UNIVERSIDAD NACIONAL DE COLOMBIA

MAESTRÍA EN CIENCIAS ESTADÍSTICA

Departamento de Estadística
Facultad de Ciencias

Análisis Multivariado de Datos

Integrantes:

Luis David Hernández Pérez	C.C. 1193549963
Daniel Felipe Villa Rengifo	C.C. 1005087556

Medellín, Colombia
Semestre 2024-02

Medellín, Enero 31 de 2025

Tabla de contenidos

Los datos utilizados son los pertenecientes al equipo-07

Punto-01:

Considere la matriz de datos asignada, la cual corresponde a un conjunto de datos simulados de un vector $\mathbf{x}$ normal 6-variado con parámetros dados por:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 4 & 0 & 2 & 0 & 1 & 0 \\ 0 & 9 & 0 & 3 & 0 & 2 \\ 2 & 0 & 5 & 0 & 4 & 0 \\ 0 & 3 & 0 & 8 & 0 & 1 \\ 1 & 0 & 4 & 0 & 6 & 0 \\ 0 & 2 & 0 & 1 & 0 & 7 \end{bmatrix}$$

Particione $\mathbf{x}$ como sigue:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix}, \quad \text{donde:} \quad \mathbf{x}^{(1)} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}, \quad \mathbf{x}^{(2)} = \begin{bmatrix} X_4 \\ X_5 \\ X_6 \end{bmatrix}$$

(a) Realice una verificación de la normalidad: uni-variada, bi-variada y 3-variada de los datos asociados a $\mathbf{x}^{(1)}$.

Nota: Utilizar algunas de las herramientas vistas en clase sobre procesos de evaluación de la normalidad multivariada y/o herramientas que usted conozca para dichos procesos.

(b) ¿Cuáles son los estimadores de Máxima Verosimilitud de $\mu^{(1)} = \mathbb{E}[\mathbf{x}^{(1)}]$ y de $\Sigma_{11} = \mathrm{Var}(\mathbf{x}^{(1)})$?

(c) Considere la variable definida por $Y = \mathbf{a}^T\mathbf{x}^{(2)}$, con $\mathbf{a} = \begin{bmatrix} 1 & 2 & -1 \end{bmatrix}^T$:

- Obtenga los datos muéstrales (o puntuaciones) asociados a la variable Y.
- Realice la verificación de normalidad uni-variada de los datos asociados a Y.

(d) Considere el vector definido por $\mathbf{y} = \mathbf{A}\mathbf{x}^{(1)}$, con $\mathbf{A} = \begin{bmatrix} 0 & 1 & 2 \\ 2 & 0 & -1 \end{bmatrix}$:

- Obtenga los datos muestrales (o puntuaciones) asociados al vector $\mathbf{y}$.
- Realice la verificación de normalidad bi-variada de los datos asociados a $\mathbf{y}$.

Solución Punto-01

Solución (a)

```
datos_01_y_02 <- read.table("datos_puntos_01_02.txt", header = T)

# Separar los conjuntos de datos

x1 <- as.matrix(datos_01_y_02[, 1:3])  # Variables V1, V2, V3
x2 <- as.matrix(datos_01_y_02[, 4:6])  # Variables V4, V5, V6
```

Primeramente verificaremos la normalidad uni-variada por medio de la prueba de Shapiro-Wilk a los datos asociados a $X^{(1)}$.

```
# Normalidad uni-variada
apply(x1, 2, function(col) shapiro.test(col))
```

```
$V1

	Shapiro-Wilk normality test

data:  col
W = 0.99036, p-value = 0.7069


$V2

	Shapiro-Wilk normality test

data:  col
W = 0.9904, p-value = 0.7094


$V3

	Shapiro-Wilk normality test

data:  col
W = 0.99186, p-value = 0.8206
```

Las pruebas de normalidad uni-variada para las variables V_1 , V_2 y V_3 no muestran evidencia suficiente para rechazar la hipótesis de normalidad. Por lo tanto, estas variables son consistentes con una distribución normal.

```r
# Normalidad bi-variada
biv_pairs <- combn(1:3, 2, simplify = FALSE)
for (pair in biv_pairs) {
  cat(sprintf("Variables: V%d y V%d\n", pair[1], pair[2]))
  print(mshapiro.test(t(x1[, pair])))
}
```

```
Variables: V1 y V2

	Shapiro-Wilk normality test

data:  Z
W = 0.98749, p-value = 0.4862

Variables: V1 y V3

	Shapiro-Wilk normality test

data:  Z
W = 0.9933, p-value = 0.9113

Variables: V2 y V3

	Shapiro-Wilk normality test

data:  Z
W = 0.98521, p-value = 0.3424
```

Para los pares de variables analizados ($V1-V2, V1-V3, V2-V3$), no hay evidencia suficiente para rechazar la hipótesis de normalidad. Por lo tanto, se considera que las combinaciones bi-variadas de $x(1)$ son consistentes con una distribución normal.

