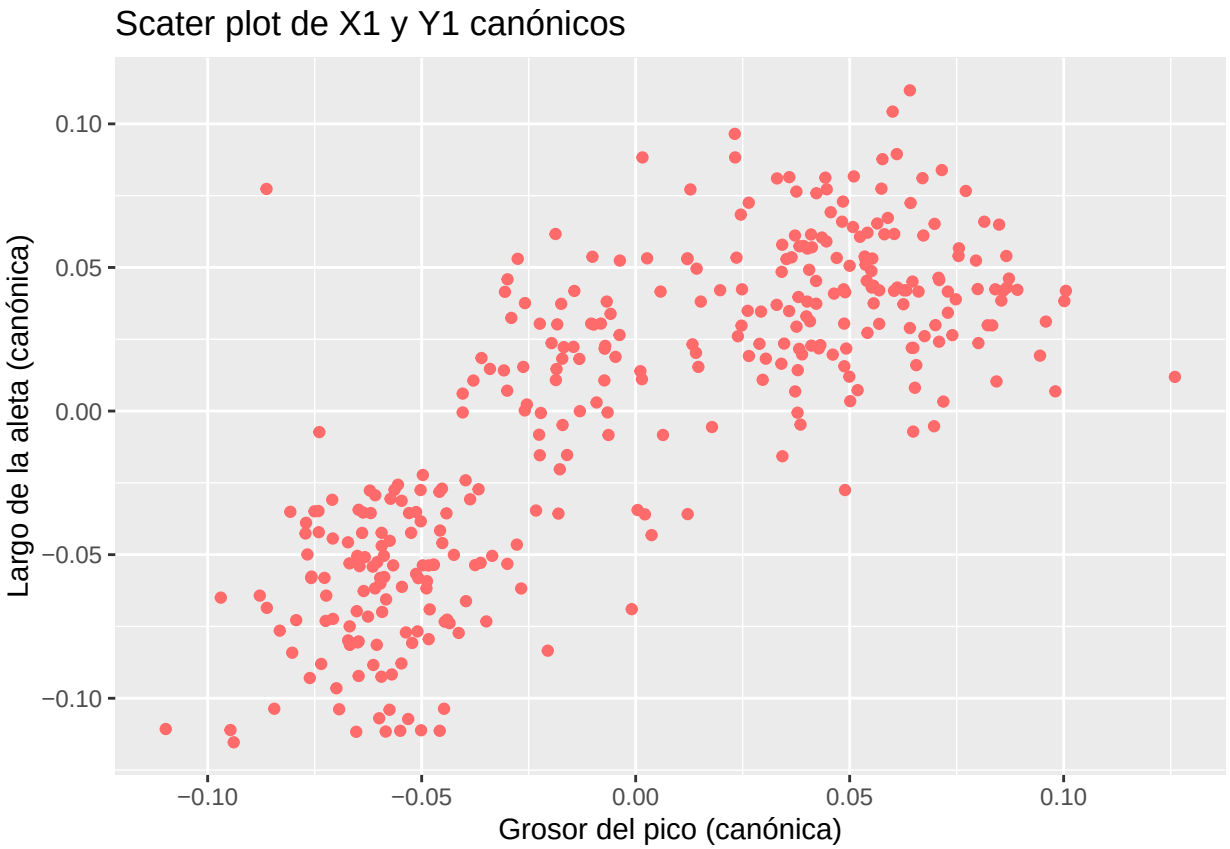
3.- Visualización de los nombres de las variables

```
colnames(ac_df)
```

```
##  [1] "ID"              "especie"         "isla"            "largo_pico_mm"
##  [5] "grosor_pico_mm"  "largo_aleta_mm"  "masa_corporal_g" "genero"
##  [9] "año"             "ac1_X"           "ac1_Y"           "ac2_X"
## [13] "ac2_Y"
```

