Act2 Matrices y vectores aleatorios C_IACD_Estadística

A01750164 | Paul Martín García Morfín

2022-10-04

Matrices y vectores aleatorios

Punto 1.

Considere la matriz de datos siguiente: $X = \begin{bmatrix} 1 & 4 & 3 \\ 6 & 2 & 6 \\ 8 & 3 & 3 \end{bmatrix}$ que consta de 3 observaciones (filas) y 3 variables (columnas)

$$b'X = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = X_1 + X_2 + X_3$$

$$c'X = \begin{bmatrix} 1 & 2 & -3 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = X_1 + 2X_2 - 3X_3$$

```
X = matrix(c(1, 6, 8, 4, 2, 3, 3, 6, 3), ncol=3)
b = matrix(c(1, 1, 1), ncol=1)
c = matrix(c(1, 2, -3), ncol=1)
tbX = t(b)%*%t(X)
tcX = t(c)%*%t(X)
```

a) Hallar la media, varianza y covarianza de b'X y c'X

```
M = matrix(c(tbX, tcX), ncol=2)
M = as.data.frame(M)
names(M) = c("b'X", "c'X")
n = 2
d = matrix(NA, ncol=2, nrow=n)
for(i in 1:n){
  d[i, ] = c(mean(M[, i]), var(M[, i]))
}
m = as.data.frame(d)
row.names(m) = c("b'X", "c'X")
names(m) = c("Media", "Varianza")
m
```

```
##     Media Varianza
## b'X    12       12
## c'X    -1       43
```

```
cov(M) # Matriz de varianza - covarianza
```

```
##     b'X c'X
## b'X  12  -3
## c'X  -3  43
```

b) Hallar el determinante de S (matriz de var-covarianzas de X)

```
det(cov(X))
```

```
## [1] 0
```

c) Hallar la matriz de varianzas-covarianzas (o porqué no se puede hallar)

```
S = cov(X)
S
```

```
##      [,1] [,2] [,3]
## [1,] 13.0 -2.5  1.5
## [2,] -2.5  1.0 -1.5
## [3,]  1.5 -1.5  3.0
```

c) Hallar los valores y vectores propios de S

```
eigen(S)
```

```
## eigen() decomposition
## $values
## [1]  1.379150e+01  3.208497e+00 -7.859007e-17
##
## $vectors
##            [,1]       [,2]       [,3]
## [1,]  0.9645458 -0.2295697 -0.1301889
## [2,] -0.2076189 -0.3555080 -0.9113224
## [3,]  0.1629288  0.9060418 -0.3905667
```

d) Argumentar si b'X y c'X son independientes o no

```
cor(M)
```

```
##            b'X        c'X
## b'X  1.0000000 -0.1320676
## c'X -0.1320676  1.0000000
```

Las variables tienen una baja correlación negativa, por otro lado, $s_{ik} \neq 0$ y $r_{ik} \neq 0$, así que no son independientes.

e) Hallar la varianza generalizada. Explicar el comportamiento de los datos de X basándose en los la variable generalizada, en los valores y vectores propios.

La varianza generalizada es el determinante de la matriz de varianzas y covarianzas S, para este caso, dicho valor es 0. Como propiedades tiene que:

- Está bien definida, ya que el determinante de la matriz S es igual que 0.
- Es una medida del volumen (k = 3) ocupado por el conjunto de datos.

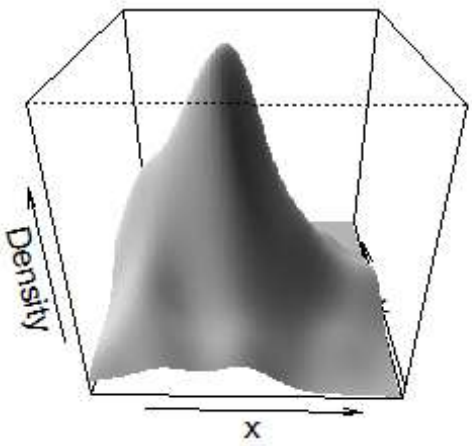
```
cor(X)

##            [,1]       [,2]       [,3]
## [1,]  1.0000000 -0.6933752  0.2401922
## [2,] -0.6933752  1.0000000 -0.8660254
## [3,]  0.2401922 -0.8660254  1.0000000
```

Las variables están relacionadas linealmente y el coeficiente de correlación es distinto de 0, así que el volumen ocupado se reduce debido a la presencia de correlación.

Punto 2.

