Análisis de Normalidad y Homocedasticidad Multivariable

Nazaret Basaldella

2024-11-29

##Introducción

En este informe se analiza la normalidad y homocedasticidad multivariable utilizando el conjunto de datos `iris` . El análisis incluye:

1. Evaluación de la normalidad univariante:
    - Histogramas con diagramas de densidad.
    - Pruebas de Shapiro-Wilk.
    - Q-Q Plots.
2. Evaluación de la normalidad multivariante:
    - Pruebas de Mardia, Henze-Zirkler y Royston.
3. Evaluación de la homocedasticidad:
    - Prueba de Levene.
    - Prueba de Box’s M.
4. Visualizaciones:
    - Heatmap de correlaciones.
    - Boxplots por especie.

```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

```
#Preparacion de los datos
# Instalar librerías necesarias si no están instaladas
if (!requireNamespace("dplyr", quietly = TRUE)) install.packages("dplyr")
if (!requireNamespace("ggplot2", quietly = TRUE)) install.packages("ggplot2")
if (!requireNamespace("tidyr", quietly = TRUE)) install.packages("tidyr")
if (!requireNamespace("MVN", quietly = TRUE)) install.packages("MVN")
if (!requireNamespace("biotools", quietly = TRUE)) install.packages("biotools")
if (!requireNamespace("heplots", quietly = TRUE)) install.packages("heplots")

# Cargar librerías
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyr)
library(MVN)
library(biotools)
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## ---
## biotools version 4.2
```

```
library(heplots)
```

```
## Loading required package: broom
```

```
##
## Attaching package: 'heplots'
```

```
## The following object is masked from 'package:biotools':
##
##     boxM
```

```
# Cargar datos
data(iris)

# Selección de variables cuantitativas y categórica
variables_cuantitativas <- iris[, c("Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Wid
th")]
variable_categorica <- iris$Species
```

1. Normalidad Multivariante

1.a. Prueba de Normalidad Univariante (Shapiro-Wilk)

```
# Prueba de Shapiro-Wilk para cada variable cuantitativa
shapiro_results <- lapply(variables_cuantitativas, shapiro.test)

# Extraer los p-valores
shapiro_p_values <- sapply(shapiro_results, function(x) x$p.value)

# Crear una tabla con los resultados
shapiro_results_df <- data.frame(
  Variable = names(shapiro_p_values),
  P_Value = shapiro_p_values,
  Normalidad = ifelse(shapiro_p_values > 0.05, "Sí", "No")
)

print(shapiro_results_df)
```

```
##                  Variable      P_Value Normalidad
## Sepal.Length Sepal.Length 1.018116e-02         No
## Sepal.Width   Sepal.Width 1.011543e-01         Sí
## Petal.Length Petal.Length 7.412263e-10         No
## Petal.Width   Petal.Width 1.680465e-08         No
```

1.b. Pruebas de Normalidad Multivariante (Mardia, Henze-Zirkler y Royston)

```
# Pruebas de normalidad multivariante usando MVN
mardia_result <- mvn(variables_cuantitativas, mvnTest = "mardia")
henze_result <- mvn(variables_cuantitativas, mvnTest = "hz")
royston_result <- mvn(variables_cuantitativas, mvnTest = "royston")

# Mostrar resultados
cat("\nResultados del Test de Mardia:\n")
```

```
##
## Resultados del Test de Mardia:
```

```
print(mardia_result$multivariateNormality)
```

```
##              Test          Statistic              p value Result
## 1 Mardia Skewness    67.430508778062 4.75799820400869e-07     NO
## 2 Mardia Kurtosis -0.230112114481001    0.818004651478012    YES
## 3             MVN               <NA>                 <NA>     NO
```

```
cat("\nResultados del Test de Henze-Zirkler:\n")
```

```
##
## Resultados del Test de Henze-Zirkler:
```

```
print(henze_result$multivariateNormality)
```

```
##            Test       HZ p value MVN
## 1 Henze-Zirkler 2.336394       0  NO
```

```
cat("\nResultados del Test de Royston:\n")
```

```
##
## Resultados del Test de Royston:
```

```
print(royston_result$multivariateNormality)
```

```
##      Test        H      p value MVN
## 1 Royston 50.39667 3.098229e-11  NO
```

2. Homocedasticidad

2.a. Prueba de Levene