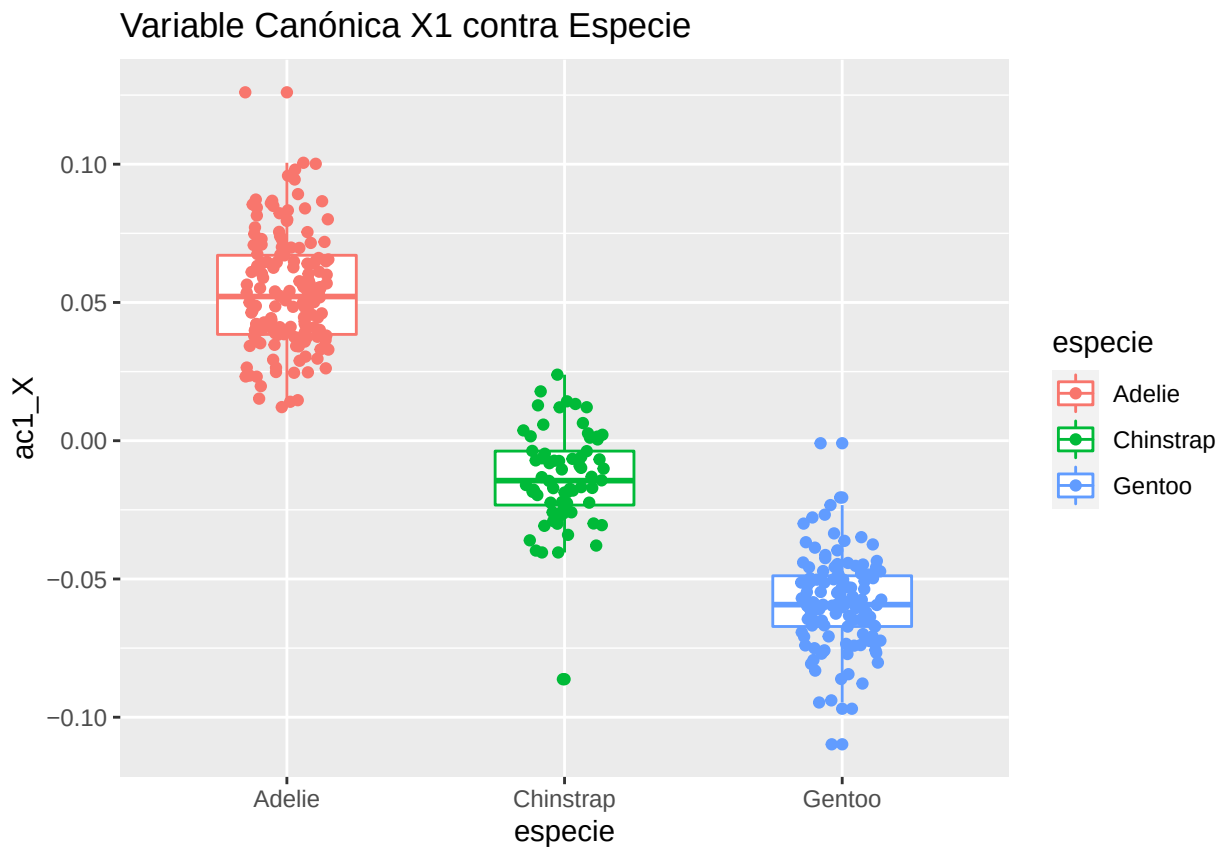
4.- Generación del grafico scater plot para la visualización de X1 y Y1

```
ac_df %>%
  ggplot(aes(x=ac1_X,y=ac1_Y))+
  geom_point(color="indianred1") +
  xlab("Grosor del pico (canónica)") +
  ylab("Largo de la aleta (canónica)") +
  ggtitle("Scater plot de X1 y Y1 canónicos")
```



