mahalanobis

Edgar Ortiz Mota

2022-05-29

Cargar los datos

```
ventas= c( 1054, 1057, 1058, 1060, 1061, 1060, 1061,
           1062, 1062, 1064, 1062, 1062, 1064, 1056,
           1066, 1070)
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)

dim(datos)
```

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

```
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 ...
```

```
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
```

Calculo 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 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(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
```

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

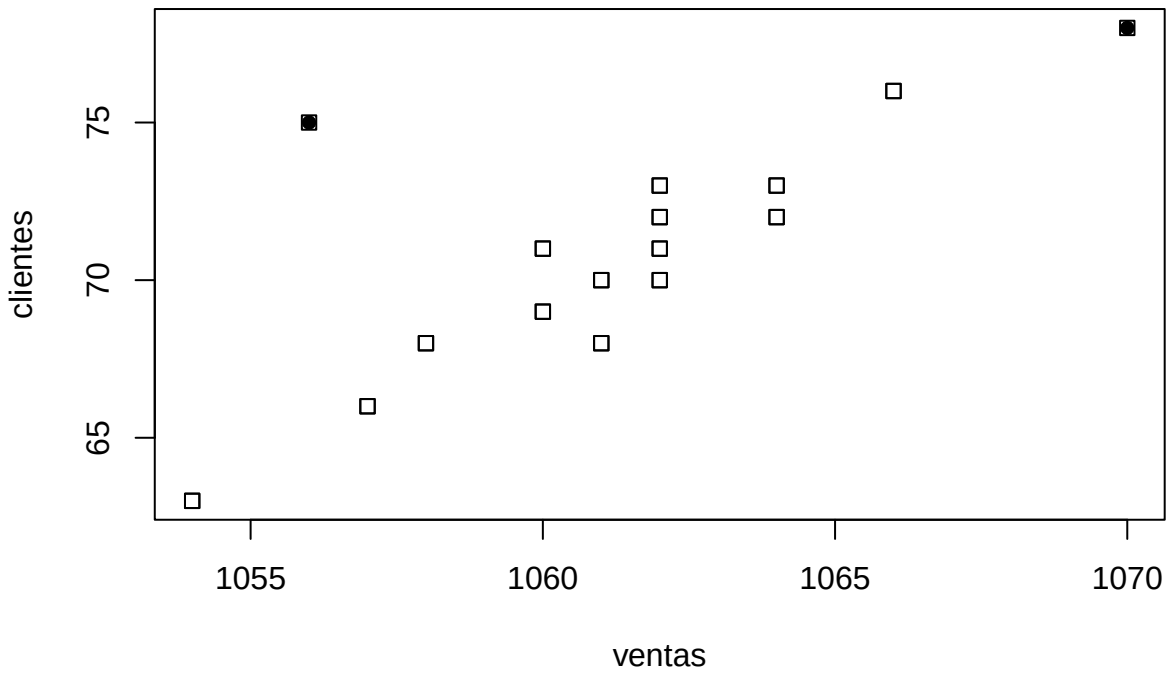
```
outlier2 <- rep(FALSE , nrow(datos))
outlier2[mah.ordenacion[1:num.outliers]] <- TRUE
```

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)
```



