

Análisis Canónico

Maisy Samai Vázquez Sánchez

2022-05-29

Análisis Canónico

Introducción

El análisis canónico es una herramienta multivariante, la cual estudia la relación entre dos variables que están divididas en grupos. Este resulta como una generalización del modelo de regresión múltiple, cuyo objetivo es establecer la relación entre un conjunto de variables predictoras y un conjunto de variables respuesta. # Preparación de la matriz 1.- Se trabajó con la matriz **penguins**, es una matriz de datos cuantitativos con especies cualitativas donde se muestra información de pingüinos.

```
library(tidyverse)
library(readxl)
penguins=read_excel("C:\\Users\\Maisy\\Documents\\Multivariada\\penguins.xlsx")
```

Exploración de la matriz

2.- Dimensión de la matriz.

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

La matriz cuenta con 344 observaciones y 9 variables. 3.- Tipo de variable

```
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 ...
```

La base de datos se conforma por 4 variables de tipo carácter y 5 numéricas.

4.- Nombre de las variables

```
colnames(penguins)
```

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

5.- Se buscan valores perdidos en la matriz

```
anyNA(penguins)
```

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

No existen valores nulos en la matriz

6.- Generación de las variables **X**

```
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
```

7.- 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

1.- Visualización de la matriz X

```
ac$xcoef
```

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

2.- Visualización de la matriz Y

```
ac$ycoef
```

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

3.- Visualización de la correlación canónica

```
ac$cor
```

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

4.- 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]
```

5.- 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
```

6.- Correlación canónica entre las variables X1 y Y1

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

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

7.- 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.- Calculo 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)
```

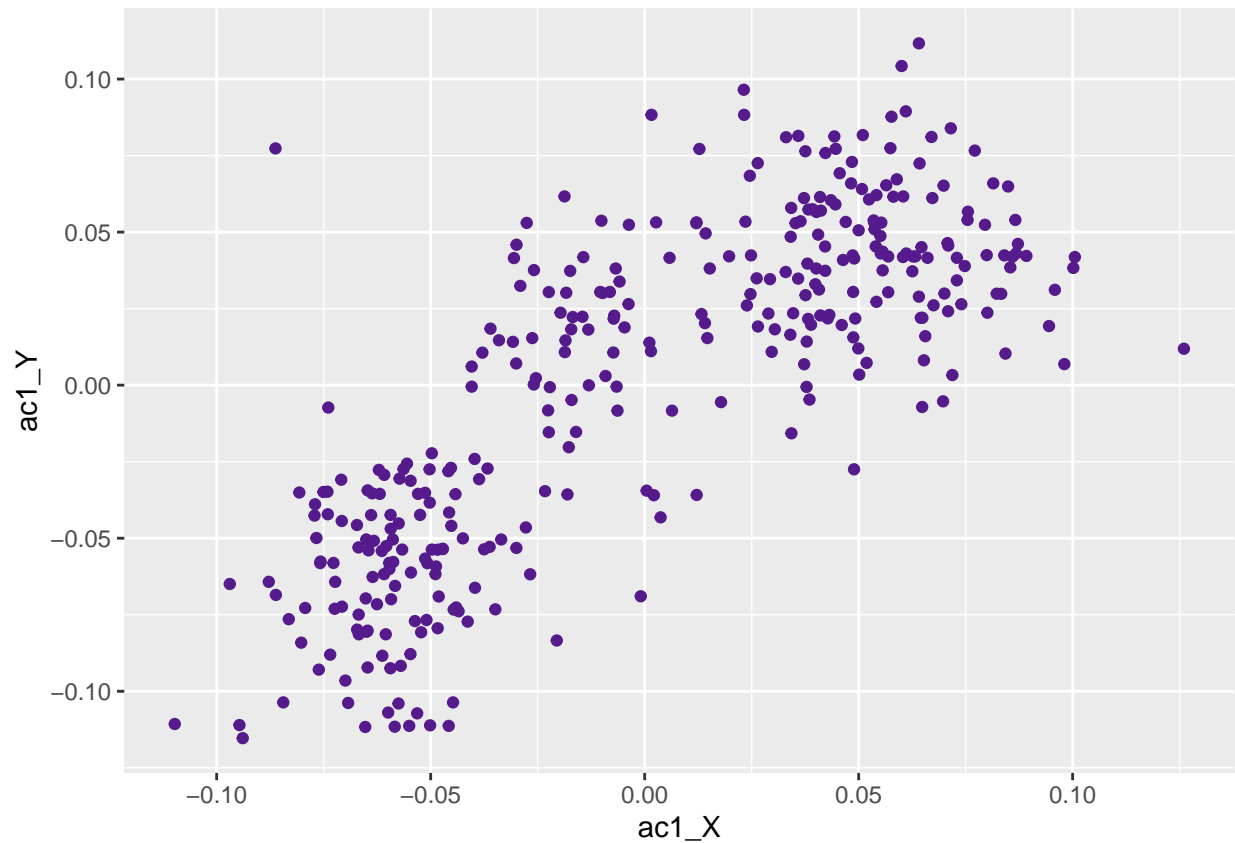
3.- Visualización de los nombres de las variables

