Actividad 1.5 Combinaciones lineales y Validación de la Normal Multivariada

Raúl Correa Ocañas

2024-08-17

Ejercicio 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).

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

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

a) Hallar la media, varianza y covarianza de X

```
X = matrix(c(1, 4, 3, 6, 2, 6, 8, 3, 3), nrow = 3, byrow = TRUE)

print("Media de X")
```

```
## [1] "Media de X"
```

```
colMeans(X)
```

```
## [1] 5 3 4
```

```
print("Varianzas y Covarianzas de X")
```

```
## [1] "Varianzas y Covarianzas de X"
```

```
cov(X)
```

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

```
print("Estimador Insesgado")
```

```
## [1] "Estimador Insesgado"
```

b) Hallar la media, varianza y covarianza de b^TX y c^TX

```
bX = matrix(X[,1] + X[,2] + X[,3])
cX = matrix(X[,1] + 2*X[,2] - 3*X[,3])
```

```
print("Media de b'X")
```

```
## [1] "Media de b'X"
```

```
print(mean(bX))
```

```
## [1] 12
```

```
print("Media de c'X")
```

```
## [1] "Media de c'X"
```

```
print(mean(cX))
```

```
## [1] -1
```

```
print("Covarianza de b'X")
```

```
## [1] "Covarianza de b'X"
```

```
print(cov(bX))
```

```
##      [,1]
## [1,]   12
```

```
print("Covarianza de c'X")
```

```
## [1] "Covarianza de c'X"
```

```
print(cov(cX))
```

```
##      [,1]
## [1,]   43
```

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

```
det_S = det(cov(X))
print("Determinante de S")
```

```
## [1] "Determinante de S"
```

```
print(det_S)
```

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

d) Hallar los valores y vectores propios de S

```
eigen_S = eigen(cov(X))

print("Valores propios de S")
```

```
## [1] "Valores propios de S"
```

```
print(eigen_S$values)
```

```
## [1]  1.379150e+01  3.208497e+00 -7.859007e-17
```

```
print("Vectores propios de S")
```

```
## [1] "Vectores propios de S"
```

```
print(eigen_S$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
```

e) Argumentar si hay independencia entre b'X y c'X , ¿y qué ocurre con X1, X2 y X3? ¿son independientes?

```
cov(matrix(c(bX,cX), nrow = 3, byrow = FALSE))
```

```
##      [,1] [,2]
## [1,]   12   -3
## [2,]   -3   43
```

```
print("Varianza Generalizada entre b'X y c'X")
```

```
## [1] "Varianza Generalizada entre b'X y c'X"
```

```
print(det(cov(matrix(c(bX,cX), nrow = 3, byrow = FALSE))))
```

```
## [1] 507
```

```
print("Varianza Generalizada entre X1, X2, X3")
```

```
## [1] "Varianza Generalizada entre X1, X2, X3"
```