Ahora verificaremos verificaremos la normalidad 3-variada de los datos asociados a $X^{(1)}$.

```r
# Normalidad 3-variada con el test de Mardia
Mardia_x1 <- mvn(x1, mvnTest = "mardia")
Mardia_x1$multivariateNormality
```

```
             Test          Statistic           p value Result
1 Mardia Skewness   9.28226670724539 0.505541420304724    YES
2 Mardia Kurtosis -0.406880365335198 0.684095857551414    YES
3             MVN               <NA>              <NA>    YES
```

- Para la asimetría, el p-valor alto (0.5055) indica que no hay evidencia suficiente para rechazar la normalidad.
- Para la curtosis, el p-valor alto (0.6841) también sugiere que los datos cumplen con la normalidad.

- La conclusión general del test (MVN) confirma que los datos en $x(1)$ son consistentes con una distribución normal multivariada.

Solución (b)

Los estimadores de maxima verosimilitud para

$$\hat{\mu} = \underline{\bar{\mathbf{x}}}^{(1)} = \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \end{bmatrix} = \begin{bmatrix} 0.06616429 \\ -0.16939592 \\ -0.34023163 \end{bmatrix} \quad y$$

$$\hat{\Sigma}_{11} = S_n = \frac{1}{n}\sum_{i=1}^{n}\left(\underline{\mathbf{x}}^{(1)} - \underline{\bar{\mathbf{x}}}^{(1)}\right)\left(\underline{\mathbf{x}}^{(1)} - \underline{\bar{\mathbf{x}}}^{(1)}\right)' = \begin{bmatrix} 4.369791 & -0.4407255 & 2.435705 \\ -0.4407255 & 9.9162583 & -1.465279 \\ 2.4357049 & -1.4652786 & 5.628944 \end{bmatrix}$$

```
# Estimador de la media muestral (Máxima Verosimilitud de µ(1))
mu_1_hat <- colMeans(x1)
print(mu_1_hat)
```

```
         V1          V2          V3
 0.06616429 -0.16939592 -0.34023163
```

```
# Estimador de la matriz de covarianza muestral (Máxima Verosimilitud de Σ11)
sigma_11_hat <- cov(x1)
print(sigma_11_hat)
```

```
           V1         V2        V3
V1  4.3697911 -0.4407255  2.435705
V2 -0.4407255  9.9162583 -1.465279
V3  2.4357049 -1.4652786  5.628944
```

Solución (c)

Se nos pide calcular los valores de la variable Y, definida por:

$$\mathbf{Y} = \mathbf{a}^T\mathbf{x}^{(2)}, \quad con \quad \mathbf{a} = \begin{bmatrix} 1 & 2 & -1 \end{bmatrix}^T \quad y \quad \underline{\mathbf{x}}^{(2)} = \begin{bmatrix} X_4 \\ X_5 \\ X_6 \end{bmatrix}$$

Haciendo el calculo tenemos lo siguiente

```
x2_data <- as.matrix(datos_01_y_02[, c("V4", "V5", "V6")])

# vector a
a <- c(1, 2, -1)

# Calcular los valores de Y
y_values <- x2_data %*% a
```

```r
y_values[1:10,] # Primeras 10 obervaciones de los valores de Y
```

```
 [1] -4.2159  3.4235 -9.7821  2.9079  1.9770 -1.1429  6.1409 -2.8938  2.6106
[10] -1.8939
```

Ahora verifiquemos la normalidad univariada a los valores asociados a Y.

```r
# Verificación de normalidad univariada con la prueba de Shapiro-Wilk
shapiro.test(y_values)
```

```

    Shapiro-Wilk normality test

data:  y_values
W = 0.99259, p-value = 0.8698
```

Dado que el p-valor es 0.8698 que es significativamente mayor a 0.05, podriamos decir que los datos asociados a Y podrían provenir de una distribución normal, lo cual es consistente con el supuesto de normalidad.

