Cuestionario 23/10

2024-10-28

Cuestionario de prueba 23 Octubre

1. Descriptiva: correlaciones

Con los datos del archivo cuerpo.txt, indica para mujeres, la variable numérica que presenta más correlación con C_abdomen.

```
cuerpo = read.table('cuerpo.txt',header=T)
#selecciono mujeres y var. numericas
matriz = cor(cuerpo[cuerpo$sexo==0,-25])
#selecciono solo la fila de C_abdomen
round(matriz,2)[13,]
```

```
## A_hombros  A_pelvis    A_cade  AP_pecho  AD_pecho    A_codo  A_muneca A_rodilla
##      0.28      0.60      0.59      0.63      0.49      0.51      0.41      0.59
## A_tobillo C_hombros   C_pecho C_cintura C_abdomen  C_cadera   C_muslo  C_biceps
##      0.46      0.61      0.77      0.84      1.00      0.83      0.70      0.75
##   C_brazo C_rodilla  C_gemelo C_tobillo  C_muneca      edad      peso    altura
##      0.64      0.61      0.52      0.49      0.49      0.41      0.80      0.23
```

Para hallar el máximo:

```
which.max(round(matriz,2)[13,-13])
```

```
## C_cintura
##        12
```

2. Regresión: variable logaritmica

En el archivo "fev.txt" se muestran datos de un estudio sobre la Capacidad Pulmonar ("fev" Forced Expiratory Volume en litros) en 654 jóvenes entre 3 y 19 años. En el archivo además de la variable fev se incluyen las variables: age (años del individuo), ht (estatura en pulgadas), sex (cualitativa, mujer=0, hombre=1) y smoke (cualitativa, No-fumador=0, fumador=1).

Estima el modelo de regresión múltiple utilizando log(fev) como variable dependiente y el resto de las variabes como regresores.

```
fev = read.table('fev.txt',header=T)
fev$fev = log(fev$fev)
#ya tenemos la variable dependiente, ahora hacemos la regresion
lm1 = lm(fev ~ ., dat = fev)
summary(lm1)
```

```
##
## Call:
## lm(formula = fev ~ ., data = fev)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.63278 -0.08657  0.01146  0.09540  0.40701
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.943998   0.078639 -24.721  < 2e-16 ***
## age          0.023387   0.003348   6.984  7.1e-12 ***
## ht           0.042796   0.001679  25.489  < 2e-16 ***
## sex          0.029319   0.011719   2.502   0.0126 *
## smoke       -0.046068   0.020910  -2.203   0.0279 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1455 on 649 degrees of freedom
## Multiple R-squared:  0.8106, Adjusted R-squared:  0.8095
## F-statistic: 694.6 on 4 and 649 DF,  p-value: < 2.2e-16
```

Interpretación de los resultados:

- Podemos ver el coeficiente para cada variable independiente y su p-valor: si es menor a nuestra *alpha* dicha variable será significativa.
- Residual standard error: es el error medio que se comete al predecir la variable dependiente con esta ecuación de regresión.
- Degrees of freedom: nº de observaciones - nº variables independientes.
- R-Squared: porcentaje de la varianza de la variable dependiente que está explicada por el modelo.