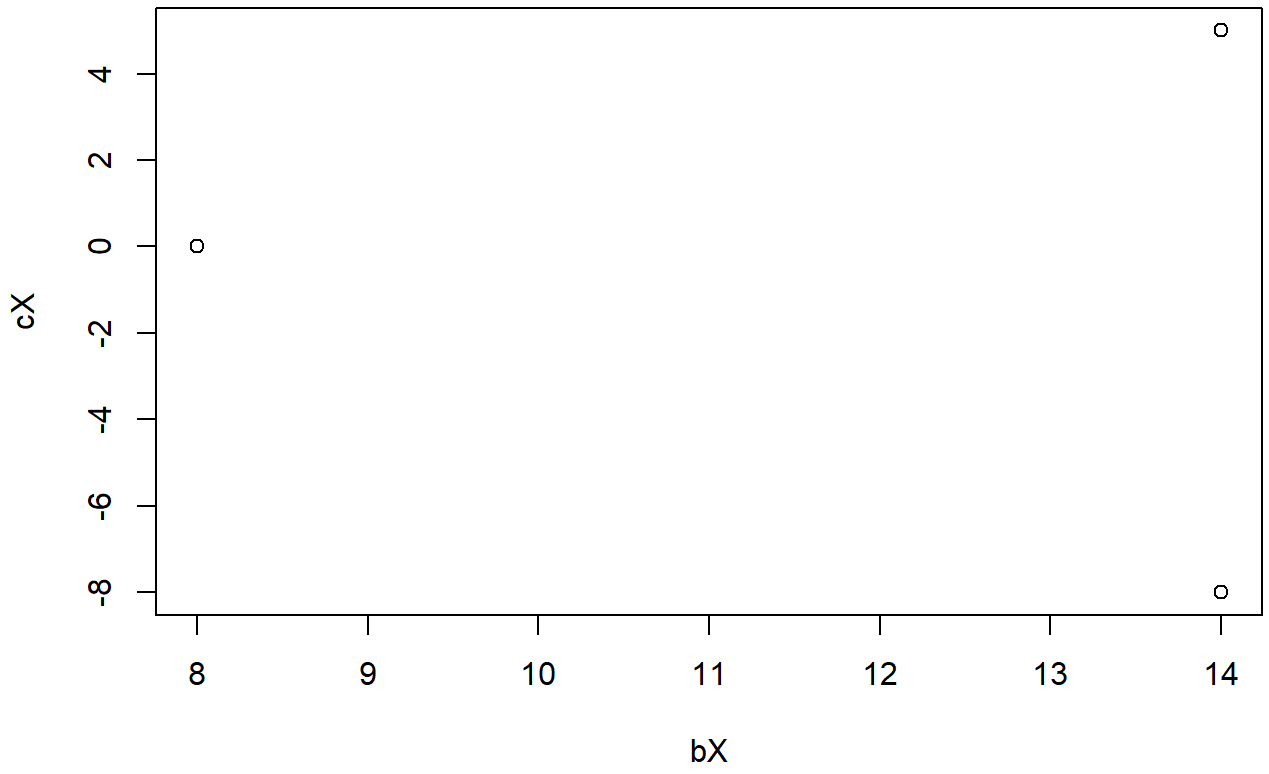
```
print(det(cov(X)))
```

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

Dos variables unicamente son consideradas como independientes cuando su covarianza es 0 y sus respectivas distribuciones son normales. La covarianza entre b^TX y c^TX es -3 y la varianza generalizada es de 507. Si b^TX y c^TX fuesen variables independientes, su covarianza sería de 0 y su varianza generalizada sería de 516 (ya que sería igual a la multiplicación de var(b^TX) y var(c^TX)). Esto sugiere sospechas de independencia entre las variables en el caso de que la muestra de las variables sean normales.

Por otro lado, las variables X_1, X_2, X_3 tienen una varianza generalizada de 0, por lo que se sabe con certeza que alguna de las variables es una combinación lineal del resto, por lo que no pueden ser independientes.

```
plot(bX, cX)
```



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

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

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

Dado que uno de los valores propios es casi 0, sugiriendo que una de las variables casí no aporta a la variabilidad de los datos. Esto combinado con el hecho de tener una varianza generalizada de 0 indica que efectivamente una de las variables es linealmente dependiente de otra o más variables.

Ejercicio 2.

Explore los resultados del siguiente código y dé una interpretación.

