Distancia de Mahalanobis

Karina Itzel Rodríguez Conde

28/5/2022

Introducción

La distancia de Mahalanobis es una medida de distancia utilizada para determinar la similitud entre dos variables aleatorias multidimensionales.

Diego Calvo

1.- 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)
```

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

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

3.- Dimensión y tipo de variables

```
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 ...
```

4.- Resumen de los datos

```
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
```

5.- 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.

5.1.- Determinar el número de outlier que queremos encontrar.

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

5.2.- 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
```

5.3.- Generar un vector boleano con los dos valores más alejados según la distancia Mahalanobis.

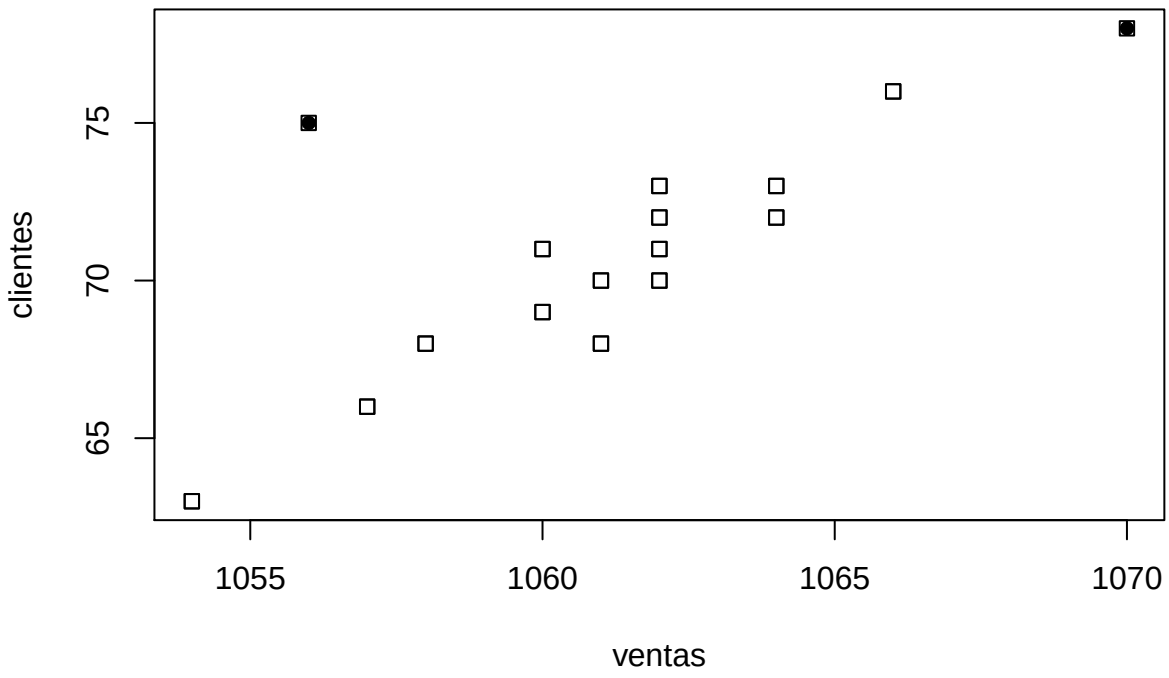
```
outlier2 <- rep(FALSE , nrow(datos))
outlier2[mah.ordenacion[1:num.outliers]] <- TRUE
```

6.- Resaltar con un punto relleno los 2 valores outliers.

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

7.- Visualizar el gráfico con los datos destacando sus outlier.