```
# Prueba de Levene para cada variable
levene_results <- lapply(variables_cuantitativas, function(x) {
  car::leveneTest(x, variable_categorica)
})

# Mostrar resultados
cat("\nResultados de la prueba de Levene:\n")
```

```
##
## Resultados de la prueba de Levene:
```

```
levene_results
```

```
## $Sepal.Length
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group   2  6.3527 0.002259 **
##       147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## $Sepal.Width
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.5902 0.5555
##       147
##
## $Petal.Length
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   2   19.48 3.129e-08 ***
##       147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## $Petal.Width
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   2  19.892 2.261e-08 ***
##       147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.b. Prueba de Box’s M

```
# Prueba de Box's M para homogeneidad de matrices de covarianza
boxm_result <- boxM(as.matrix(variables_cuantitativas), variable_categorica)

cat("\nResultado de la prueba de Box's M:\n")
```

```
##
## Resultado de la prueba de Box's M:
```

```
print(boxm_result)
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  as.matrix(variables_cuantitativas)
## Chi-Sq (approx.) = 140.94, df = 20, p-value < 2.2e-16
```

3. Visualizaciones

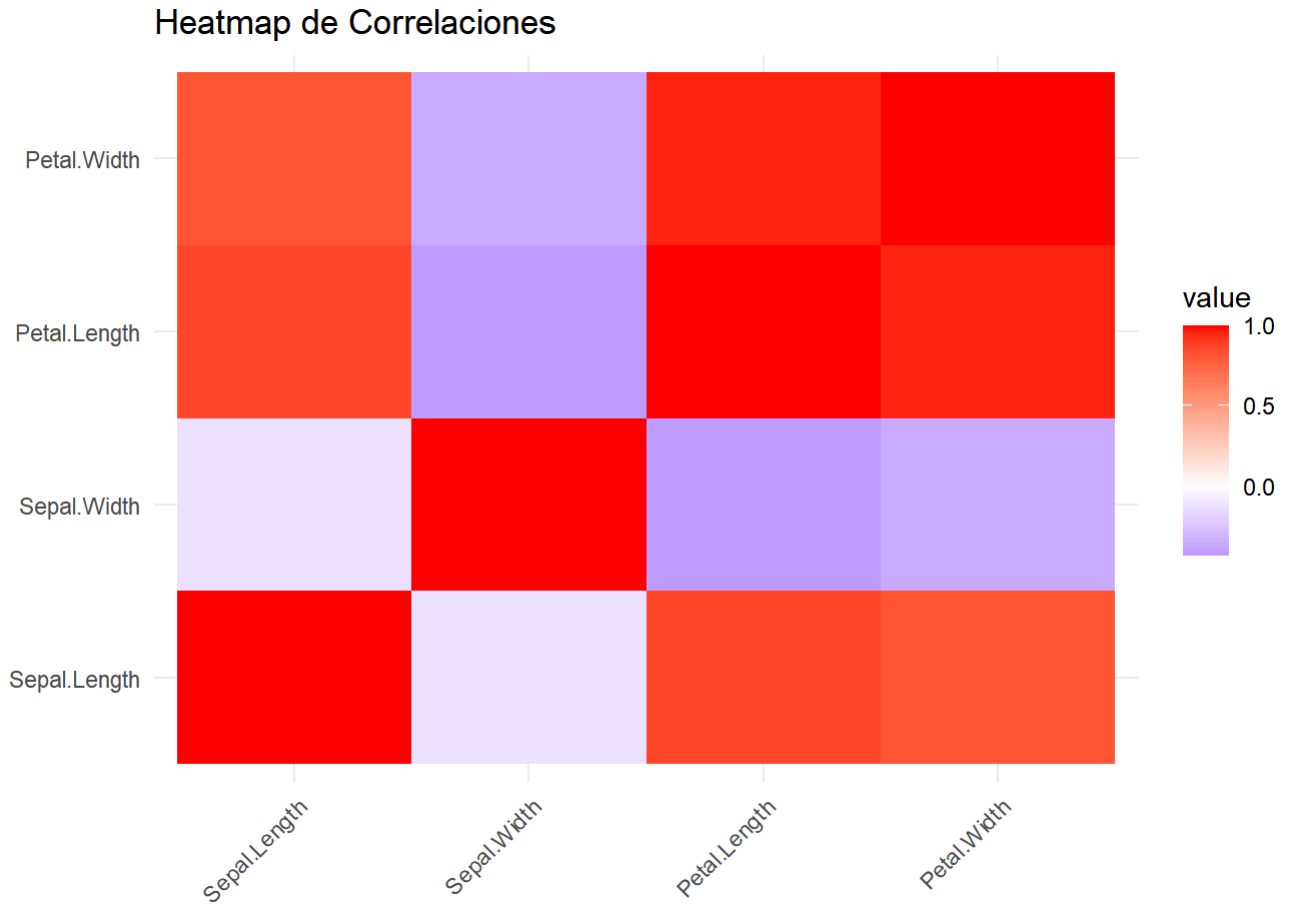
3.a. Heatmap de Correlaciones

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

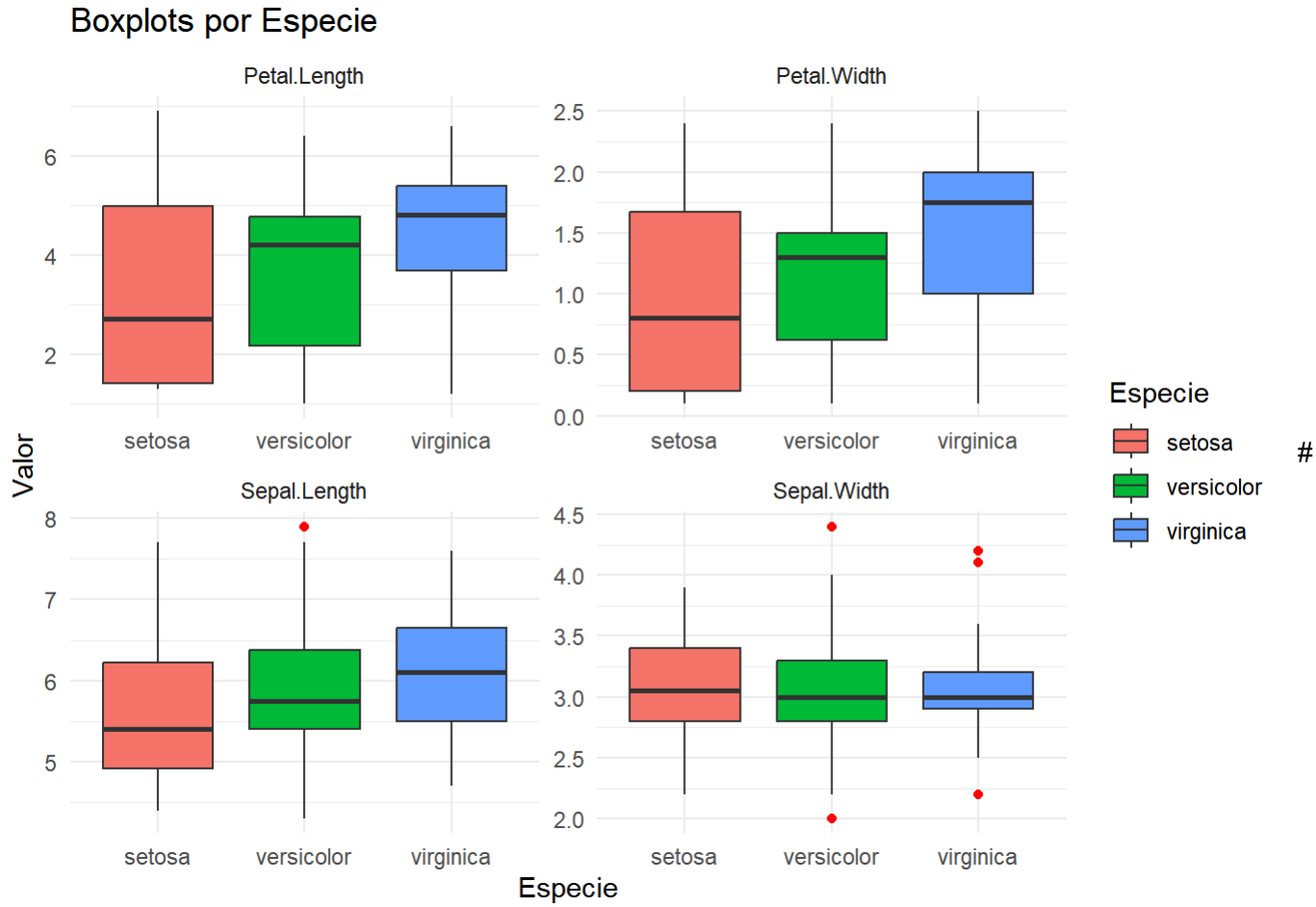
```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
# Calcular la matriz de correlación
cor_matrix <- cor(variables_cuantitativas)

# Transformar en formato largo
cor_data <- melt(cor_matrix)

# Crear el heatmap
ggplot(cor_data, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  labs(title = "Heatmap de Correlaciones", x = "", y = "") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



3.b. Boxplots por Especie