3. PCA

Con los datos correspondientes a mujeres del archivo cuerpo.txt, realiza un análisis de componentes principales de las variables estandarizadas (12 medidas de contorno, van de la 10 a la 21).

```
cuerpo = read.table('cuerpo.txt',header=T)
muj = cuerpo[cuerpo$sexo==0,10:21]
fit1 = princomp(muj,cor=T)
#porcentaje de variabilidad explicada
fit1$sdev^2/12*100
```

```
##     Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7
## 71.7101333  8.4702158  5.3954905  3.0281301  2.7992547  2.1304964  1.7344519
##     Comp.8     Comp.9    Comp.10    Comp.11    Comp.12
##  1.3672597  1.2231082  0.8797813  0.8018298  0.4598483
```

```
#correlación de CP2 con C_muslo
fit1$loadings['C_muslo','Comp.2']
```

```
## [1] 0.03580187
```

```
# ¿La solución de 3 componentes deja sin explicar
# el 12.26 % de la variable C_muslo?
source('prinfact.R')
fit2 = prinfact(muj,3) #ojo, pedimos 3 CP
fit2$loadings
```

```
##              Comp 1      Comp 2      Comp 3 communality uniquiness
## C_hombros 0.8341144  0.07902883  0.33056589   0.8112661 0.18873387
## C_pecho   0.8704837  0.29837846  0.24601729   0.9072961 0.09270386
```

```
## C_cintura 0.8613778  0.38530336 -0.01209975   0.8905768 0.10942319
## C_abdomen 0.8155069  0.39711806 -0.18047764   0.8553265 0.14467352
## C_cadera  0.9049913  0.14088841 -0.28882599   0.9222793 0.07772067
## C_muslo   0.8652339  0.03609471 -0.35699333   0.8773767 0.12262329
## C_biceps  0.8990499  0.13742210  0.13654656   0.8458205 0.15417952
## C_brazo   0.9020096 -0.11200374  0.20732423   0.8691495 0.13085046
## C_rodilla 0.8520201 -0.27918052 -0.18256733   0.8372108 0.16278918
## C_gemelo  0.8166777 -0.38485460 -0.15534004   0.8392061 0.16079392
## C_tobillo 0.7358360 -0.47805596 -0.06920591   0.7747815 0.22521848
## C_muneca  0.7879072 -0.32835170  0.33196107   0.8388107 0.16118929
```

```
#¿La puntuación (score) más alta en valor absoluto obtenida en el primer componente es 12.815?
max(fit2$scores[,1])
```

```
## [1] 12.81514
```

4. PCA

Con los datos del archivo cuerpo.txt, realiza un análisis de componentes principales (en correlaciones) de las variables 11, 12, 13, 14, 15, 16, 19, 20 utilizando los datos correspondientes a hombres. Indica el porcentaje de variablidad de la variable C_muslo que está explicada por la solución de 3 componentes.

```
hom = cuerpo[cuerpo$sexo==1,c(11,12,13,14,15,16,19,20)]
fit3 = prinfact(hom,3)
fit3$loadings
```

```
##              Comp 1       Comp 2     Comp 3 communality uniquiness
## C_pecho   0.8355899  0.258326273  0.3146210   0.8639293 0.13607071
## C_cintura 0.8446576  0.374917189 -0.2640320   0.9237223 0.07627774
## C_abdomen 0.8372765  0.385364030 -0.2845534   0.9305081 0.06949194
## C_cadera  0.9160802  0.066831440 -0.1408075   0.8634961 0.13650390
## C_muslo   0.8308266 -0.240620238  0.1830164   0.7816659 0.21833410
## C_biceps  0.7439800 -0.002641495  0.6139807   0.9304855 0.06951454
## C_gemelo  0.7602146 -0.522285486 -0.1289044   0.8673246 0.13267536
## C_tobillo 0.7113649 -0.446300884 -0.2579350   0.7717549 0.22824510
```

5. LDA

El conjunto de datos iris (archivo: "lirios.txt") contiene la información de 150 lirios, de los cuales: 50 son setosa, 50 son versicolor y 50 son virginica. Disponemos de cuatro medidas en cada observación: el largo y ancho del sépalo y pétalo, en centímetros. Realiza el análisis discriminante utilizando species como variable respuesta en función de las cuatro variables anteriormente indicadas, sin estandarizar.

```
library(MASS)
lirios = read.table('lirios.txt',header=T)
lirios$Species = factor(lirios$Species)
lda1 = lda(Species~., data = lirios)
lda1
```

```
## Call:
## lda(Species ~ ., data = lirios)
##
## Prior probabilities of groups:
##     setosa versicolor  virginica
##  0.3333333  0.3333333  0.3333333
##
## Group means:
```

```
##            Sepal.Length Sepal.Width Petal.Length Petal.Width
## setosa            5.006       3.428        1.462       0.246
## versicolor        5.936       2.770        4.260       1.326
## virginica         6.588       2.974        5.552       2.026
##
## Coefficients of linear discriminants:
##                     LD1         LD2
## Sepal.Length  0.8293776  0.02410215
## Sepal.Width   1.5344731  2.16452123
## Petal.Length -2.2012117 -0.93192121
## Petal.Width  -2.8104603  2.83918785
##
## Proportion of trace:
##    LD1    LD2
## 0.9912 0.0088
```

```
#El centroide de la especie Setosa,
#en la variable Sepal.Length se sitúa en 6.006
lda1$means
```

```
##            Sepal.Length Sepal.Width Petal.Length Petal.Width
## setosa            5.006       3.428        1.462       0.246
## versicolor        5.936       2.770        4.260       1.326
## virginica         6.588       2.974        5.552       2.026
```

```
#vemos que la media de Sepal.Length para la clase setosa es 5.006
```