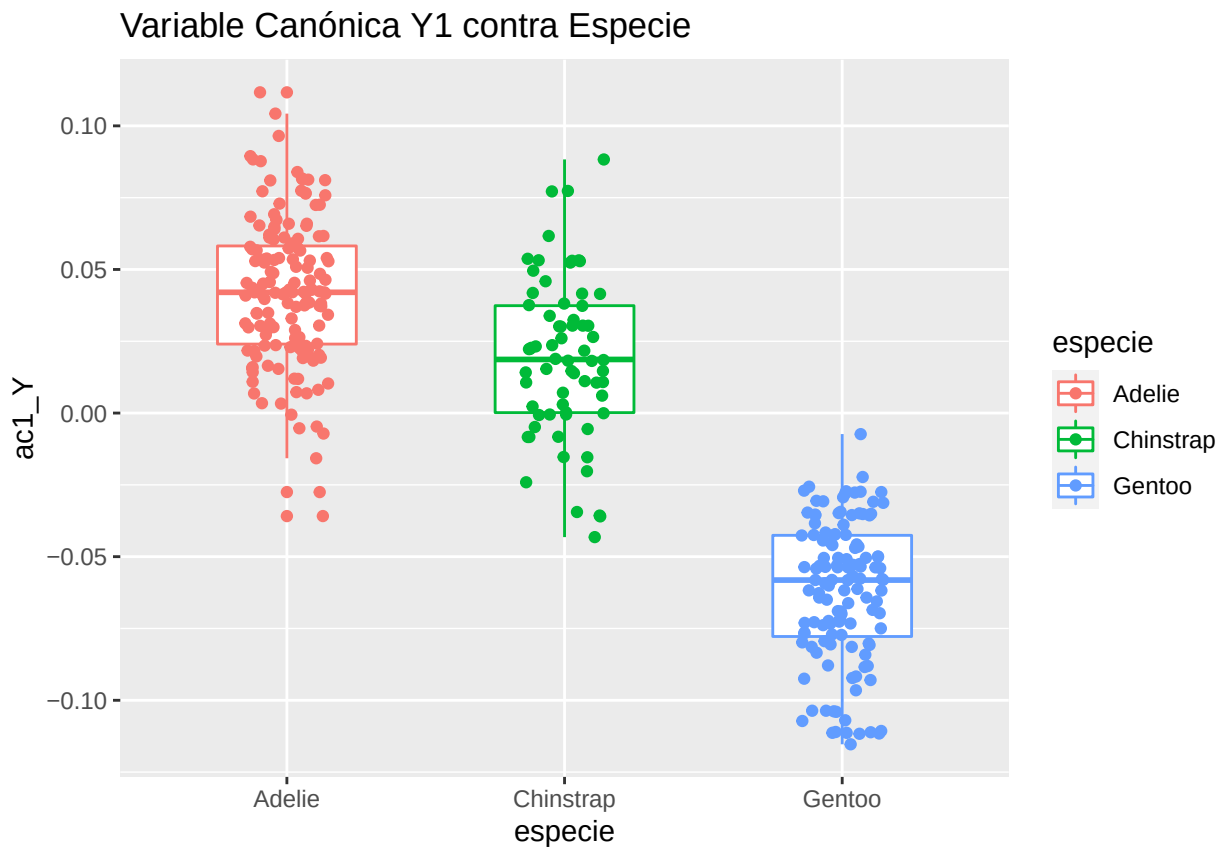
5.- Generación de un boxplot

```
ac_df %>%
  ggplot(aes(x=especie,y=ac1_X, color=especie))+
  geom_boxplot(width=0.5)+
  geom_jitter(width=0.15)+
  ggtitle("Variable Canónica X1 contra Especie")
```