Solución (d)

Se nos pide calcular los valores de la variable Y, definida por:

$$\mathbf{Y} = \mathbf{A}\underline{\mathbf{x}}^{(1)}, \quad con \quad \mathbf{A} = \begin{bmatrix} 0 & 1 & 2 \\ 2 & 0 & -1 \end{bmatrix} \quad y \quad \underline{\mathbf{x}}^{(1)} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

Haciendo el calculo tenemos que

```r
# Seleccionar las columnas correspondientes a x(1)
x1_data <- as.matrix(datos_01_y_02[, c("V1", "V2", "V3")])

# Definir la matriz A
A <- matrix(c(0, 1, 2,2, 0, -1),
            nrow = 2, byrow = TRUE)

# Calcular los valores de Y = A * x(1)
y_values_d <- x1_data %*% t(A)


# Primeras 10 filas de los valores de Y
y_values_d[1:10, ]
```

```
         [,1]    [,2]
 [1,] -4.2824 -0.1716
 [2,] -0.5055  1.0423
 [3,] -4.0639  5.8175
```

```
 [4,]  2.8040  4.4765
 [5,] -5.4089  3.7831
 [6,] -2.9337 -1.8006
 [7,]  3.4953  1.3161
 [8,] -2.9879 -0.8263
 [9,]  2.9011 -0.8607
[10,]  3.7672  1.2241
```

Ahora verifiquemos la normalidad-bivariada a los valores asociados a Y.

```r
# Aplicar prueba de Mardia para normalidad bivariada
mardia_test_d <- mvn(data = as.data.frame(y_values_d),
                   mvnTest = "mardia")

mardia_test_d$multivariateNormality
```

```
             Test         Statistic           p value Result
1 Mardia Skewness   4.0057558363536 0.40522744241763    YES
2 Mardia Kurtosis 0.143080994084797 0.88622621501471    YES
3             MVN              <NA>             <NA>    YES
```

- Para la asimetría, el p-valor alto (0.4052) indica que no hay evidencia suficiente para rechazar la normalidad.
- Para la curtosis, el p-valor alto (0.8862) también sugiere que los datos cumplen con la normalidad.
- La conclusión general del test (MVN) confirma que los datos asociados a Y son consistentes con una distribución normal multivariada.

Punto-02:

A partir de los dos conjuntos de datos asociados a $\mathbf{x}^{(1)}$ y $\mathbf{x}^{(2)}$, realice los siguientes puntos:

(a) Hallar $\mu_{1|2} = \mathbb{E}[\mathbf{x}^{(1)} \mid \mathbf{x}^{(2)}]$.

(b) A partir de (a), ¿cuál es la matriz de coeficientes que resulta del ajuste de un Modelo de Regresión Lineal Multivariado (MRL-Multivariado) de $\mathbf{x}^{(1)}$ versus $\mathbf{x}^{(2)}$?

(c) Utilizando teoría de modelos lineales, ajuste el MRL-Multivariado de $\mathbf{x}^{(1)}$ versus $\mathbf{x}^{(2)}$. Compare los coeficientes de dicho modelo ajustado con los obtenidos en (b).

Solución Punto-02:

Solución (a)

Se nos pide hallar $\mu_{1|2} = E[\underline{\mathbf{x}}^{(1)} \mid \underline{\mathbf{x}}^{(2)}]$

Sabemos que

$$\hat{\mu}_{1|2} = \underline{\hat{\mu}}^{(1)} + \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}\left(\underline{\mathbf{x}}^{(2)} - \underline{\hat{\mu}}^{(2)}\right)$$

Haciendo el calculo tenemos que

```
# Separar los conjuntos de datos
x1 <- as.matrix(datos_01_y_02[, 1:3])  # Variables V1, V2, V3
x2 <- as.matrix(datos_01_y_02[, 4:6])  # Variables V4, V5, V6

# Calcular las medias muestrales
mu1 <- colMeans(x1)
mu2 <- colMeans(x2)

# Calcular matrices de covarianza
S11 <- cov(x1)
S22 <- cov(x2)
S12 <- cov(x1, x2)
S21 <- t(S12)

# Cálculo de la media condicional:  _{1|2} = E[x(1) | x(2)]
mu_1_given_2 <- function(x2_val) {
  mu1 + S12 %*% solve(S22) %*% (x2_val - mu2)
}

# Aplicar la función a los valores observados de x(2)
est_mu_1_given_2 <- t(apply(x2, 1, mu_1_given_2))

# Primeras 10 observaciones
est_mu_1_given_2 %>% head(10)
```

```
            [,1]        [,2]        [,3]
 [1,] -1.3553765  4.37016752 -2.03552612
 [2,]  0.4128138 -0.84645970 -0.11007908
 [3,] -0.8155162  0.03930673 -2.16221587
 [4,]  0.9067611 -0.86654084  2.20719574
 [5,] -1.3863161  5.14597979 -2.78992467
 [6,]  0.3771059 -0.94286047  0.58990915
 [7,]  0.5327162  0.47273639  0.99986015
 [8,] -0.4323892  0.56521761 -1.40687991
 [9,]  0.8358329 -1.09426221  1.75353399
[10,]  0.1331770 -0.63346006 -0.08993782
```