Ejercicio 2

```
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)
```

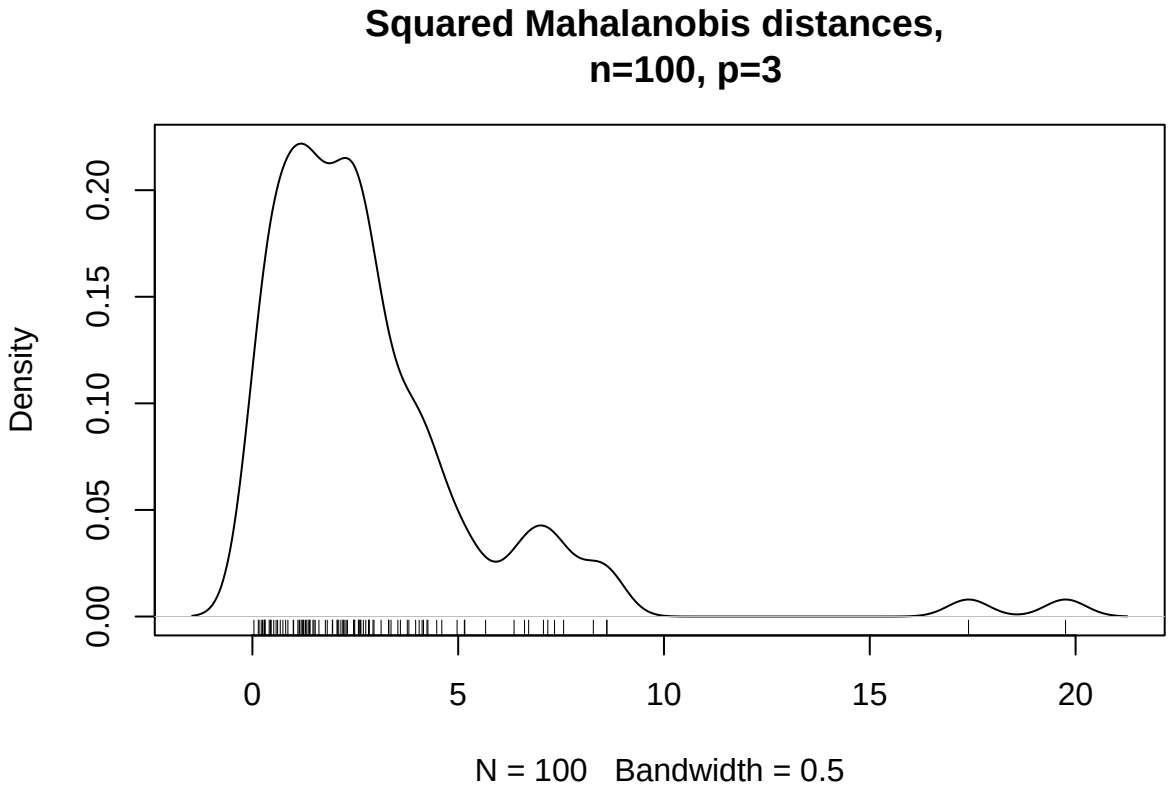
```
## [1] 5.37037
```

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

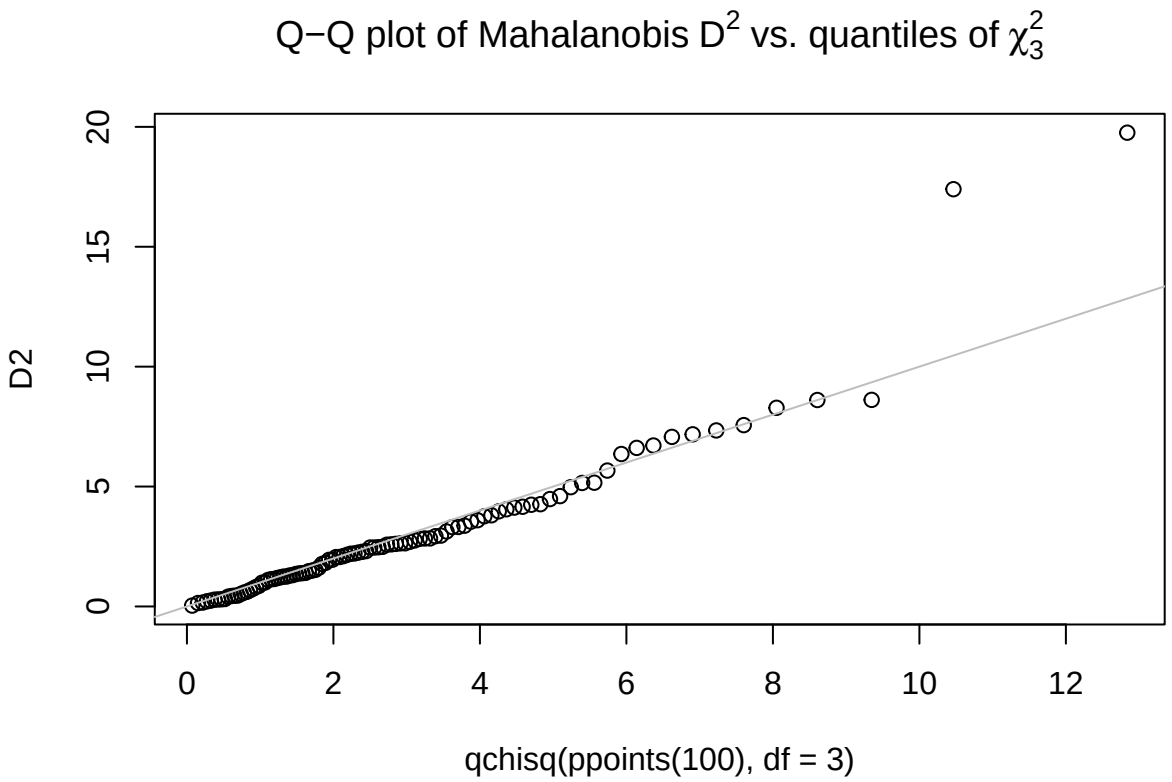
Here, D^2 = usual squared Euclidean distances

```
Sx <- cov(x)
D2 <- mahalanobis(x, colMeans(x), Sx)
```

```
plot(density(D2, bw = 0.5),
     main="Squared Mahalanobis distances,
     n=100, p=3") ; rug(D2)
```




```
qqplot(qchisq(ppoints(100), df = 3), D2,
       main = expression("Q-Q plot of Mahalanobis" * ~D^2 *
                         " vs. quantiles of" * ~ chi[3]^2))
abline(0, 1, col = 'gray')
```



