Analisis exploratorio multivariado

Table of contents

Analisis exploratorio multivariante

Medidas Multivariantes

Vector de Medias

```
data("iris")
str(iris)
```

```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

```
summary(iris)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
```

```
 virginica :50
```

Ahora se procede a calcular e interpretar el vector de medias para las variables cuantitativas

```
subconjunto = iris[, 1:4]
p<-ncol(subconjunto) # variable que contiene el número de variables
n <- nrow(subconjunto)
vectormedias1 = colMeans(subconjunto) #Primera opción
vectormedias1
```

```
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
    5.843333     3.057333     3.758000     1.199333
```

Estas medidas representan los valores medios de las diferentes características de las flores de iris en el conjunto de datos. Son útiles para tener una idea general de cómo son las dimensiones promedio de las flores en el dataset. Ahora veamos la diferencia entre realizar una lectura de datos y una interpretación de la información.

Interpretación:Podemos observar que las dimensiones promedio de los sépalos y pétalos varían entre las especies de iris. Por ejemplo, si comparamos la longitud promedio de los pétalos, vemos que es más alta que la de los sépalos. Esto podría indicar diferencias morfológicas entre las especies de iris. También podemos notar que la longitud promedio de los pétalos es más grande que la de los sépalos, mientras que el ancho promedio de los sépalos es más grande que el de los pétalos. Esto podría indicar diferencias en la forma y proporciones de los diferentes órganos de la flor.Las diferencias en las dimensiones promedio de los sépalos y pétalos entre las especies de iris podrían indicar diferencias en la forma y tamaño de las flores entre las especies. Estas diferencias podrían ser útiles para la clasificación y discriminación de las especies de iris. El vector de medias proporciona información útil sobre las dimensiones promedio de los sépalos y pétalos en el dataset iris, lo que nos permite realizar inferencias sobre las características morfológicas de las diferentes especies de iris incluidas en el conjunto de datos.

Matriz de Varianzas y Covarianzas “S”

Para su calculo se utiliza la función cov() que calcula la matriz S corregida. Siempre usaremos el subconjunto de variables númericas.La matriz de varianzas y covarianzas del dataset iris proporciona información sobre cómo las diferentes características (longitud y ancho de los sépalos y pétalos) están relacionadas entre sí. Cada elemento de la matriz representa la

covarianza entre dos características específicas. La diagonal principal de la matriz contiene las varianzas de cada característica, mientras que los elementos fuera de la diagonal principal contienen las covarianzas entre pares de características.

```
S= cov(subconjunto)
S
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    0.6856935  -0.0424340    1.2743154   0.5162707
Sepal.Width    -0.0424340   0.1899794   -0.3296564  -0.1216394
Petal.Length    1.2743154  -0.3296564    3.1162779   1.2956094
Petal.Width     0.5162707  -0.1216394    1.2956094   0.5810063
```

Medidas Glovales de Variabilidad

Estas medidas de varianza nos permiten entender la variabilidad en las características de las flores del dataset iris desde diferentes perspectivas. Proporcionan información valiosa sobre la dispersión de los datos y pueden ayudar a contextualizar los resultados obtenidos en el análisis posterior.

```
##Varianza total de los datos de la base de datos llamada iris
VT<-sum(diag(S))
VT
```

```
[1] 4.572957
```

```
##Varianza media de los datos de la base llamada iris
VM<-VT/p
VM
```

```
[1] 1.143239
```

```
##varianza generalizada
VG<-det(S)
VG
```

```
[1] 0.00191273
```

```r
## Desviación Generalizada
DG<-(VG)^(1/2)
DG
```

```
[1] 0.04373476
```

```r
##Varianza efectiva :
VE<-(VG)^(1/p)
VE
```

```
[1] 0.2091286
```

```r
##Desviación promedio:
DP <- (VG)^(1/(2*p))
DP
```

```
[1] 0.4573058
```

La varianza total representa la variabilidad total en el conjunto de datos. En el caso del dataset iris, esta medida nos indica cuánto varían las características de las flores en general. Una varianza total más alta sugiere una mayor variabilidad en las características de las flores entre las diferentes observaciones. La varianza media representa la variabilidad promedio en las características del dataset. En este caso, una varianza media de aproximadamente 1.143239 indica que, en promedio, las características de las flores en el dataset iris varían en torno a este valor.

Variabilidad con Distancias

```r
#Distancia de Mahalanobis
x<-as.matrix(subconjunto)
media<-colMeans(subconjunto)
matriz.media<-matrix(media,nrow = n,ncol=p,byrow = TRUE)
SI<-solve(S)
distm<-((x-matriz.media)%*%SI)%*%t((x-matriz.media))
#distm
```

Matriz de Correlaciones

La matriz de correlaciones R es una matriz cuadrada y simétrica, y además semidefinida positiva, que tiene unos en la diagonal principal y fuera de ella los coeficientes de correlación lineal entre pares de variables.Para su calculo utilizaremos la función "cor()".