Solucion (b)

```
# Matriz de coeficientes del modelo de regresión
B_hat <- S12 %*% solve(S22)
print(B_hat)
```

```
              V4          V5          V6
V1 -0.058743490  0.25345173 -0.07353095
V2  0.553898748 -0.07750982  0.19987111
V3  0.003132685  0.66344081  0.04630720
```

- Las filas de la matriz (B) representan las variables dependientes ((V1, V2, V3)).
- Las columnas de la matriz (B) representan las variables independientes ((V4, V5, V6)).

Los valores en la matriz (B) indican cómo cada variable independiente ((V4, V5, V6)) afecta, en promedio, a cada variable dependiente ((V1, V2, V3)).

Por ejemplo:

- El coeficiente $B_{(1,4)} = -0.0587$ indica que un aumento de una unidad en (V4) disminuye (V1) en (0.0587) unidades, manteniendo constantes (V5) y (V6).

El coeficiente $B_{(2,5)} = -0.0775$ indica que un aumento de una unidad en (V5) disminuye (V2) en (0.0775) unidades, manteniendo constantes (V4) y (V6).

- El coeficiente $B_{(3,5)} = 0.6634$ muestra que (V3) aumenta considerablemente cuando (V5) incrementa en una unidad.

Solución (c)

```r
# Ajustar el modelo de regresión lineal multivariado
modelo <- lm(cbind(V1, V2, V3) ~ V4 + V5 + V6, data = datos_01_y_02)

# Coeficientes obtenidos del modelo ajustado
coef_modelo <- coef(modelo)

print(coef_modelo)
```

```
                     V1          V2           V3
(Intercept)  0.16445841 -0.50094574 -0.226650667
V4          -0.05874349  0.55389875  0.003132685
V5           0.25345173 -0.07750982  0.663440812
V6          -0.07353095  0.19987111  0.046307196
```

- La matriz (B) calculada coincide exactamente con los coeficientes del modelo ajustado.

- Esto valida que los cálculos teóricos y el modelo ajustado son consistentes.

Conclusión general

La matriz (B) y los coeficientes del modelo ajustado confirman que las relaciones entre las variables independientes ((V4, V5, V6)) y las dependientes ((V1, V2, V3)) son adecuadamente modeladas por un modelo de regresión lineal multivariado.

Punto-03:

Para este punto, considere los dos conjuntos de datos asignados, los cuales corresponden a datos simulados de los vectores normales 3-variados independientes $\mathbf{x}_1$ y $\mathbf{x}_2$, con vector de medias y matriz de varianza-covarianza dados por:

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 4 & 2 & 0 \\ 2 & 3 & 1 \\ 0 & 1 & 5 \end{bmatrix}$$

Es decir, los dos conjuntos de datos son simulaciones de los vectores:

$$\mathbf{x}_1 \sim N_3(\mu, \Sigma), \quad \mathbf{x}_2 \sim N_3(\mu, \Sigma), \quad \mathrm{Cov}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{O}_{3\times 3}$$

Considere las siguientes combinaciones lineales de $\mathbf{x}_1$ y $\mathbf{x}_2$:

$$\mathbf{v}_1 = \mathbf{x}_1 + 2\mathbf{x}_2, \quad \mathbf{v}_2 = 2\mathbf{x}_1 - \mathbf{x}_2$$

(a) Obtenga los datos muéstrales (o puntuaciones) asociados a los vectores $\mathbf{v}_1$ y $\mathbf{v}_2$.

(b) Realice la verificación de normalidad 3-variada de los datos asociados a $\mathbf{v}_1$.
Nota: Utilizar algunas de las herramientas vistas en clase sobre procesos de evaluación de la normalidad multivariada y/o herramientas que usted conozca para dichos procesos.

(c) Realice la verificación de normalidad 6-variada de los datos asociados al vector:

$$\mathbf{v} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix}$$

Nota: Utilizar algunas de las herramientas vistas en clase sobre procesos de evaluación de la normalidad multivariada y/o herramientas que usted conozca para dichos procesos.