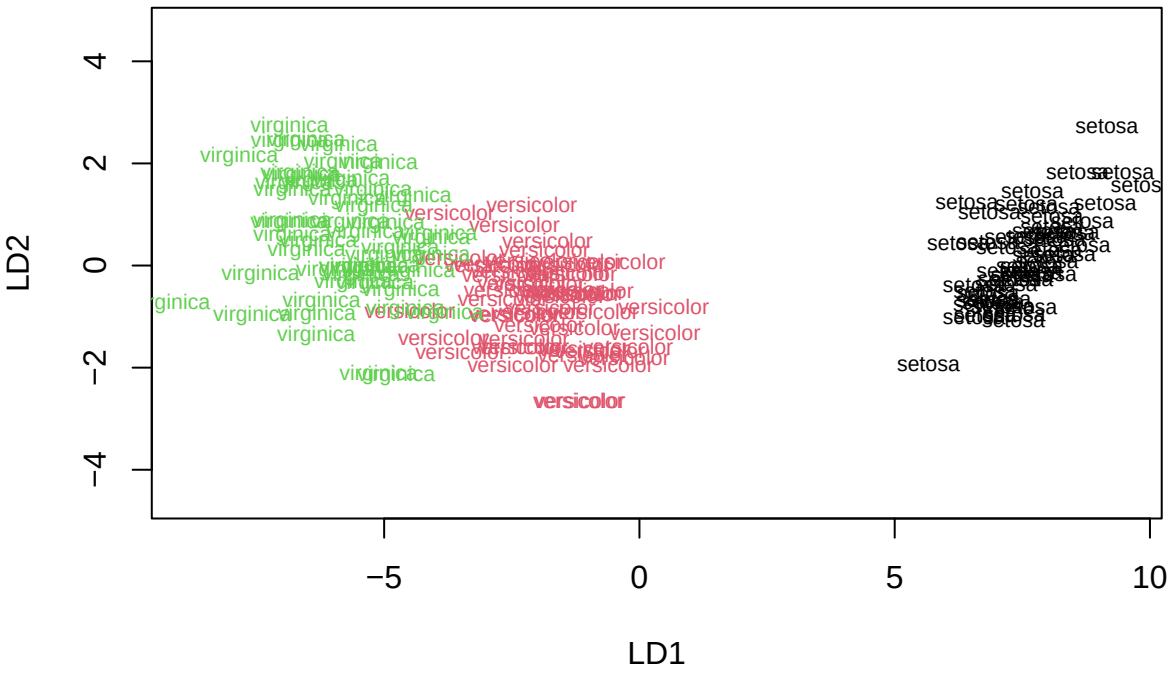
```
#El coeficiente de la función discriminante 1,
#para la variable Sepal.Width, vale 1.53
lda1$scaling
```

```
##                     LD1         LD2
## Sepal.Length  0.8293776  0.02410215
## Sepal.Width   1.5344731  2.16452123
## Petal.Length -2.2012117 -0.93192121
## Petal.Width  -2.8104603  2.83918785
```

```
#vemos que el coeficiente es 1.5344731
```

```
#podemos predecir el mismo dataset usando Leave-One-Out CV
p1 = predict(lda1)
```

```
#Gráfico
plot(lda1, col=as.integer(lirios$Species))
```




```
#De las 150 observaciones, hay tres de ellas
#que se han clasificado incorrectamente
library(multiUS)
lda2 = ldaPlus(lirios[,1:4],grouping=lirios$Species)
lda2$class
```

```
## $orgTab
##             pred
## orig         setosa versicolor virginica Sum
##   setosa         50          0         0  50
##   versicolor      0         48         2  50
##   virginica       0          1        49  50
##   Sum            50         49        51 150
##
## $perTab
##             pred
## orig         setosa versicolor virginica Sum
##   setosa        100          0         0 100
##   versicolor      0         96         4 100
##   virginica       0          2        98 100
##
## $corPer
## [1] 98
```

