Distancia de Mahalanobis

Arleth Michell Morales García

2022-05-21

DISTANCIA DE MAHALANOBIS

Cargar los datos

```
ventas= c( 1054, 1057, 1058, 1060, 1061, 1060, 1061, 1062, 1062, 1064, 1062, 1062, 1064, 1056, 1066, 10
clientes= c(63, 66, 68, 69, 68, 71, 70, 70, 71, 72, 72, 73, 73, 75, 76, 78)
```

Utilizamos la función data.frame() para crear un juego de datos en R

```
datos <- data.frame(ventas ,clientes)
```

Exploración de la matriz

```
dim(datos)
```

```
## [1] 16  2
```

La matriz cuenta con 16 observaciones y 2 variables

```
str(datos)
```

```
## 'data.frame':    16 obs. of  2 variables:
##  $ ventas  : num  1054 1057 1058 1060 1061 ...
##  $ clientes: num  63 66 68 69 68 71 70 70 71 72 ...
```

Las variables son númericas

```
summary(datos)
```

```
##      ventas        clientes
##  Min.   :1054   Min.   :63.00
##  1st Qu.:1060   1st Qu.:68.75
##  Median :1062   Median :71.00
##  Mean   :1061   Mean   :70.94
##  3rd Qu.:1062   3rd Qu.:73.00
##  Max.   :1070   Max.   :78.00
```

En la venta el valor máximo es de 1070 y el valor mínimo es de 1054

CÁLCULO DE LA DISTANCIA

El método de distancia Mahalanobis mejora el método clásico de distancia de Gauss eliminando el efecto que pueden producir la correlación entre las variables a analizar.

Determinar el nÃºmero de outlier que queremos encontrar.

```
num.outliers <- 2
```

Ordenar los datos de mayor a menor distancia, según la métrica de ## Mahalanobis.

```
mah.ordenacion <- order(mahalanobis(datos , colMeans( datos), cov(datos)), decreasing=TRUE)
mah.ordenacion
```

```
##  [1] 14 16  1 15  2  5  3 10 13  8 12  4  6  7  9 11
```

Los datos 14, 16 y 1 tienen una mayor distancia de Mahalanobis y los datos 7, 9 y 11 tienen una menor distancia

Generar un vector boleano los dos valores más alejados segun la distancia Mahalanobis.

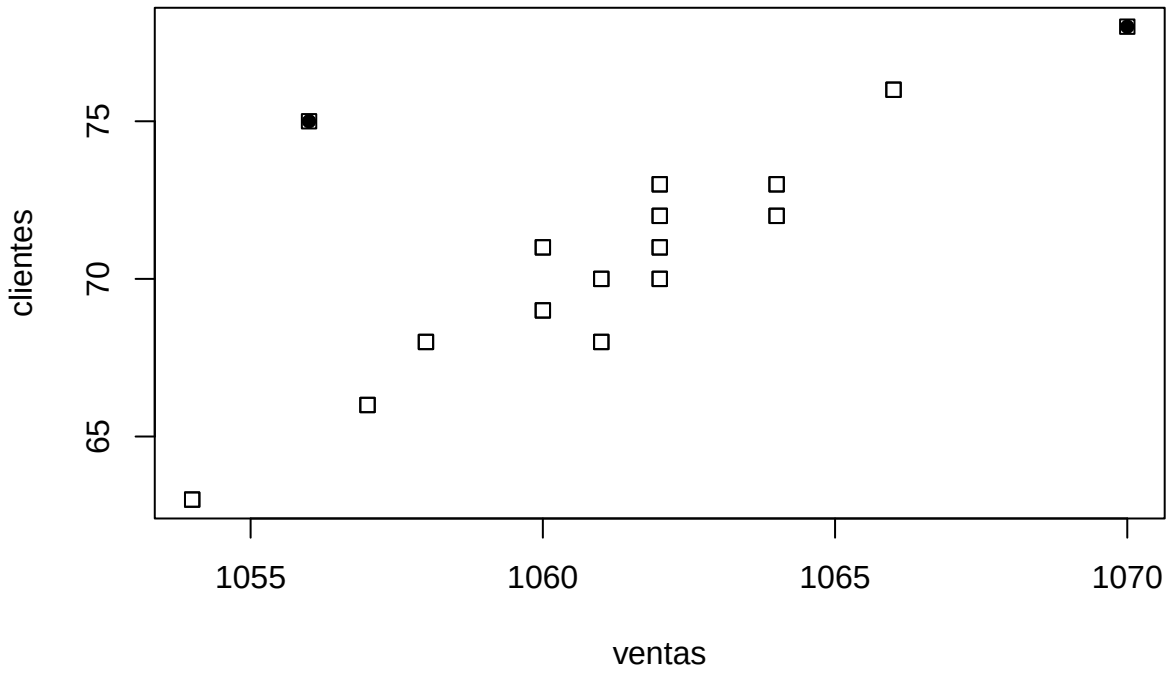
```
outlier2 <- rep(FALSE , nrow(datos))
outlier2[mah.ordenacion[1:num.outliers]] <- TRUE
```