Explore los resultados del siguiente código y dé una interpretación (se sugiere intersertarlo en un trozo de R en Rmarkdown para que dé varias ventanas de salida de resutados):

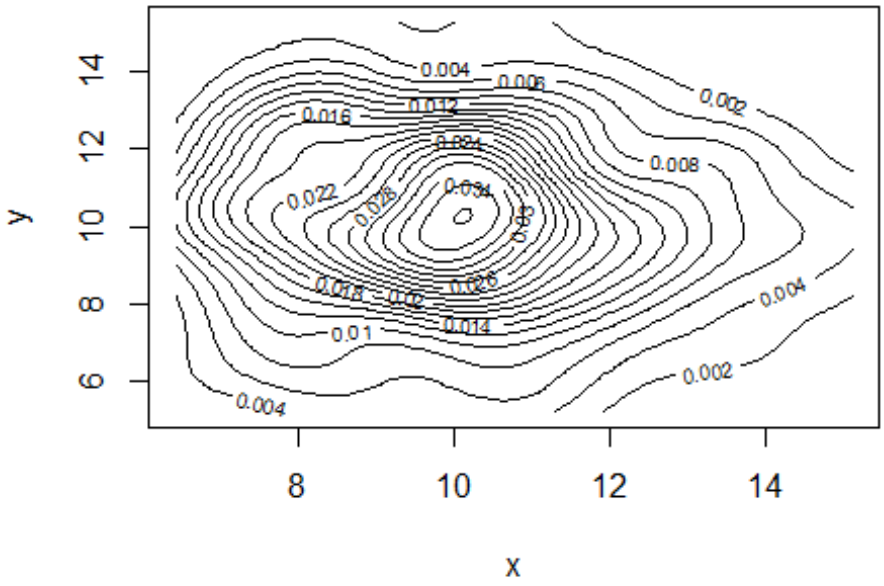
```
library(MVN)
x = rnorm(100, 10, 2)
y = rnorm(100, 10, 2)
datos = data.frame(x, y)
mvn(datos, mvnTest = "hz", multivariatePlot = "persp")
```




```
## $multivariateNormality
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.3097139 0.9477426 YES
##
## $univariateNormality
```

```
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling     x        0.5889    0.1216    YES
## 2 Anderson-Darling     y        0.3112    0.5470    YES
##
## $Descriptives
##     n      Mean  Std.Dev    Median      Min      Max     25th     75th
## x 100  9.886173 1.915586  9.953151 6.430227 15.11195 8.372673 10.89561
## y 100 10.114667 2.091789 10.078900 5.228911 15.25345 8.670948 11.63443
##          Skew    Kurtosis
## x  0.40918720 -0.25033308
## y -0.04499271 -0.05783738
```

```
mvn(datos, mvnTest = "hz", multivariatePlot = "contour")
```




```
## $multivariateNormality
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.3097139 0.9477426 YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling     x        0.5889    0.1216    YES
## 2 Anderson-Darling     y        0.3112    0.5470    YES
##
## $Descriptives
##     n      Mean  Std.Dev    Median      Min      Max     25th     75th
## x 100  9.886173 1.915586  9.953151 6.430227 15.11195 8.372673 10.89561
## y 100 10.114667 2.091789 10.078900 5.228911 15.25345 8.670948 11.63443
```

```
##          Skew    Kurtosis
## x  0.40918720 -0.25033308
## y -0.04499271 -0.05783738
```

Se usa la prueba de Henze-Zirkler, la cual está basada en la distancia funcional no negativa, es decir, mide la distancia entre dos funciones de distribución. Si los datos presentan una distribución normal multivariada, la prueba estadística se distribuye aproximadamente como una lognormal. En este caso se cumple con normalidad multivariante ya que el p-value > 0.05.

El estadístico Anderson-Darling mide qué tan bien siguen los datos una distribución específica. En este caso se evalua normalidad. En esta prueba se tiene un p-value > 0.05, así que se acepta H (normalidad para X y para Y).

Se tiene un bajo sesgo para ambas variables (hay simetría), además, por los valores de curtosis podemos decir que tiene una forma mesocúrtica.

Punto 3.

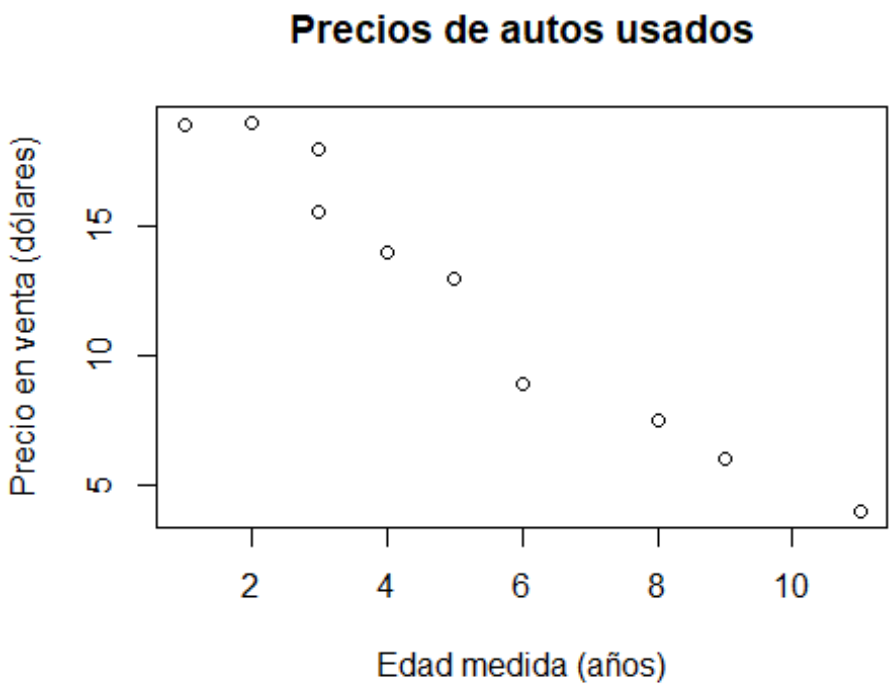
Un periódico matutino enumera los siguientes precios de autos usados para un compacto extranjero con edad medida en años y precio en venta medido en miles de dólares.