```
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 gráfico scatter plot para la visualización de X1 y Y1

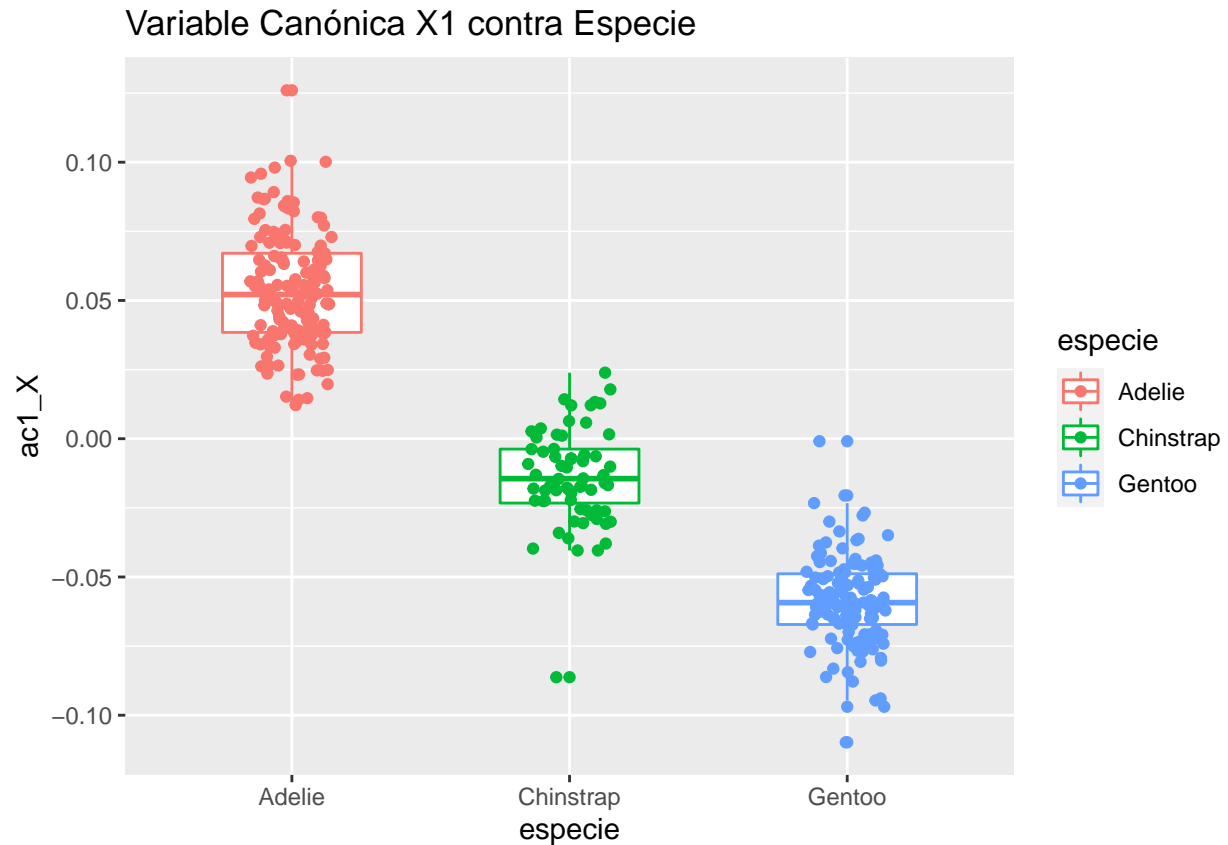
```
ac_df %>%
  ggplot(aes(x=ac1_X,y=ac1_Y))+
  geom_point(color="#551A8B")
```