Resaltar con un punto relleno los 2 valores outliers.

```
colorear.outlier <- outlier2 * 16
```

Visualizar el gráfico con los datos destacando sus outlier.

```
plot(datos , pch=0)
points(datos , pch=colorear.outlier)
```



SEGUNDO EJEMPLO MAHALANOBIS

```
require(graphics)
```

```
ma <- cbind(1:6, 1:3)
(S <-  var(ma))
```

```
##      [,1] [,2]
## [1,]  3.5  0.8
## [2,]  0.8  0.8
```

```
mahalanobis(c(0, 0), 1:2, S)
```

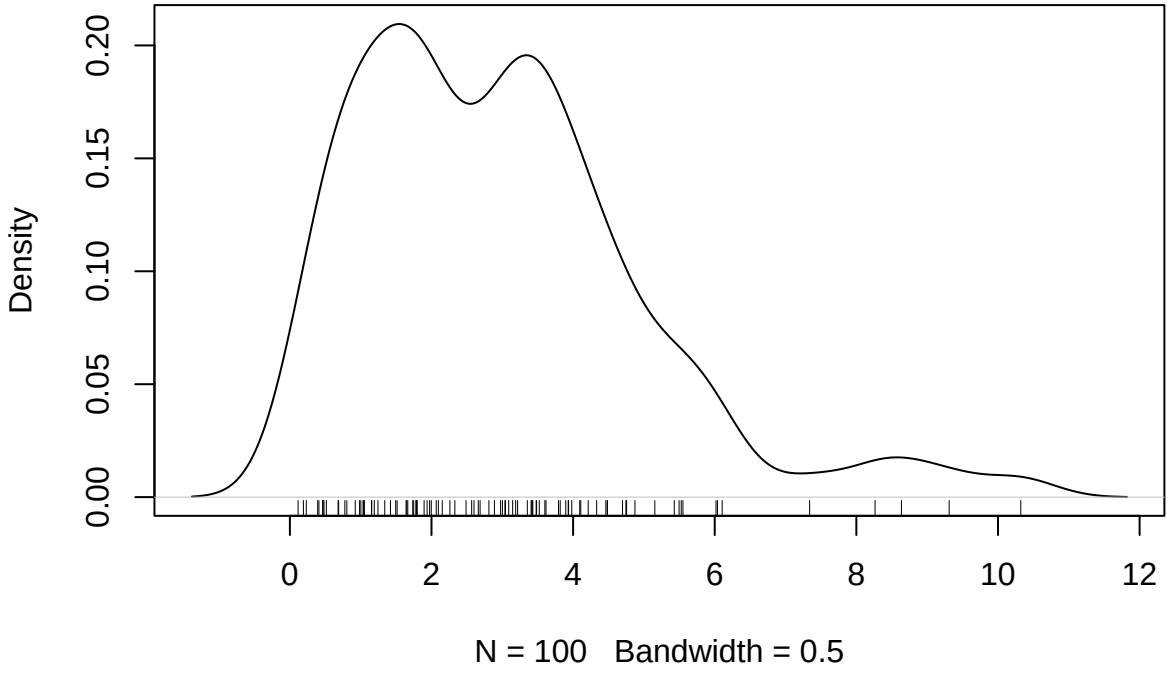
```
## [1] 5.37037
```

```
x <- matrix(rnorm(100*3), ncol = 3)
stopifnot(mahalanobis(x, 0, diag(ncol(x))) == rowSums(x*x))
```

