Transformaciones

Diego Alberto Baños Lopez
A01275100

21-08-2023

Leer el archivo de trabajo: datos de McDonald

```
# Cargamos las librerias que se usaran para esta actividad
library(MASS)
library(e1071)
library(nortest)
library(VGAM)
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```
# Cargaremos los datos del CSV
M <- read.csv("./mc-donalds-menu-1.csv")
```

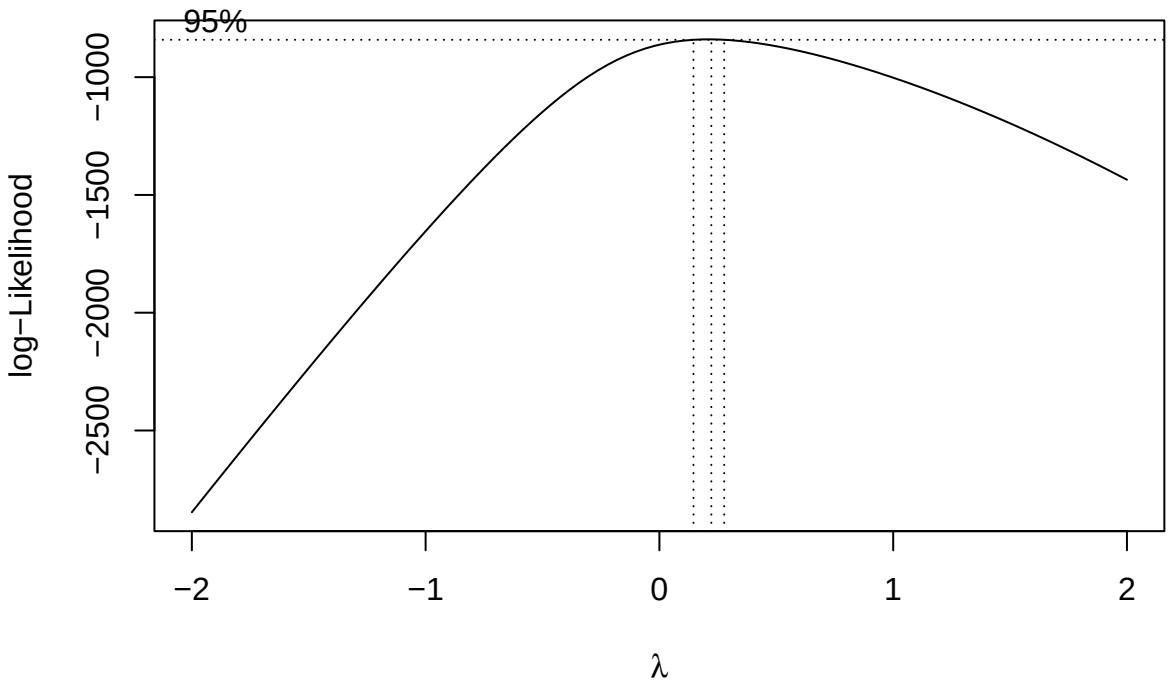
Transformaciones Box-Cox

Usando modelo exacto y aproximado

```
# Cargamos la columna Sodium
sodium_data <- M$Sodium

# Ajustar los valores no positivos
min_value <- min(sodium_data)
if (min_value <= 0) {
  sodium_data <- sodium_data + abs(min_value) + 1
}

# Transformación Box-Cox
boxcox_result <- boxcox(sodium_data ~ 1, lambda = seq(-2, 2, 0.1))
```




```
lambda_optimal <- boxcox_result$x[which.max(boxcox_result$y)]

# Aplicamos las transformaciones
sodium_transformed_exact <- ((sodium_data^lambda_optimal) - 1) / lambda_optimal
sodium_transformed_approx <- sqrt(sodium_data + 1)
```