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



Proporcionado por Help R

```
require(graphics)
```

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

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

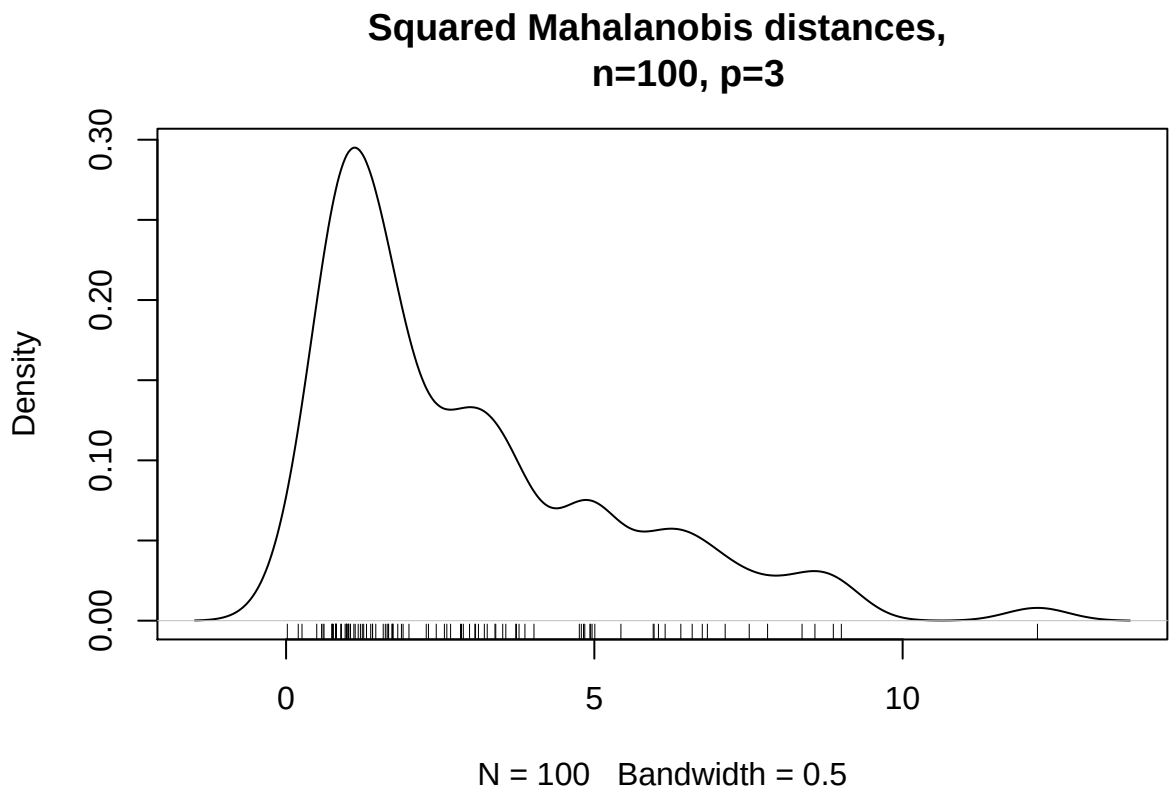
```
mahalanobis(c(0, 0), 1:2, S)
```

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

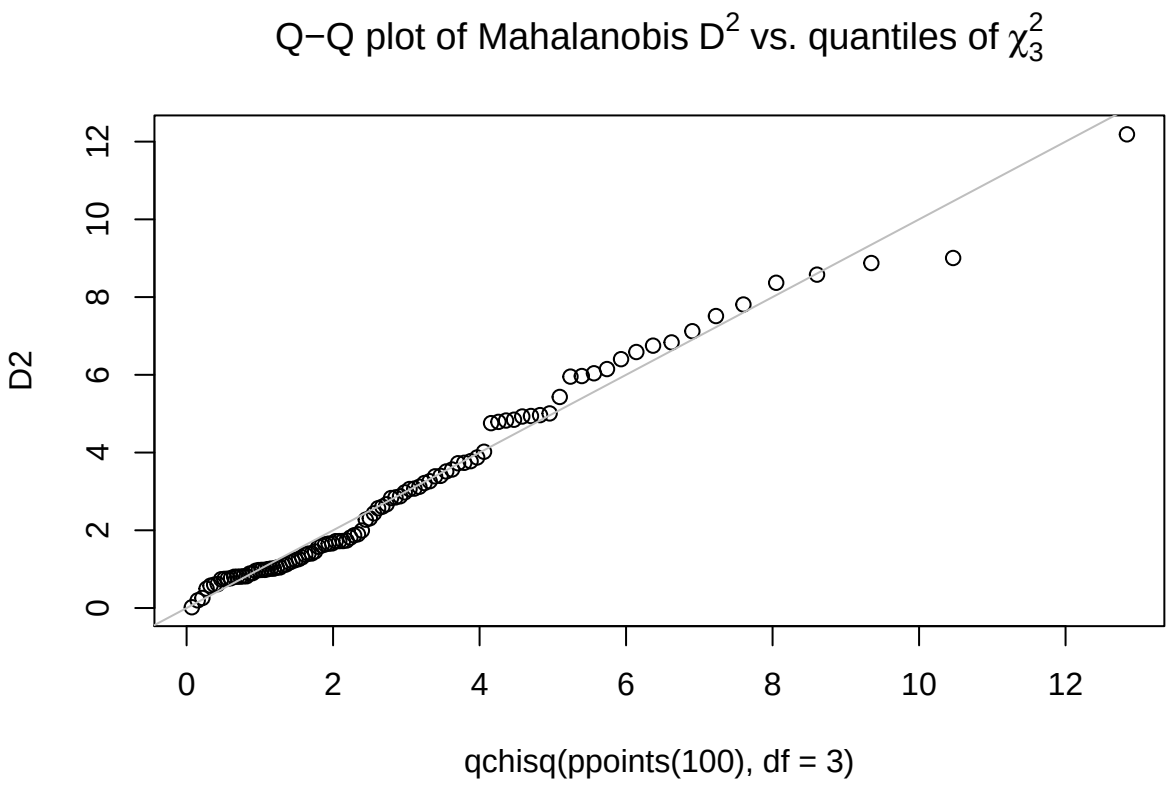
```
x <- matrix(rnorm(100*3), ncol = 3)
stopifnot(mahalanobis(x, 0,
                      diag(ncol(x))) == rowSums(x*x))
```