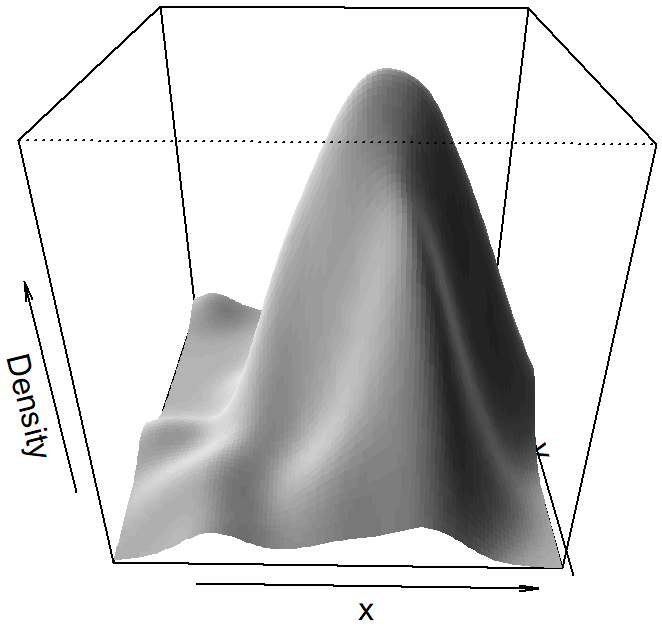
```
x = rnorm(100, 10, 2)
y = rnorm(100, 10, 2)

datos = data.frame(x,y)
datos
```

```
##             x         y
## 1   12.741917 12.401931
## 2    8.870604 12.089502
## 3   10.726257  7.993583
## 4   11.265725 13.696964
## 5   10.808537  8.666453
## 6    9.787751 10.211028
## 7   13.023044  9.155488
## 8    9.810682  9.755300
## 9   14.036847 10.376386
## 10   9.874572 10.238322
## 11  12.609739  9.949815
## 12  14.573291 10.216145
## 13   7.222279  9.029130
## 14   9.442422  8.991566
## 15   9.733357  6.677802
## 16  11.271901  9.235333
## 17   9.431494  8.974699
## 18   4.687089 15.403782
## 19   5.119066  7.275768
## 20  12.640227 10.274512
## 21   9.386723  7.012750
## 22   6.437383  7.059129
## 23   9.656165 10.249405
## 24  12.429349  8.006722
## 25  13.790387  9.996355
## 26   9.139062  9.143482
## 27   9.485461  8.772657
## 28   6.473674  5.950644
## 29  10.920195  7.550504
## 30   8.720010 10.359033
## 31  10.910900 11.135241
## 32  11.409675  9.014245
## 33  12.070207 10.000126
## 34   8.782147 12.245779
## 35  11.009910 12.879711
## 36   6.565983  7.805772
## 37   8.431082  9.765361
## 38   8.298185 12.402997
## 39   5.171585  9.060541
## 40  10.072245  9.895061
## 41  10.411997  9.827785
## 42   9.277885  8.224642
## 43  11.516326  9.110632
## 44   8.546590  9.941110
## 45   7.263438  9.172262
## 46  10.865636 12.226772
## 47   8.377214  9.038014
## 48  12.888203  9.133662
## 49   9.137108 11.393725
## 50  11.311296  7.887263
## 51  10.643851  9.918603
```