Ejercicio 3

Diseñar un ejercicio utilizando la distancia de Mahalanobis.

se observaran las puntuaciones de una persona respecto a la calificacion de su comida, el timpo que tardo en prepararlos y el costo de la comida que va de 0 a 100 pesos

creacion de los datos

```
cm= data.frame(puntuacion =c(78,98,89,87,86,96,69,80,87,98),
                   minutos=c(30,20,21,16,45,34,23,24,26,23),
                     costo=c(88,90,67,98,78,67,90,80,79,82))
cm
```

```
##    puntuacion minutos costo
## 1          78      30    88
## 2          98      20    90
## 3          89      21    67
## 4          87      16    98
## 5          86      45    78
## 6          96      34    67
## 7          69      23    90
## 8          80      24    80
## 9          87      26    79
## 10         98      23    82
```

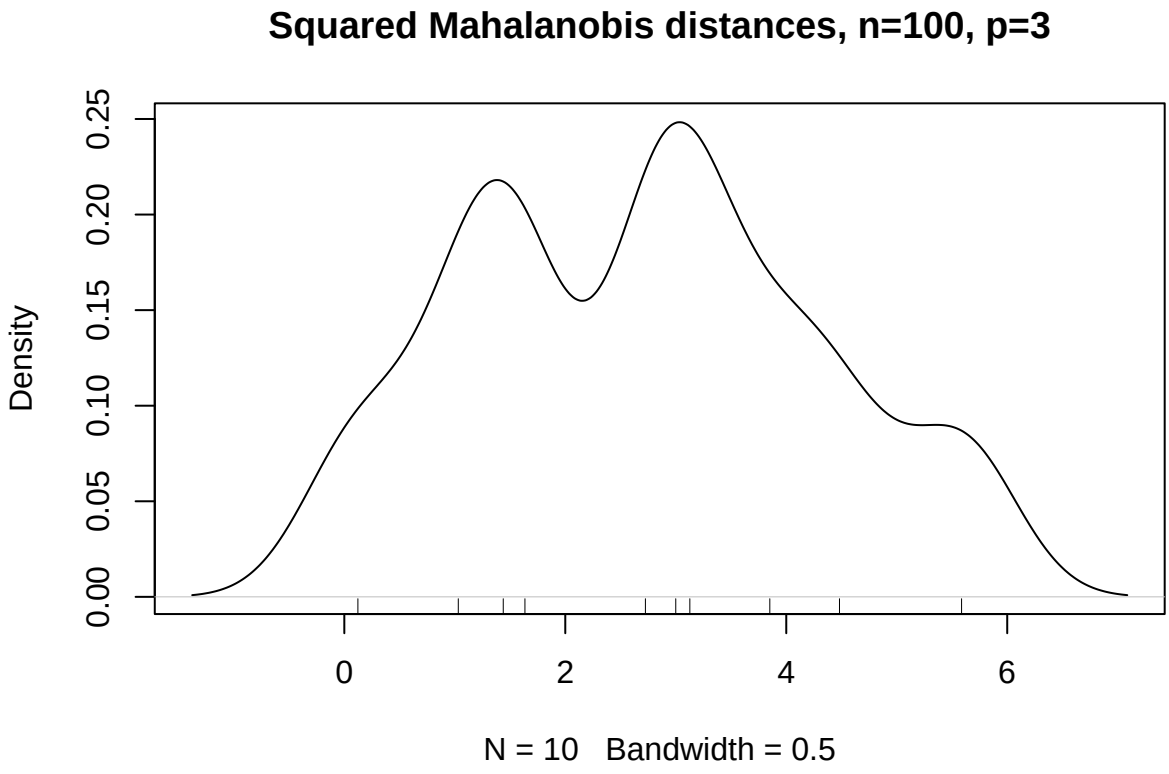
calcular las distanciad de mahalanobis para cada observación

```
D2= mahalanobis(cm,colMeans(cm),cov(cm))
```