6. Regresión simple

Con los datos del archivo cuerpo.txt estima el modelo de regresión simple entre el Peso (variable dependiente) y A_codo como regresor. Un individuo pesa 55.5 kg y la medida de su A_codo es 11.2 cm. Calcula el error de predicción (residuo) del modelo para esta persona.

```
rsim = lm(peso ~ A_codo, data = cuerpo)
nueva = data.frame(A_codo=11.2)
55.5-predict(rsim,nueva)
```

```
##       1
## 3.62301
```

7. Regresión múltiple

```
rmult = lm(peso ~ A_muneca + C_pecho + C_gemelo +
             C_tobillo + C_muneca + sexo, data = cuerpo)
summary(rmult)
```

```
##
## Call:
## lm(formula = peso ~ A_muneca + C_pecho + C_gemelo + C_tobillo +
##     C_muneca + sexo, data = cuerpo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1544  -2.7975  -0.2356   2.6109  18.5181
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -81.556591   3.588659 -22.726  < 2e-16 ***
## A_muneca      1.121290   0.423753   2.646  0.00840 **
## C_pecho       0.821394   0.037911  21.666  < 2e-16 ***
## C_gemelo      1.349320   0.112678  11.975  < 2e-16 ***
## C_tobillo     0.613891   0.191720   3.202  0.00145 **
## C_muneca      0.009522   0.373800   0.025  0.97969
## sexo         -0.446404   0.687679  -0.649  0.51654
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.411 on 500 degrees of freedom
## Multiple R-squared:  0.892,  Adjusted R-squared:  0.8907
## F-statistic: 688.5 on 6 and 500 DF,  p-value: < 2.2e-16
```

- *"Todos los regresores en el modelo tienen efecto significativo"* FALSO
- *A igualdad del resto de los regresores, hay diferencia significativa en el peso de hombres y mujeres igual a -0.446 kg* FALSO: no es significativa
- *La desviación típica de los residuos es 4.411 y sus unidades son cm* FALSO: las unidades son kg: los residuos son el error en la variable dependiente (peso).
- *A igualdad del resto de los regresores, el aumento de un kg en el peso de una persona, incrementa 1.121 cm, la variable A_muneca* FALSO: el peso incremeta 1.121 kg por cada cm que incrementa A_muneca

8. Regresión múltiple

```
rmul2 = lm(peso ~ sexo + C_muslo, data = cuerpo)
summary(rmul2)
```

```
##
## Call:
## lm(formula = peso ~ sexo + C_muslo, data = cuerpo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -21.3884  -3.6732  -0.2318   3.0220  22.1476
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -44.45503    3.37873  -13.16   <2e-16 ***
## sexo         18.82584    0.52350   35.96   <2e-16 ***
## C_muslo       1.83677    0.05873   31.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.874 on 504 degrees of freedom
## Multiple R-squared:  0.8071, Adjusted R-squared:  0.8063
## F-statistic:  1054 on 2 and 504 DF,  p-value: < 2.2e-16
```