```
R = round(cor(subconjunto),2) #Se utiliza la función round para redondear a 2 decimales cada
R
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length         1.00       -0.12         0.87        0.82
Sepal.Width         -0.12        1.00        -0.43       -0.37
Petal.Length         0.87       -0.43         1.00        0.96
Petal.Width          0.82       -0.37         0.96        1.00
```

Medidas Globales de Dependencia Lineal

```
#Coeficiente de dependencia
CD= det(R)
CD
```

```
[1] 0.00950873
```

```
#Dependencia Global
DRp = 1- (CD)^ 1/(p-1)
DRp
```

```
[1] 0.9968304
```

```
#Coeficiente de correlación promedio
CCP = sqrt(DRp)
CCP
```

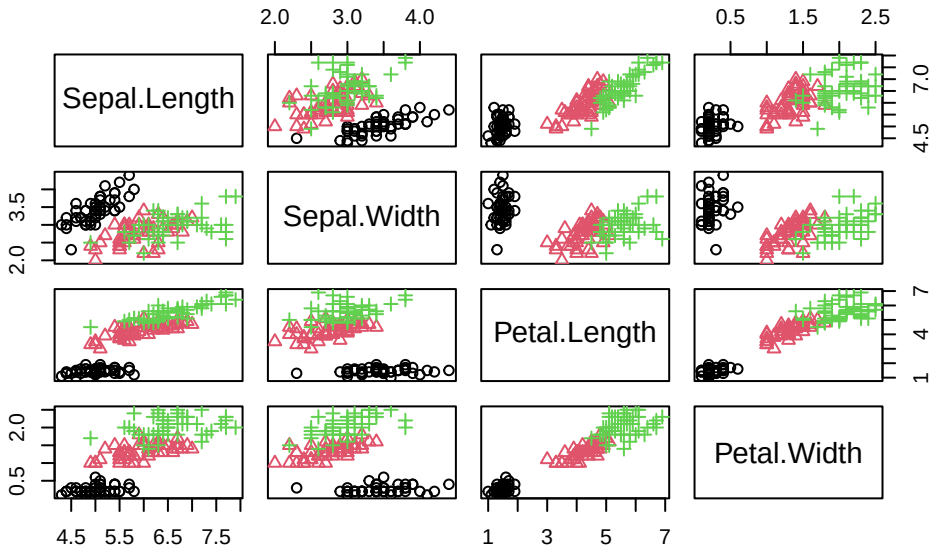
```
[1] 0.998414
```

Análisis Gráfico.

En este apartado la interpretación se deja al estudiante y se mostrará el código para la creación de los diferentes gráficos multivariantes.

Matrices de dispersión.

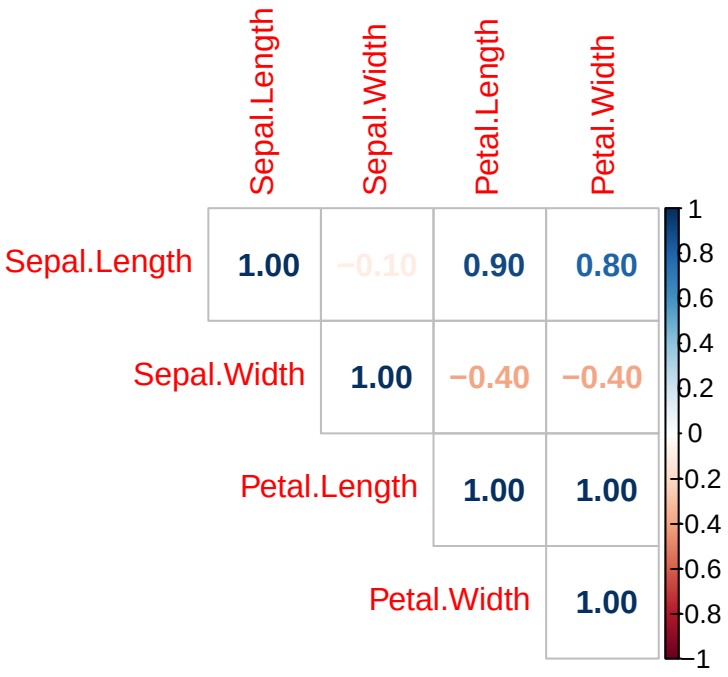
```
pairs(iris[,1:4],pch=as.numeric(iris$Species),col=iris$Species)
```