Se puede observar en el gráfico una relación positiva y considerablemente fuerte, de primera instancia podríamos decir que las variables están relacionadas.

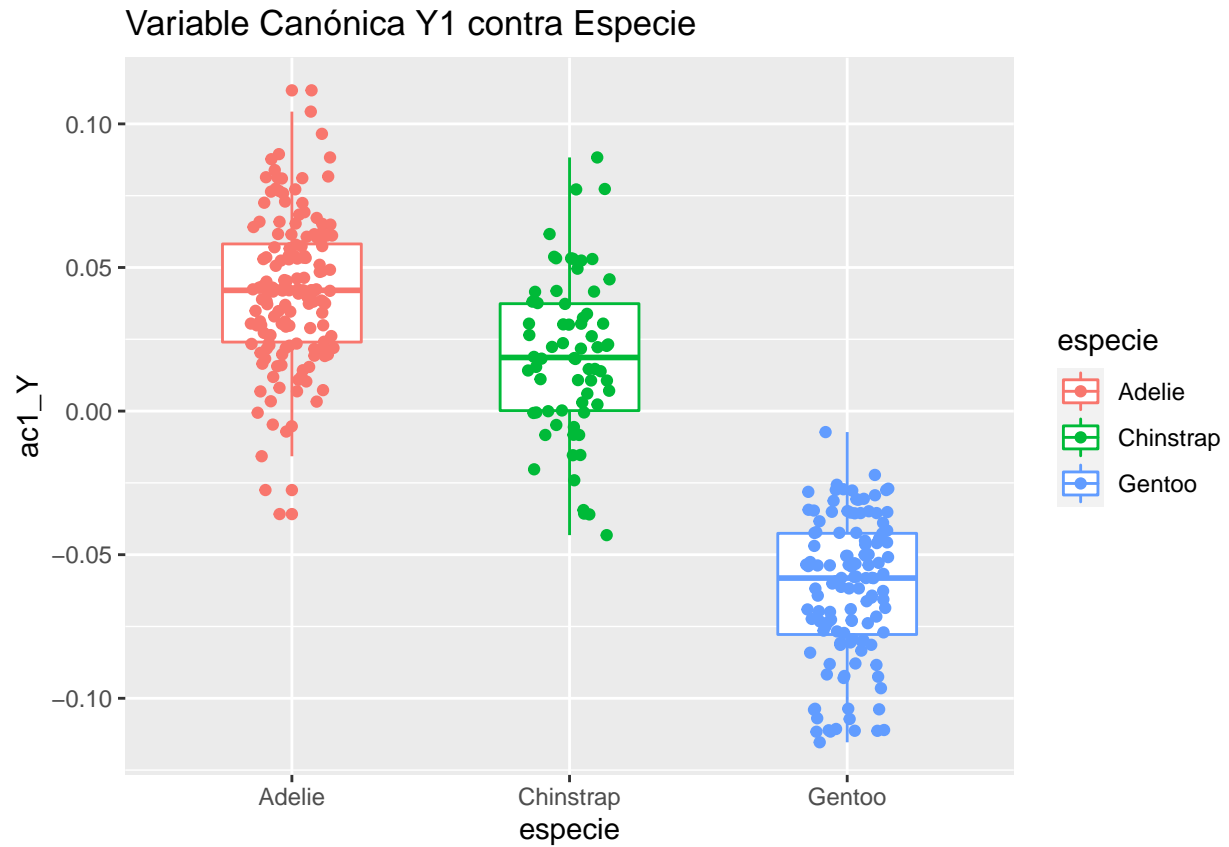
5.- Generación de un boxplot.

```
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")
```