9. LDA

El archivo "cars.txt" tiene datos de 391 coches y contiene las siguientes variables:mpg, engine, horse, weight, accel, origin, cylinders. Todas las variables son continuas excepto origin que es cualitativa y representa el lugar de fabricación del vehículo. Tiene la siguiente codificación: 1 coches americanos (USA), 2 coches europeos (EUR) y 3 coches japoneses (JAP). Hay 244 coches USA, 68 coches UER y 79 coches JAP. Realiza el análisis discriminante entre los tres tipos de coches según su origen (variable origin) utilizando como variables explicativas las siguientes mpg (1), engine (2), weight (4), accel (5), cylinders (7)

```
cars = read.table('cars.txt',header=T)
(ldacars = ldaPlus(cars[,c(1,2,4,5,7)],grouping=cars$origin))
```

```
## Call:
## ldaPlus(cars[, c(1, 2, 4, 5, 7)], grouping = cars$origin)
##
## Prior probabilities of groups:
##         1         2         3
## 0.6240409 0.1739130 0.2020460
##
## Group means:
##        mpg   engine   weight    accel cylinders
## 1 20.07869 247.2807 3366.918 14.97582  6.270492
## 2 27.60294 109.6324 2433.471 16.79412  4.161765
## 3 30.45063 102.7089 2221.228 16.17215  4.101266
##
## Coefficients of linear discriminants:
##                      LD1          LD2
## mpg         0.0479925753  0.126704267
## engine     -0.0195654724  0.013992246
## weight      0.0006767203 -0.001745456
## accel      -0.1326637769 -0.120266049
## cylinders   0.1861326580  0.395961793
##
## Proportion of trace:
##    LD1    LD2
## 0.9505 0.0495
```

```
#Las dos variables con más peso (positivo o negativo) en la
#primera función discriminante estandarizada es engine, weight

#OJO: coeficientes estandarizados
```

```
ldacars$standCoefWithin
```

```
##                  LD1        LD2
## mpg        0.3061845  0.8083517
## engine    -1.5498053  1.1083431
## weight     0.4566936 -1.1779442
## accel     -0.3532551 -0.3202426
## cylinders  0.2518874  0.5358425
```

```
nueva = data.frame(mpg=33,engine=91,weight=1795,
                   accel=17.4,cylinders=4)
predict(ldacars,nueva)
```

```
## $class
## [1] 3
## Levels: 1 2 3
##
## $posterior
##              1         2         3
## [1,] 0.2295695 0.2272863 0.5431441
##
## $x
##           LD1      LD2
## [1,] 1.156975 1.014418
```

```
#lo clasifica como JAP (clase 3)
```

```
#La segunda coordenada del centroide para los
#coche europeos es 0.0210113 en valor absoluto.
ldacars$centroids
```

```
##          LD1         LD2
## 1 -0.7128957  0.01081138
## 2  1.0759601 -0.38620689
## 3  1.2757121  0.29903914
```

```
#vemos que es -0.38620689
```

```
#El número de funciones discriminantes
#significativas con nivel de significación alpha=0.05 es 1.
ldacars$sigTest
```

```
##           WilksL        F df1 df2           p
## 1 to 1 0.5815768 55.39867   5 385 2.69889e-43
```