```
#install.packages("corrplot")
library(corrplot)
```

```
corrplot 0.92 loaded
```

```
correlacion<-round(cor(subconjunto), 1)
corrplot(correlacion, method="number", type="upper")
```



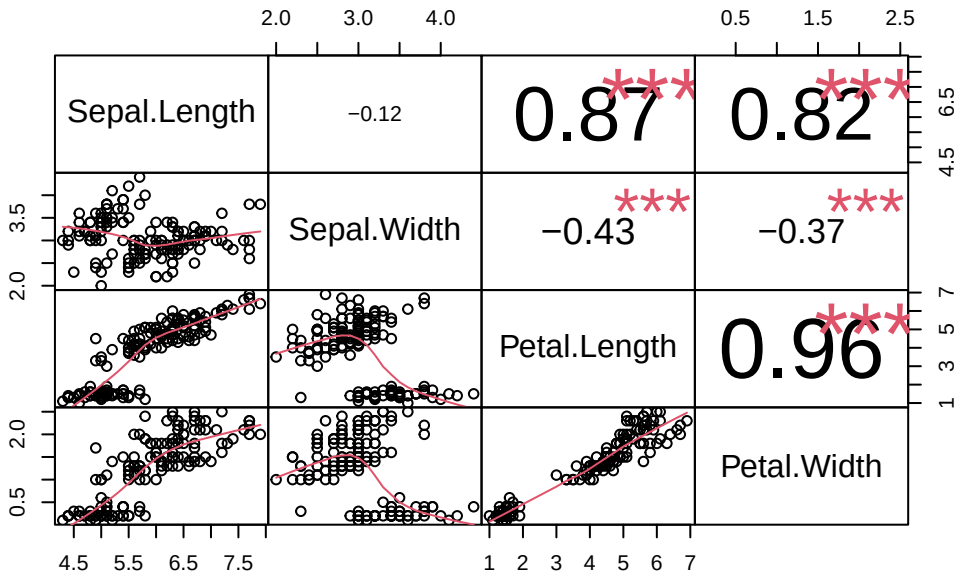
Gráficos de dispersión para calcular los coeficientes de una sola vez y ver si son estadísticamente significativos.

```
library(PerformanceAnalytics)
```

```
Cargando paquete requerido: xts

Cargando paquete requerido: zoo


Adjuntando el paquete: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric


Adjuntando el paquete: 'PerformanceAnalytics'

The following object is masked from 'package:graphics':

    legend
```

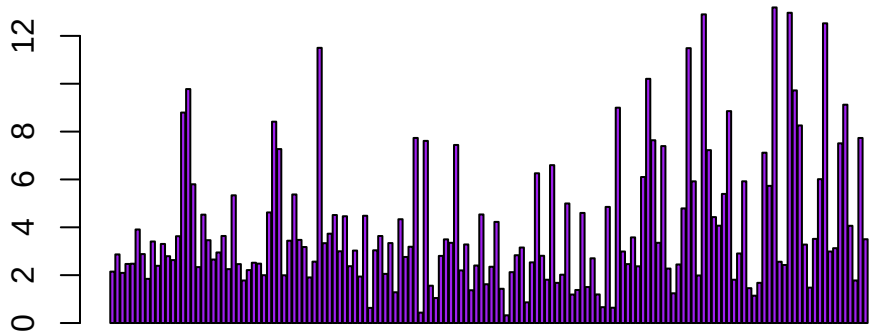
```
chart.Correlation(subconjunto, histogram = F, pch = 19)
```

```
Warning in par(usr): argument 1 does not name a graphical parameter
Warning in par(usr): argument 1 does not name a graphical parameter
Warning in par(usr): argument 1 does not name a graphical parameter
Warning in par(usr): argument 1 does not name a graphical parameter
Warning in par(usr): argument 1 does not name a graphical parameter
Warning in par(usr): argument 1 does not name a graphical parameter
```



Distancias de Mahalanobis

```
Vmedias <- colMeans(subconjunto)
numDatos <- nrow(subconjunto)
MatrVarCov <- cov(subconjunto)*(numDatos-1)/numDatos
dismahalanobis <- mahalanobis(subconjunto, Vmedias, MatrVarCov)
barplot(dismahalanobis, main = "Distancias de Mahalanobis", col =
"purple")
```


Distancias de Mahalanobis
0
2
4
6
8
12