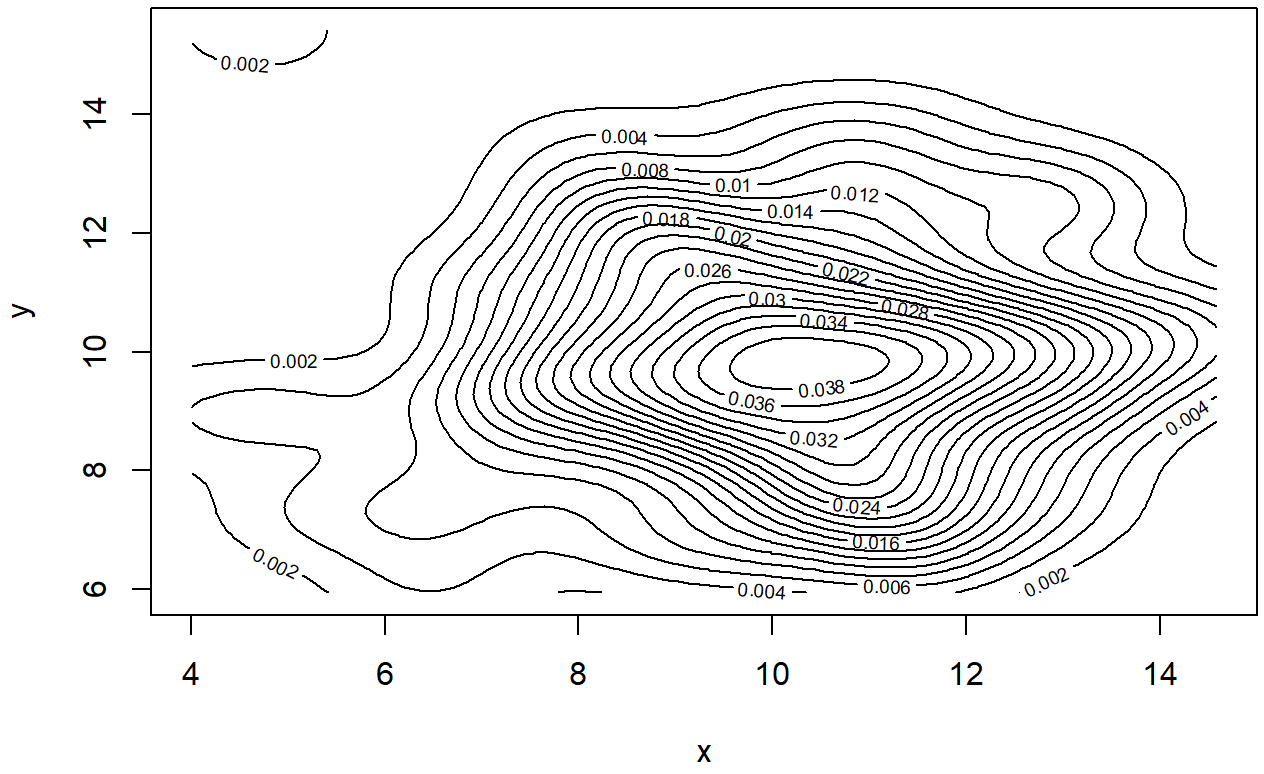
```
## 52   8.432322  6.896910
## 53  13.151455 12.334339
## 54  11.285799  9.452709
## 55  10.179521  9.064309
## 56  10.553101  7.523495
## 57  11.358578  9.984476
## 58  10.179666  8.399436
## 59   4.013820  8.933015
## 60  10.569766 12.575350
## 61   9.265531  9.648948
## 62  10.370461  7.856435
## 63  11.163647 10.326414
## 64  12.799474  9.274523
## 65   8.545416 11.180027
## 66  12.605085 12.864844
## 67  10.671696  8.014615
## 68  12.077012 10.909301
## 69  11.841457 10.169796
## 70  11.441756 11.791131
## 71   7.913762  9.540444
## 72   9.819627 11.673238
## 73  11.247036  6.509888
## 74   8.092953 13.378918
## 75   8.914342 11.729556
## 76  11.161993  9.698448
## 77  11.536357  7.101986
## 78  10.927535 11.286017
## 79   8.228447 10.966388
## 80   7.800438  9.987289
## 81  13.025414 10.302912
## 82  10.515843  8.831782
## 83  10.176880 10.737613
## 84   9.758207 10.589309
## 85   7.611342  9.441481
## 86  11.223994  7.327527
## 87   9.565720 11.401498
## 88   9.634487 11.108393
## 89  11.866693  8.327387
## 90  11.643546  6.810824
## 91  12.784233 10.409917
## 92   9.047652  9.309824
## 93  11.300697 10.505223
## 94  12.782221  7.411995
## 95   7.778422  8.081659
## 96   8.278415 12.171550
## 97   7.736523 10.807550
## 98   7.081572 11.172975
## 99  10.159965 13.630457
## 100 11.306409 10.257643
```