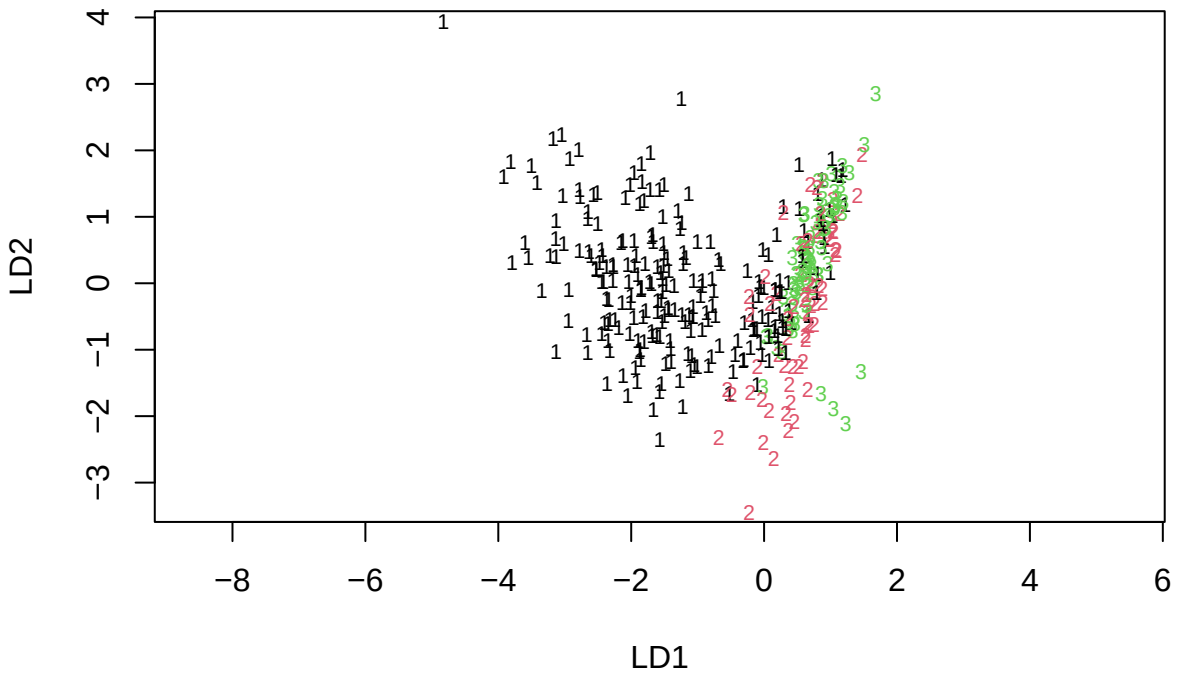
```
#si, solo hay una función LD significativa
```

```
#matriz de confusion
ldacars$class
```

```
## $orgTab
##      pred
## orig    1   2   3 Sum
##   1   210   8  26 244
##   2    11  27  30  68
##   3     4  17  58  79
##   Sum 225  52 114 391
##
```

```
## $perTab
##     pred
## orig          1          2          3        Sum
##    1  86.065574   3.278689  10.655738 100.000000
##    2  16.176471  39.705882  44.117647 100.000000
##    3   5.063291  21.518987  73.417722 100.000000
##
## $corPer
## [1] 75.44757
```

```
plot(ldacars,col=as.integer(cars$origin))
```



10. LDA 'cuerpo.txt'

```
ldacuerpo = ldaPlus(cuerpo[,10:17],grouping=cuerpo$sexo)
ldacuerpo$standCoefWithin
```

```
##                 LD1
## C_hombros  0.3685152
## C_pecho   -0.1164560
## C_cintura  1.0347362
## C_abdomen -0.7198295
## C_cadera  -0.2266642
## C_muslo   -0.8284417
## C_biceps   0.2333304
## C_brazo    0.6957187
```

11 LDA 'penguins.txt'

```
penguins = read.table('penguins.txt',header=T)
(ldapen = ldaPlus(penguins[,3:6],grouping=penguins$species))
```

```
## Call:
```

```
## ldaPlus(penguins[, 3:6], grouping = penguins$species)
##
## Prior probabilities of groups:
##    Adelie Chinstrap    Gentoo
## 0.4415205 0.1988304 0.3596491
##
## Group means:
##           culmen_length_mm culmen_depth_mm flipper_length_mm body_mass_g
## Adelie             38.79139        18.34636          189.9536    3700.662
## Chinstrap          48.83382        18.42059          195.8235    3733.088
## Gentoo             47.50488        14.98211          217.1870    5076.016
##
## Coefficients of linear discriminants:
##                             LD1          LD2
## culmen_length_mm   0.08832666 -0.417870885
## culmen_depth_mm   -1.03730494 -0.021004854
## flipper_length_mm  0.08616282  0.013474680
## body_mass_g        0.00129952  0.001711436
##
## Proportion of trace:
##   LD1   LD2
## 0.866 0.134
```

```
#El análisis muestra que son importantes
#(significativas con alfa = 0.05)
#las tres funciones discriminates
ldapen$sigTest
```

```
##            WilksL        F df1 df2             p
## 1 to 2 0.01878543 528.8705   8 672 4.160107e-284
## 2 to 2 0.30092718 260.9574   3 337  1.622851e-87
```

```
#FALSO: las dos son significativas pero es que solo hay 2
```