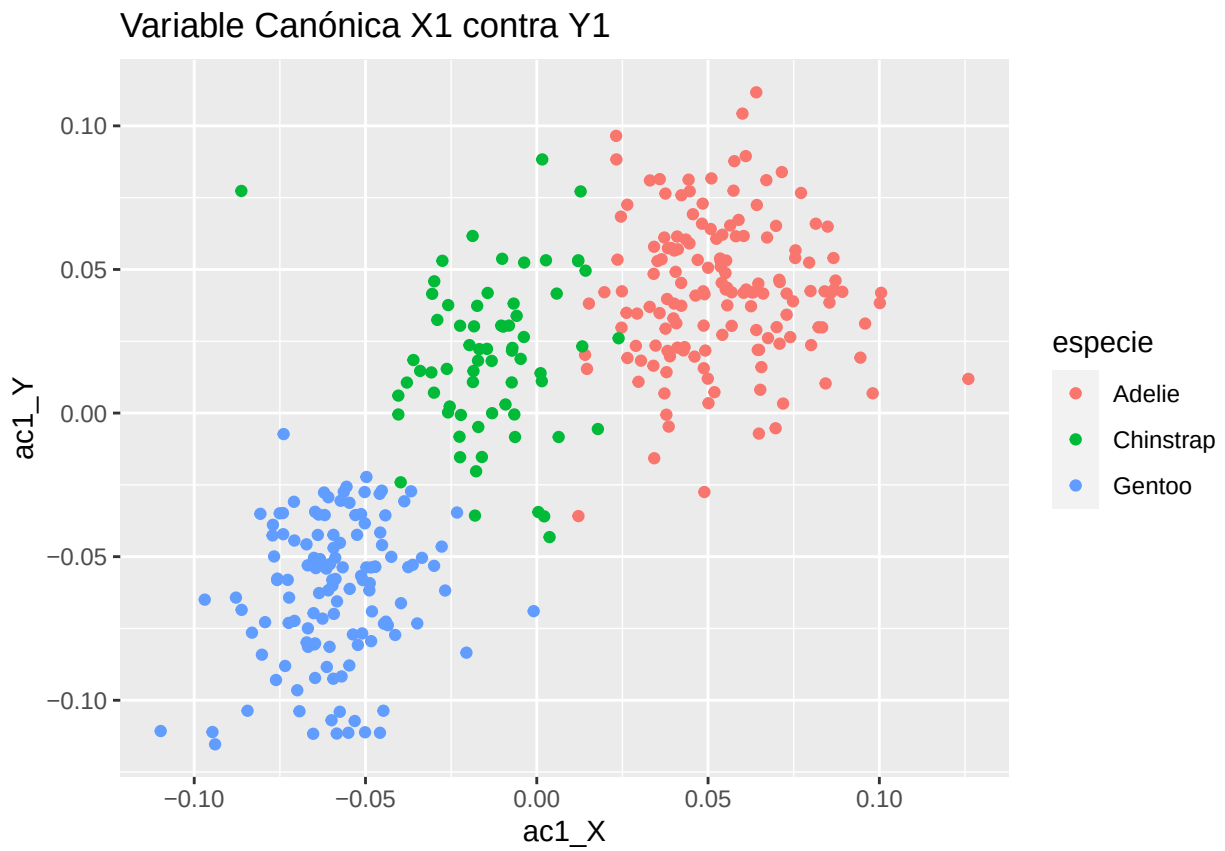
Se observa una correlacion entre la variable canónica X1 y la variable latente Especie.

```
ac_df %>%
  ggplot(aes(x=especie,y=ac1_Y, color=especie))+
  geom_boxplot(width=0.5)+
  geom_jitter(width=0.15)+
  ggtitle("Variable Canónica Y1 contra Especie")
```