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




```
## $multivariateNormality
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.4456005 0.7182024 YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling     x        0.4619    0.2533    YES
## 2 Anderson-Darling     y        0.3549    0.4542    YES
##
## $Descriptives
##     n      Mean  Std.Dev    Median      Min      Max     25th     75th
## x 100 10.065030 2.082714 10.179594 4.013820 14.57329 8.766613 11.32312
## y 100  9.825033 1.808347  9.861423 5.950644 15.40378 8.817001 10.92357
##         Skew    Kurtosis
## x -0.4689891  0.14533218
## y  0.3044610 -0.02820991
```

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




```
## $multivariateNormality
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.4456005 0.7182024 YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling     x        0.4619    0.2533    YES
## 2 Anderson-Darling     y        0.3549    0.4542    YES
##
## $Descriptives
##     n      Mean  Std.Dev    Median      Min      Max     25th     75th
## x 100 10.065030 2.082714 10.179594 4.013820 14.57329 8.766613 11.32312
## y 100  9.825033 1.808347  9.861423 5.950644 15.40378 8.817001 10.92357
##         Skew    Kurtosis
## x -0.4689891  0.14533218
## y  0.3044610 -0.02820991
```

La prueba de Henze-Zirkler es una prueba de normalidad multivariada en la que su hipótesis nula es que la distribución es normal. Al tener un valor de p de 0.718, no se tiene suficiente evidencia estadística para rechazar la hipótesis nula y por lo tanto se infiere que la distribución conjunta de las variables efectivamente es normal. También se hicieron pruebas de Anderson-Darling para cada variable para probar sus respectivas normalidades univariadas. Estas tuvieron valores p 0.462 y 0.355, ambas sin tener suficiente evidencia estadistica para rechazar la hipótesis nula. Por ende se puede inferir normalidad univariada para cada variable. La curtosis es cercana a cero para cada variable, y sus sesgos son relativamente cercanos a cero también. Observando las

gráficas, la forma de la distribución efectivamente parece normal bivariada y sus contornos confirman este comportamiento. Por lo tanto, se puede inferir que se cumplen los supuestos de la normal multivariada y por lo tanto la distribución conjunta es normal.

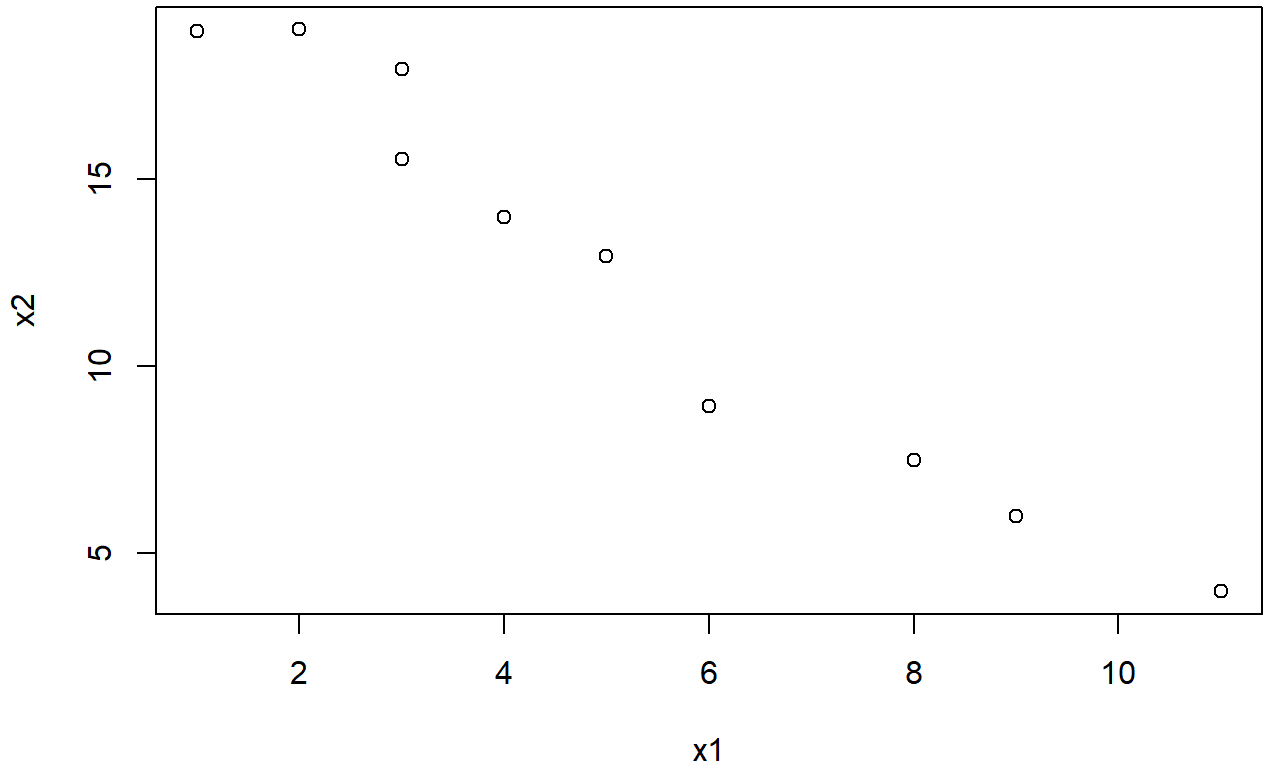
Ejercicio 3.