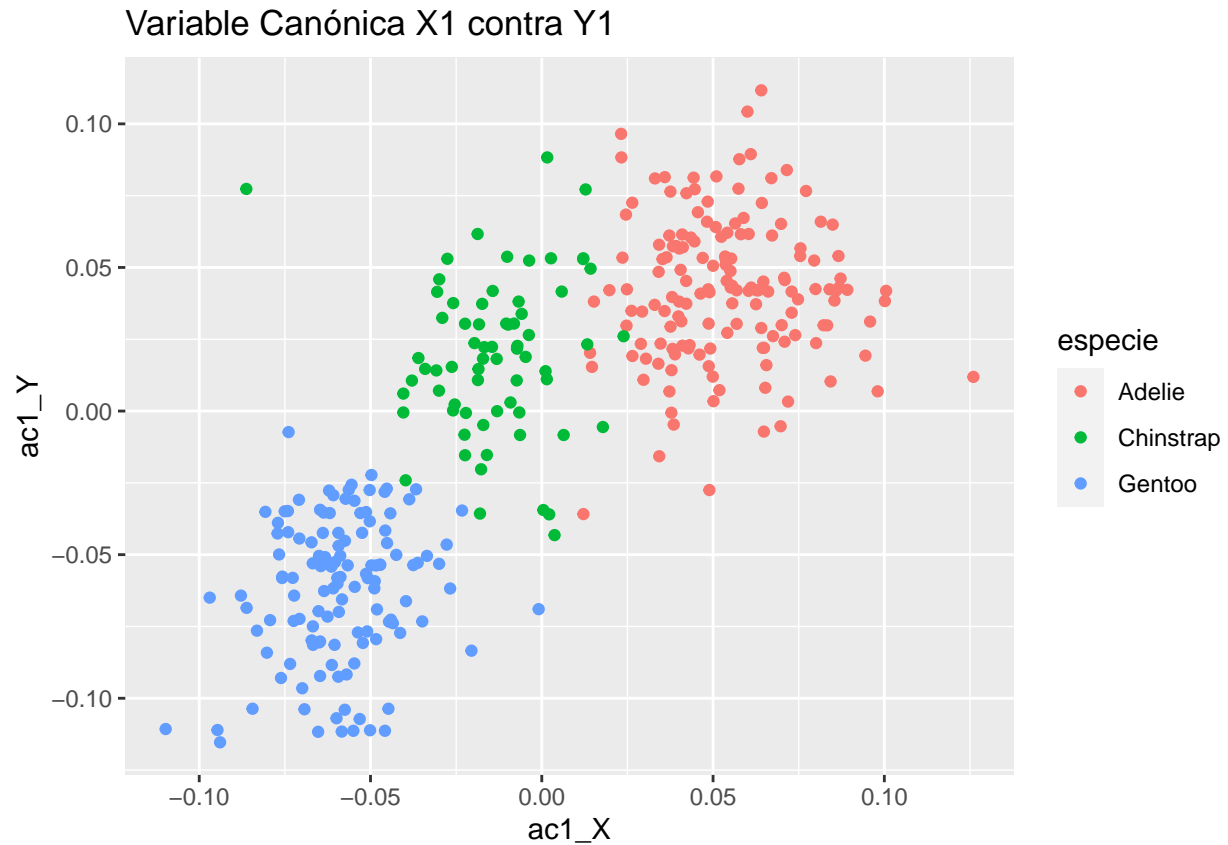
Se observa una correlación entre la variable canónica X1 y la variable latente **Especie**, podemos notar como hay similitudes entre las medianas de la especie Chinstrap y Gentoo, dado que las líneas de los 3 se encuentran en medio de la caja se considera que hay una distribución simétrica, sin embargo también podemos ver que hay valores que sobre salen de las cajas.

```
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")
```



Se puede observar como ahora las especies de Adelie y Chinstrap tienen más similitudes y la simetría se mantiene e incluso ahora hay menos valores fuera de las cajas.