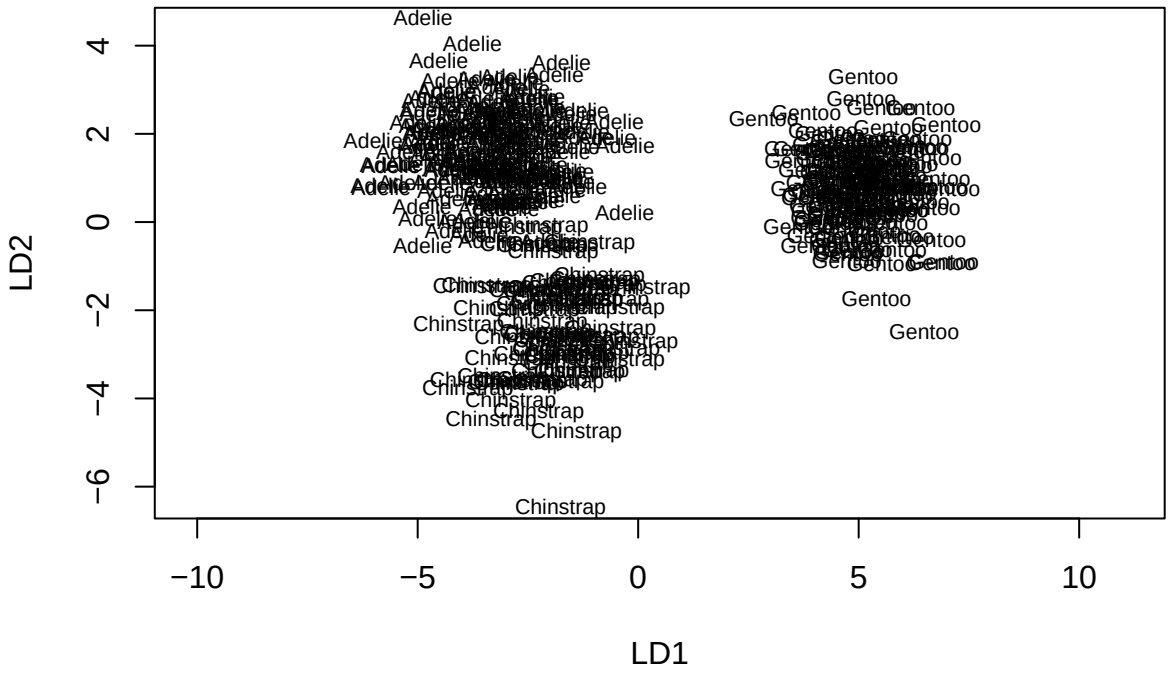
```
#Los coeficientes de la funciones discriminates
#estandarizadas son todos menores que uno en valor absoluto
ldapen$standCoefWithin
```

```
##                            LD1         LD2
## culmen_length_mm   0.2614340 -1.23683654
## culmen_depth_mm   -1.1626360 -0.02354274
## flipper_length_mm  0.5722528  0.08949247
## body_mass_g        0.6007348  0.79115277
```

```
#FALSO
```

```
plot(ldapen,col=as.integer(penguins$species))
```

```
## Warning in eqscplot(X[, 1L:2L], xlab = xlab, ylab = ylab, type = "n", ...): NAs
## introduced by coercion
```




```
#Hay más del 23% de observaciones de la especie
#Gentoo mal clasificadas
ldapen$class
```

```
## $orgTab
##            pred
## orig        Adelie Chinstrap Gentoo Sum
##   Adelie       150         1      0 151
##   Chinstrap      3        65      0  68
##   Gentoo         0         0    123 123
##   Sum          153        66    123 342
##
## $perTab
##            pred
## orig             Adelie   Chinstrap      Gentoo         Sum
##   Adelie     99.3377483   0.6622517   0.0000000 100.0000000
##   Chinstrap   4.4117647  95.5882353   0.0000000 100.0000000
##   Gentoo      0.0000000   0.0000000 100.0000000 100.0000000
##
## $corPer
## [1] 98.83041
```