Solución: Punto-03

Solución (a)

Tenemos que

$$\mathbf{v}_1 = \mathbf{x}_1 + 2\mathbf{x}_2, \quad y \quad \mathbf{v}_2 = 2\mathbf{x}_1 - \mathbf{x}_2$$

Haciendo los calculos tenemos que

```
# cargando la muestras de datos de X(1) y X(2)
x1_data_03 <- read.table("equipo_07_muestra1_datos_03.txt")
x2_data_03 <- read.table("equipo_07_muestra2_datos_03.txt")
```

```r
# Calculando a V1 y V2
v1 <- x1_data_03 + 2*x2_data_03
v2 <- 2*x1_data_03 - x2_data_03
```

Ahora vizualicemos las primeras 10 observaciones asociadas a $\mathbf{v}_1$ y $\mathbf{v}_2$.

```r
v1 %>% head(10)
```

```
        V1      V2       V3
1  -1.8490 -2.5414 -13.9794
2  -2.6448 -1.0773   1.9142
3   0.4960 -2.9083  -6.0066
4   1.5260  1.3293   0.3419
5  -3.0829  0.9189  -0.3715
6   1.8622  0.3212  -8.8208
7   7.2936  7.2650  -3.2885
8   3.2889  9.1205  -2.4975
9   1.4521 -1.5028   5.1043
10  1.5850  0.1867   3.9368
```

```r
v2 %>% head(10)
```

```
        V1      V2       V3
1  -2.7475 -3.5153   0.2252
2  14.1824  9.5144  -3.4876
3   4.6625 -2.0701   8.4588
4   7.1300 -0.5934   2.2308
5  -2.0778 -3.4112  -3.0700
6   0.6509  0.0249  -9.7021
7  -1.9483 -5.7180   4.2535
8   1.5703 -2.2480 -14.4340
9   4.3702  2.0014  -1.5494
10  4.0490 -1.3071   0.9296
```

Solución (b)

Verifiquemos la normalidad 3-variada de los datos asociados $\mathbf{v}_1$.

```r
# Aplicar prueba de Mardia para normalidad 3-variada
mardia_test_v1 <- mvn(data = as.data.frame(v1),
                   mvnTest = "mardia")

mardia_test_v1$multivariateNormality
```

```
             Test        Statistic           p value Result
1 Mardia Skewness 10.2906401172815 0.415375186124852    YES
2 Mardia Kurtosis 1.09536098434238  0.27335851982769    YES
3             MVN             <NA>              <NA>    YES
```

- Para la asimetría, el p-valor alto (0.4153) indica que no hay evidencia suficiente para rechazar la normalidad.
- Para la curtosis, el p-valor alto (0.2733) también sugiere que los datos cumplen con la normalidad.
- La conclusión general del test (MVN) confirma que los datos asociados a $\mathbf{v}_1$ son consistentes con una distribución normal multivariada.

Solución (c)

Veamos primeramente si $\underline{\mathbf{x}}_1$ y $\underline{\mathbf{x}}_2$ cumplen el supuesto de nomralidad 3-variada.

```r
mardia_test_x1 <- mvn(data = as.data.frame(x1_data_03), mvnTest = "mardia")
mardia_test_x1$multivariateNormality
```

```
             Test         Statistic            p value Result
1 Mardia Skewness   20.595282974089 0.0240992905964353     NO
2 Mardia Kurtosis 0.992560426477898  0.320924218144901    YES
3             MVN              <NA>               <NA>     NO
```

- Para la asimetría, el p-valor bajo (0.0.0240) indica que si hay evidencia suficiente para rechazar la normalidad.
- Para la curtosis, el p-valor alto (0.3209) también sugiere que los datos cumplen con la normalidad.
- La conclusión general del test (MVN) confirma que los datos asociados a $\underline{\mathbf{x}}_1$ no son consistentes con una distribución normal multivariada.

```r
mardia_test_x2 <- mvn(data = as.data.frame(x2_data_03), mvnTest = "mardia")
mardia_test_x2$multivariateNormality
```

```
             Test        Statistic           p value Result
1 Mardia Skewness 15.1833963323184  0.12551724045642    YES
2 Mardia Kurtosis 1.18675599467803 0.235323881097516    YES
3             MVN             <NA>              <NA>    YES
```

Con base en los resultados de la prueba de Mardia,no hay evidencia suficiente para rechazar la hipótesis de normalidad 3-variada. Por lo tanto, los datos asociados a $\mathbf{x}_2$ se ajustan a una distribución normal 3-variada.

Dado que los datos asociados a $\mathbf{x}_1$ no se distribuyen normal 3-variando, por tanto los datos asociados al vector $\mathbf{v}$ que se compone de combinaciones lineales de $\mathbf{x}_1$ y $\mathbf{x}_2$ no cumpliran el supuesto de normalidad 6-variada.