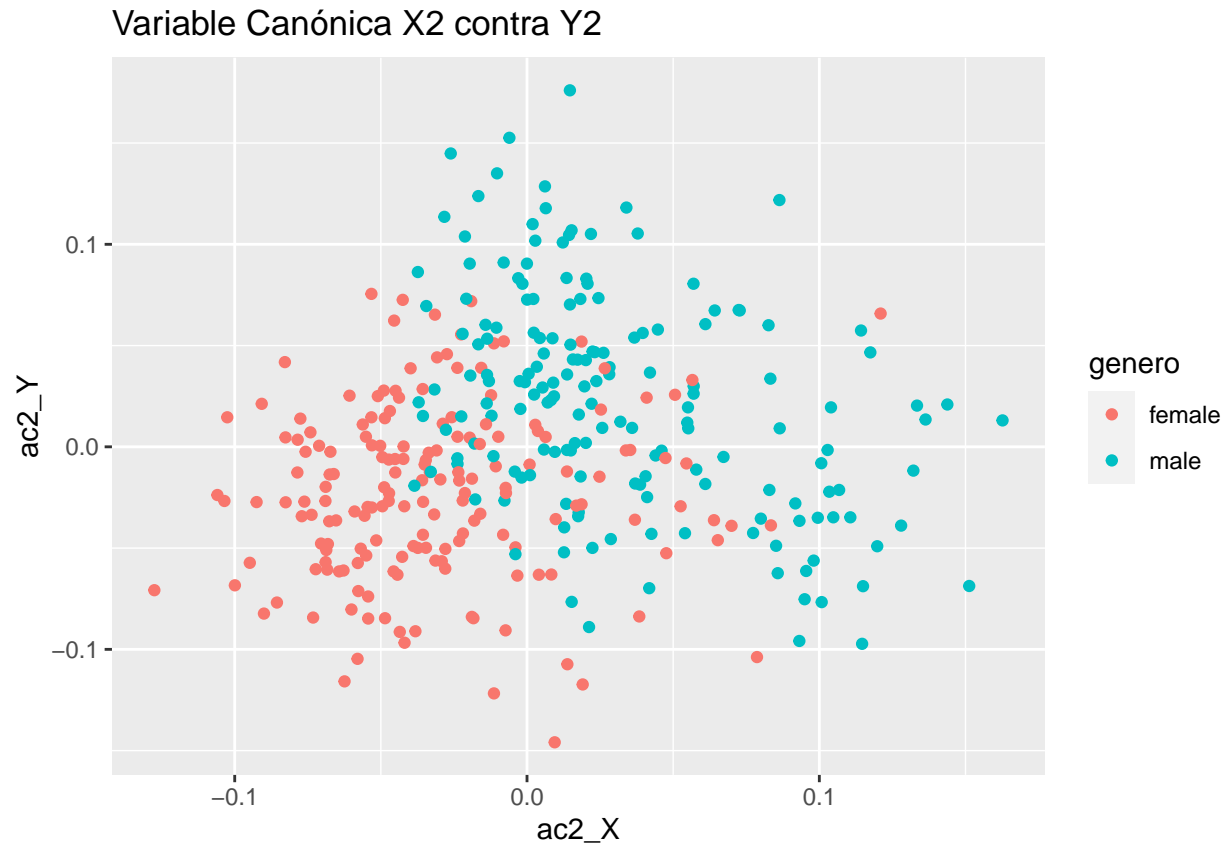
```
ac_df %>%
  ggplot(aes(x=ac1_X, y=ac1_Y, color=especie)) +
  geom_point() +
  ggtitle("Variable Canónica X1 contra Y1")
```



En este gráfico se puede observar como los grupos se separan bastante bien, por lo que no hay mucho traslape, sin embargo se puede notar como un valor se aleja bastante de su propio grupo (Chinstrap). Se aprecia una relación medianamente fuerte y positiva.

6.- Scatter plot con las variables canónicas X2 y Y2 separadas por genero.

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

En esta ultima figura **No** se identifica correlación entre el conjunto de variables X2 y Y2 separadas por genero.