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 = 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)
```

a) Construya un diagrama de dispersión

```
plot(x1, x2)
```



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

La covarianza de los datos tiene que ser negativa, debido a que se observa una relación negativa (mientras X1 incrementa, X2 decrementa y viceversa).

c) Calcular el cuadrado de las distancias de Mahalanobis

```
X = matrix(c(x1,x2), ncol = 2, byrow = FALSE)
mu = colMeans(X)

sigma = cov(X)

mahalanobis_dist = mahalanobis(X, center = mu, cov = sigma)
mahalanobis_dist
```

```
##  [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.

```
chi_sq_crit = qchisq(0.5, df = 2)

is_in_contour = (mahalanobis_dist <= chi_sq_crit)
mean(is_in_contour)
```

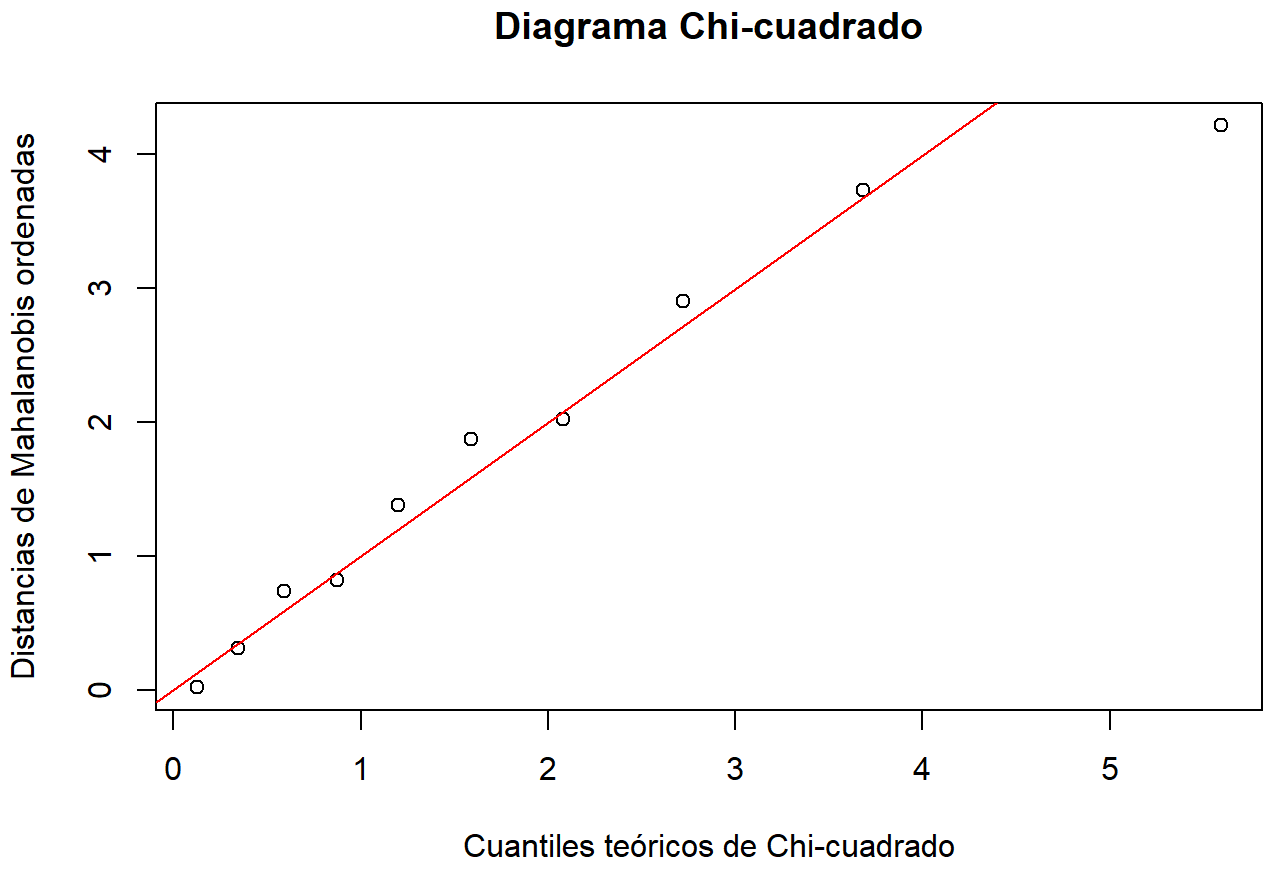
```
## [1] 0.5
```

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

```
mahalanobis_dist_sorted = sort(mahalanobis_dist)

chi_square_quantiles = qchisq(ppoints(length(mahalanobis_dist_sorted)), df = 2)

plot(chi_square_quantiles, mahalanobis_dist_sorted,
     main = "Diagrama Chi-cuadrado",
     xlab = "Cuantiles teóricos de Chi-cuadrado",
     ylab = "Distancias de Mahalanobis ordenadas")
abline(0, 1, col = "red")
```



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