x1: 1, 2, 3, 3, 4, 5, 6, 8, 9, 11

x2: 18.95, 19.00, 17.95, 15.54, 14.00, 12.95, 8.94, 7.49, 6.00, 3.99

a) Construya un diagrama de dispersión

```
x1 = c(1, 2, 3, 3, 4, 5, 6, 8, 9, 11)
x2 = c(18.95, 19.00, 17.95, 15.54, 14.00, 12.95, 8.94, 7.49, 6.00, 3.99)
plot(x=x1, y=x2,  main = "Precios de autos usados", xlab = "Edad medida
(años)", ylab = "Precio en venta (dólares)")
```



b) Inferir el signo de la covarianza muestral a partir del gráfico.

Por la pendiente que se forma en los datos, la covarianza tendrá un signo negativo.

```
datos = data.frame(x1, x2)
S = cov(datos)
S
```

```
##           x1        x2
## x1  10.62222 -17.71022
## x2 -17.71022  30.85437
```

c) Calcular el cuadrado de las distancias estadísticas

$(x_j - \bar{x})' S^{-1} (x_j - \bar{x})$ con $x'_j = [x_{j1}, x_{j2}]$. Nota: para el cálculo de la distancia de Mahalanobis, usa: mahalanobis(A,medias,S).

```
medias = colMeans(datos)
d2M = mahalanobis(datos, medias, S)
d2M
```

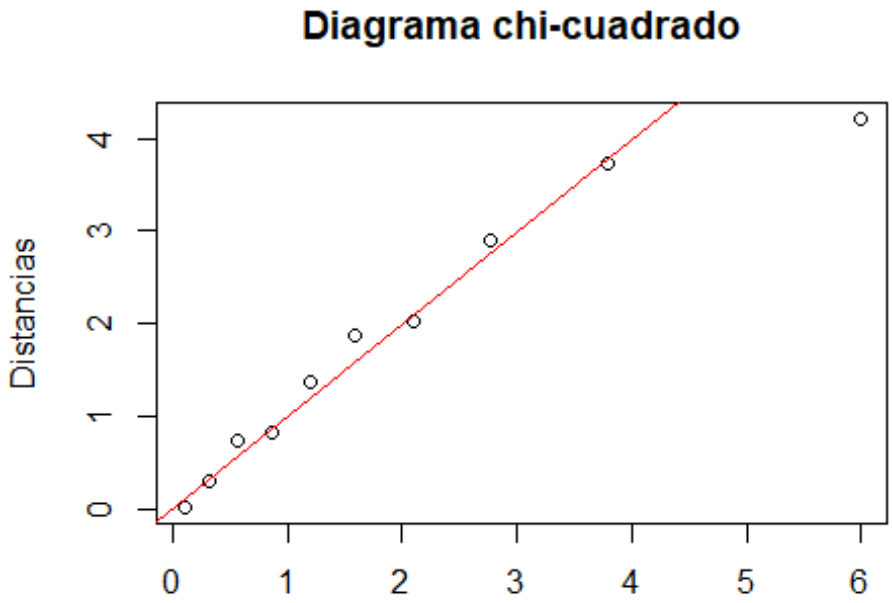
```
##  [1] 1.8753045 2.0203262 2.9009088 0.7352659 0.3105192 0.0176162
3.7329012
##  [8] 0.8165401 1.3753379 4.2152799
```

d) Usando las anteriores distancias, determine la proporción de las observaciones que caen dentro del contorno de probabilidad estimado del 50% de una distribución normal bivariada.

```
length(d2M[d2M <= qchisq(0.5, df=2)])/length(d2M)

## [1] 0.5
```

e) Ordene las distancias del inciso c y construya un diagrama chi-cuadrado

```
plot(qchisq(((1:nrow(datos))-1/2)/nrow(datos), df=2), sort(d2M), xlab="",
ylab="Distancias", main="Diagrama chi-cuadrado")
abline(a=0, b=1, col="red")
```



f) Dados los resultados anteriores, ¿serían argumentos para decir que los datos son aproximadamente normales bivariados?

Se podrían obtener algunas conclusiones preeliminares, sin embargo, considero que haría falta hacer las pruebas de normalidad, con cálculos de probabilidad y contornos de densidad para llegar a una conclusión más acertada acerca de la normalidad bivariada.