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')
```



Diseño de un ejercicio utilizando la distancia de Mahalanobis.

Se incluye:

1.- Planteamiento del problema (Cuál es el problema que se va a resolver).

Utilizando la matriz state.x77, se quiere saber la distancia de Mahalanobis de la población y su esperanza de vida. Qué expectativas de vida tiene la población de Estados Unidos.

2.- Datos simulados o matriz precargada en R.

Se utiliza la matriz precargada en R de state.x77

MÉTODO:

1.- Cargar los datos

```
datos <- as.data.frame(state.x77)
datos <- datos[,cbind(1,4)] #se utiliza la variable de Population y Life Exp.
```

2.- Dimensión y tipo de variables

```
dim(datos)
```

```
## [1] 50  2
```

```r
str(datos)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ Population: num  3615 365 2212 2110 21198 ...
##  $ Life Exp  : num  69 69.3 70.5 70.7 71.7 ...
```

3.- Resumen de los datos

```r
summary(datos)
```

```
##    Population       Life Exp
##  Min.   :  365   Min.   :67.96
##  1st Qu.: 1080   1st Qu.:70.12
##  Median : 2838   Median :70.67
##  Mean   : 4246   Mean   :70.88
##  3rd Qu.: 4968   3rd Qu.:71.89
##  Max.   :21198   Max.   :73.60
```

4.- 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.

5.1.- Determinar el número de outlier que queremos encontrar.

```r
num.outliers <- 10
```

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

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

```
##  [1]  5 32 40 24 11 43 10 38 28 13 44 18 34 23  2 35 27  1 16 15 33 48  7 49 41
## [26] 22  8 12 39 37 50 45  6  9 26 31 19 21 29 30 47 17 46 42  3  4 36 20 14 25
```

5.3.- Generar un vector boleano con los dos valores más alejados según la distancia Mahalanobis.

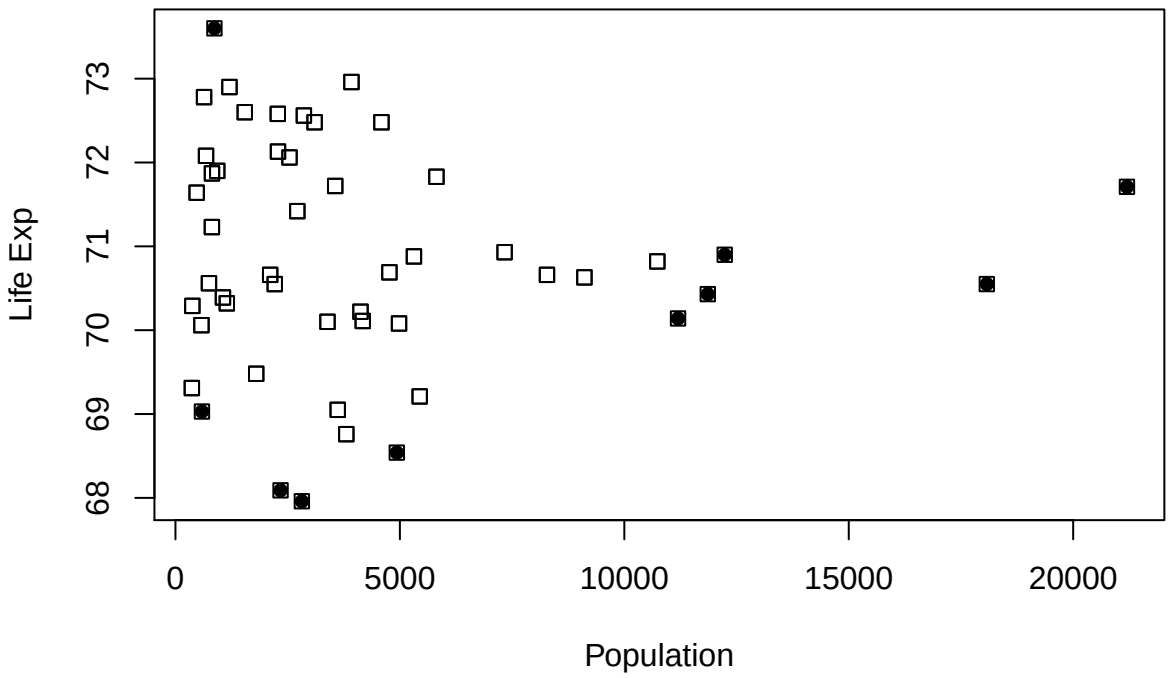
```r
outlier2 <- rep(FALSE , nrow(datos))
outlier2[mah.ordenacion[1:num.outliers]] <- TRUE
```

6.- Resaltar con un punto relleno los 10 valores outliers.

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

7.- Visualizar el gráfico con los datos destacando sus outlier.

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



3.- Interpretación. En el gráfico se identifican los outliers. Las distancias de cada uno están algo lejanas, hay outliers que se separan del resto de los datos y otros que, a pesar de que están cerca de la población, mantienen cierta lejanía. Pero la mayoría de los datos se mantiene cerca.