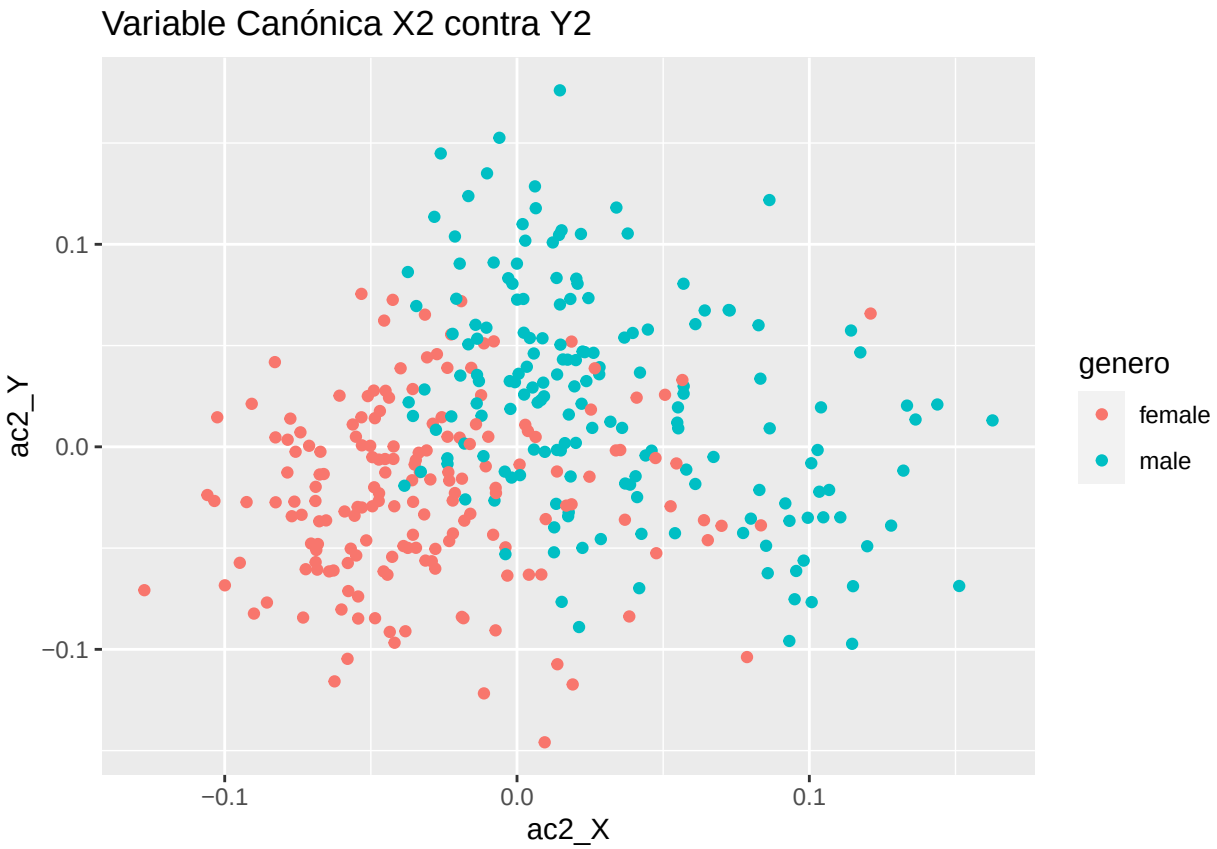
6.- Construcción de un scater plot con las variables X1 y Y1, separados por especie

```
ac_df %>%
  ggplot(aes(x=ac1_X,y=ac1_Y, color=especie))+
  geom_point()+
  ggtitle("Variable Canónica X1 contra Y1")
```



7.- Scater plot con las variables canónicas X2 y Y2 separadas por género.

```
ac_df %>%
  ggplot(aes(x=ac2_X,y=ac2_Y, color=genero))+
  geom_point()+
  ggtitle("Variable Canónica X2 contra Y2")
```



No se identifica correlación entre el conjunto de variables X2 y Y2 separadas por género.

8.- Generación de la ecuación canónica

```
ac$xcoef
```

```
##                       [,1]       [,2]
## grosor_pico_mm  0.03098538 0.04615243
## largo_pico_mm  -0.03746177 0.04107014
```

```
ac$ycoef
```

```
##                          [,1]       [,2]
## largo_aleta_mm  -0.055220261 -0.0951545
## masa_corporal_g  0.001411466  0.1100076
```

9.- Sustitución en la ecuación canónica general **U1** = 0.0309(grosor del pico) + 0.0461 (Largo del pico)

V1 = -0.0552(Largo de la aleta) + 0.0014 (masa corporal)