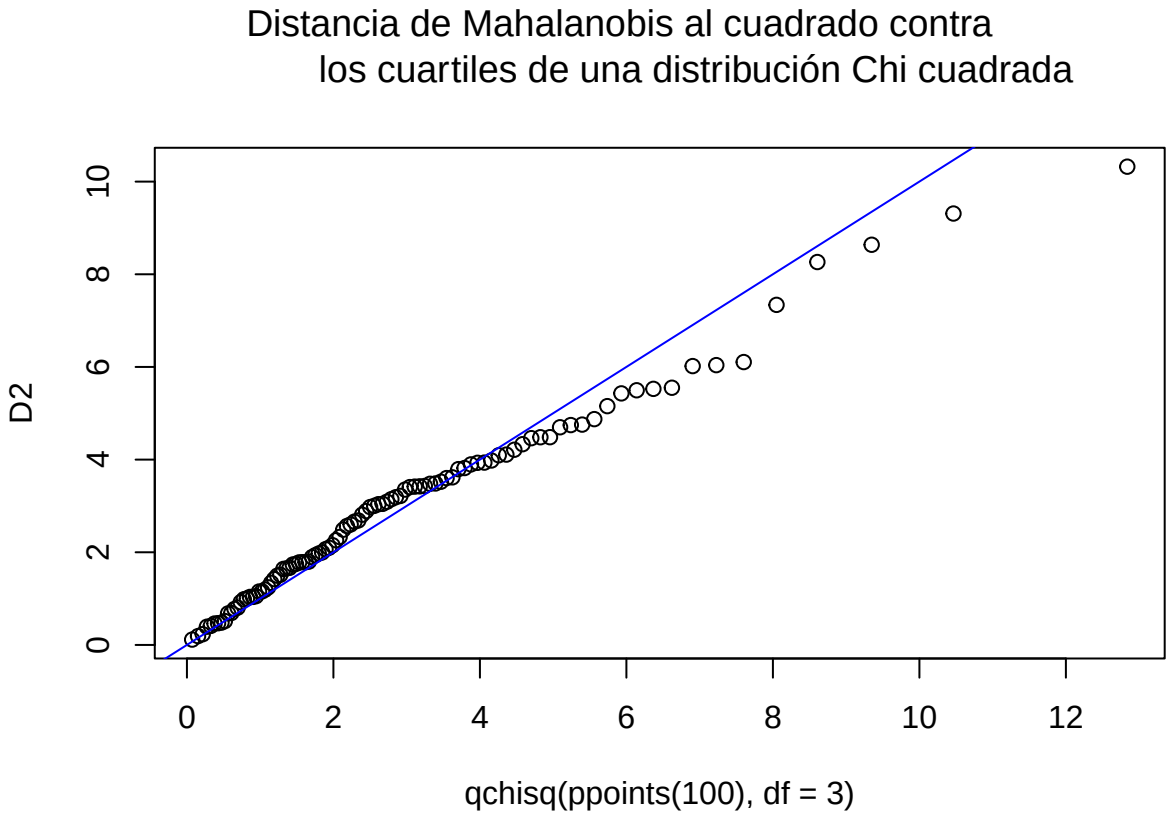
##- Here, D^2 = usual squared Euclidean distances

```
Sx <- cov(x)
D2 <- mahalanobis(x, colMeans(x), Sx)
plot(density(D2, bw = 0.5),
     main="Gráfico de densidad de la distancia cuadrada de Mahalanobis, n=100, p=3") ; rug(D2)
```

Gráfico de densidad de la distancia cuadrada de Mahalanobis, n=100,




```
qqplot(qchisq(ppoints(100), df = 3), D2,
       main = expression("Distancia de Mahalanobis al cuadrado contra
       los cuartiles de una distribución Chi cuadrada"))
abline(0, 1, col = 'blue')
```



TERCER EJERCICIO: DISEÑAR TU PROPIO EJERCICIO

Trabajé con una matriz de arboles, extraída del paquete datasets que se encuentra precargada en R

Se seleciona la matriz de datos.

```
arboles<-datasets::trees
```

Exploración de la matriz

```
colnames(arboles)
```

```
## [1] "Girth"  "Height" "Volume"
```

Cuenta con tres variables

Girt = Circunferencia

Height = Altura

Volume = Volumen

```
dim(arboles)
```

```
## [1] 31  3
```

La matriz cuenta con 31 observaciones y 3 variables, para este ejercicio solo necesitamos dos variables.

Seleccionamos las variables con las que vayamos a trabajar.

```
arboles1<- arboles[,2:3]
```

Hacemos un resumen con la nueva de la matriz de datos

```
summary(arboles1)
```

```
##      Height       Volume
##  Min.   :63   Min.   :10.20
##  1st Qu.:72   1st Qu.:19.40
##  Median :76   Median :24.20
##  Mean   :76   Mean   :30.17
##  3rd Qu.:80   3rd Qu.:37.30
##  Max.   :87   Max.   :77.00
```

La altura maxima de un árbol es de 87 y la mínima de 63.

CALCULO DE LA DISTANCIA DE MAHALANOBIS

Determinar el número de outlier que queremos encontrar.

```
num.outliers <- 2
```

Ordenar los datos de mayor a menor distancia, según la métrica de ## Mahalanobis.

```
mah.ordenacion <- order(mahalanobis(arboles1 , colMeans( arboles1), cov(arboles1)), decreasing=TRUE)
mah.ordenacion
```

```
##  [1] 31 20 18  3  6 28  2  5 17  7 27 26 29 30 24  1  9 14 11  4 25 19  8 23 15
## [26] 22 12 10 13 16 21
```

Los datos 31,20 y 18 tienen una mayor distancia de Mahalanobis y los datos 13, 16 y 21 tienen una menor distancia de Mahalanobis

Generar un vector boleano los dos valores más alejados segun la distancia Mahalanobis.

```
outlier2 <- rep(FALSE , nrow(arboles1))
outlier2[mah.ordenacion[1:num.outliers]] <- TRUE
```

Resaltar con un punto relleno los 2 valores outliers.

```
colorear.outlier <- outlier2 * 16
```

Visualizar el gráfico con los datos destacando sus outlier.

```
plot(arboles1 , pch=0)
points(arboles1 , pch=colorear.outlier)
```