```
# Crear boxplots para cada variable cuantitativa por especie
variables_cuantitativas %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Valor") %>%
  mutate(Especie = rep(variable_categorica, times = ncol(variables_cuantitativas))) %>%
  ggplot(aes(x = Especie, y = Valor, fill = Especie)) +
  geom_boxplot(outlier.color = "red", outlier.shape = 16) +
  facet_wrap(~Variable, scales = "free") +
  labs(title = "Boxplots por Especie", x = "Especie", y = "Valor") +
  theme_minimal()
```



Conclusiones

Normalidad Univariante

1. Según los resultados de la prueba de Shapiro-Wilk:
   - **Sepal.Length** ($p = 0.01$): No cumple con la normalidad.
   - **Sepal.Width** ($p = 0.10$): Cumple con la normalidad.
   - **Petal.Length** ($p < 0.001$): No cumple con la normalidad.
   - **Petal.Width** ($p < 0.001$): No cumple con la normalidad.
2. Los **Q-Q Plots** confirman estas observaciones, mostrando que **Sepal.Width** tiene la distribución más cercana a la normalidad, mientras que las variables relacionadas con los pétalos presentan fuertes desviaciones.

Normalidad Multivariante

1. Resultados de las pruebas multivariantes:

- **Mardia Skewness** ($p < 0.001$): Los datos presentan asimetría significativa y no cumplen con la normalidad.
- **Mardia Kurtosis** ($p = 0.818$): No se detecta curtosis significativa.
- **Henze-Zirkler** ($p = 0$): Los datos no cumplen con la normalidad multivariante.
- **Royston** ($p < 0.001$): Tampoco se cumple la normalidad multivariante.

2. En general, los datos **no cumplen con la normalidad multivariante**, principalmente debido a la asimetría detectada.

Homocedasticidad

1. **Prueba de Levene**:
   - **Sepal.Width** ($p = 0.5555$): Presenta homogeneidad de varianzas.
   - Las demás variables (**Sepal.Length**, **Petal.Length**, **Petal.Width**) tienen $p < 0.05$, indicando heterogeneidad de varianzas.
2. **Box's M Test**:
   - $p < 0.001$: Las matrices de covarianza no son homogéneas entre grupos.

Visualizaciones

1. **Boxplots por Especie**:
   - Las variables relacionadas con los pétalos (**Petal.Length** y **Petal.Width**) muestran las mayores diferencias entre especies, lo que sugiere que son útiles para clasificar las especies.
   - **Sepal.Length** también discrimina entre especies, aunque menos pronunciadamente.
   - **Sepal.Width** tiene el mayor solapamiento entre grupos y es menos discriminante.
2. **Heatmap de Correlaciones**:
   - Las variables de pétalos están altamente correlacionadas entre sí, con correlaciones cercanas a 1.
   - **Sepal.Width** muestra baja correlación con las demás variables, indicando que aporta información diferente.

Resumen General

- Ninguna de las variables cumple con la normalidad multivariante, y solo **Sepal.Width** cumple con la normalidad univariante.
- Las pruebas de homocedasticidad indican que **Sepal.Width** presenta homogeneidad de varianzas, mientras que las demás no.
- Las variables relacionadas con los pétalos son las más útiles para discriminar entre especies debido a sus claras diferencias entre grupos.