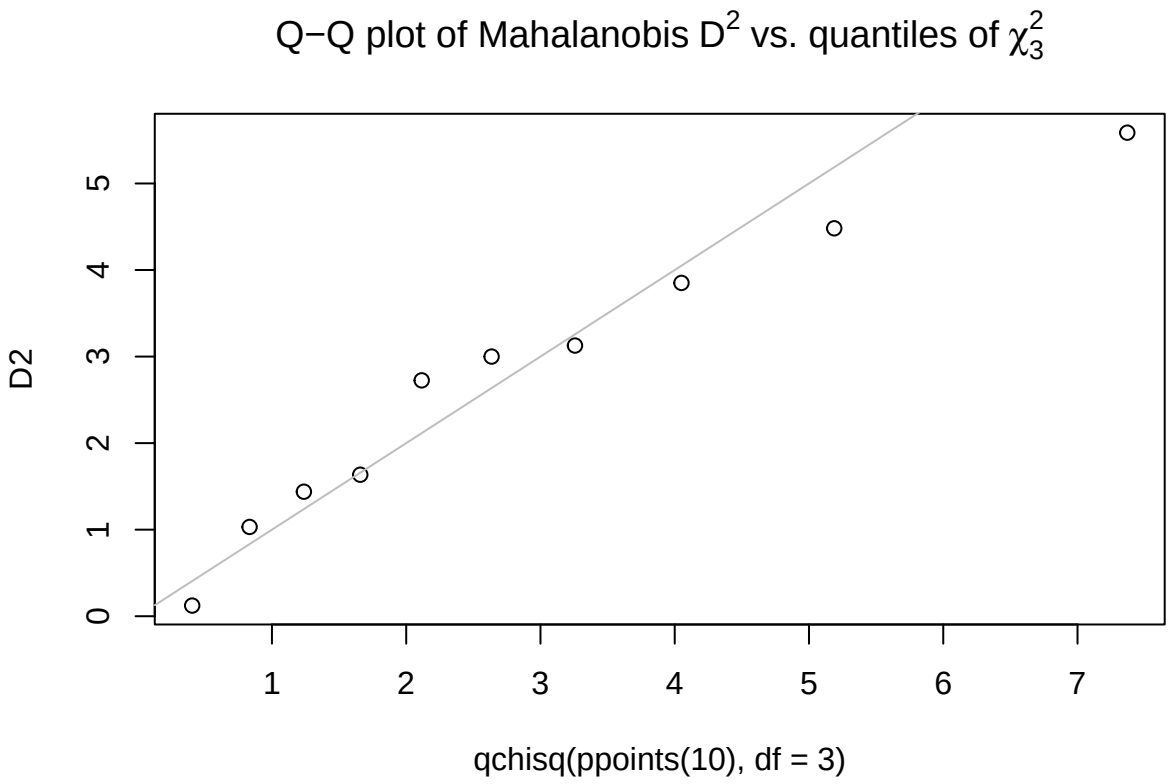
```
D2
```

```
##  [1] 1.4385830 2.9999759 4.4817841 3.1273018 5.5866793 2.7252444 3.8510558
##  [8] 1.0315759 0.1229968 1.6348031
```

```
     plot(density(D2, bw = 0.5),
          main="Squared Mahalanobis distances, n=100, p=3") ; rug(D2)
```




```
     qqplot(qchisq(ppoints(10), df = 3), D2,
            main = expression("Q-Q plot of Mahalanobis" * ~D^2 *
                              " vs. quantiles of" * ~ chi[3]^2))
abline(0, 1, col = 'gray')
```



Como se puede observar exicten datos que atipicos que no se distribuyen correctamente, por lo que afectaria nuestro estudio.