Según lo observado en la gráfica de QQPlot, se tienen fuertes sospechas de que la data es normalmente bivariada. Los puntos están cerca a la linea que sería lo esperado de una distribución normal bivariada, y al comparar si el 50% de los datos observados correspondian a lo esperado en la teoría, se tuvieron resultados que afirman la sospecha. Sería más convincente hacer pruebas de normalidad multivariada para afirmar la sospecha.

Ejercicio 4.

Si X es un vector aleatorio con X_1, X_2, X_3 son tres variables conjuntamente normales, no independientes, con b, un vector de 3 constantes, b_1, b_2, b_3, y c, otro vector de 3 constantes, c_1, c_2, c_3, demuestra que las variables $V_1 = b^{'}X$ y $V_2 = c^{'}X$ son independientes si $b^{'}c = 0$.

Recordemos que $E(X) = E(X)$ donde α es un vector escalar. Si $Cov(b^TX, c^TX) = \frac{1}{n}(b^TX - E(b^TX)(c^TX - E(c^TX))$, entonces $Cov(b^TX, c^TX) = \frac{1}{n}(b^TX - b^TE(X))(c^TX - c^TE(X))$. Factorizando las constantes:

$$Cov(b^TX, c^TX) = \frac{1}{n}b^T(X - E(X))^T(X - E(X))c = b^TCov(X)c$$

Bien se sabe que la matriz de covarianzas es simétrica, por lo que el producto $b^TCov(X)c$ debe dar 0. Recordando que las variables son conjuntamente normales, y observando que la covarianza entre V_1, V_2 es cero, hemos demostrado que son independientes.