Analisis de Normalidad

Aqui se analizara la normalidad para sodio

```
# Comparación de medidas estadísticas
print("/*------------Medidas------------*/")
```

```
## [1] "/*------------Medidas------------*/"
```

```
summary(sodium_data)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   108.5   191.0   496.8   866.0  3601.0
```

```r
print("*----------------------------------*")
```

```
## [1] "*----------------------------------*"
```

```r
summary(sodium_transformed_exact)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.249   9.958  11.008  15.730  23.267
```

```r
print("*----------------------------------*")
```

```
## [1] "*----------------------------------*"
```

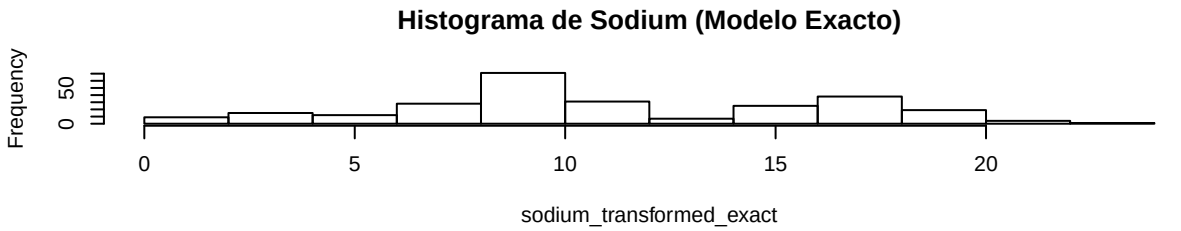
```r
summary(sodium_transformed_approx)
```

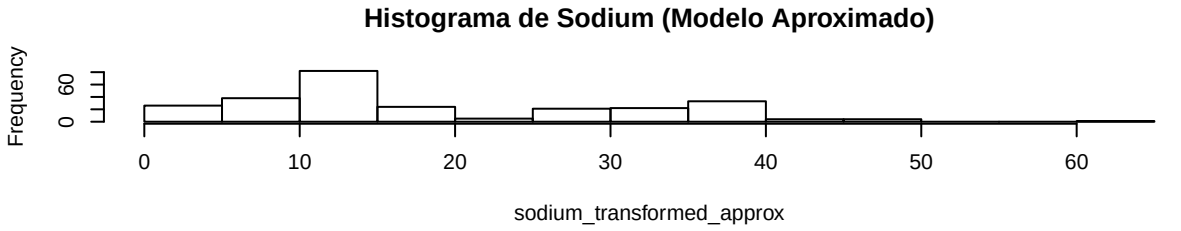
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.414  10.462  13.856  18.651  29.445  60.017
```

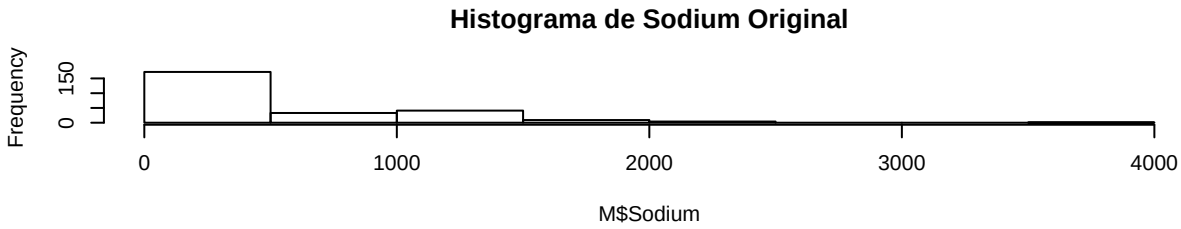
```r
print("/*--------------------------------*/")
```

```
## [1] "/*--------------------------------*/"
```

```r
# Histogramas
par(mfrow = c(3, 1))
hist(sodium_transformed_exact, col = 0, main = "Histograma de Sodium (Modelo Exacto)")
hist(sodium_transformed_approx, col = 0, main = "Histograma de Sodium (Modelo Aproximado)")
hist(M$Sodium, col = 0, main = "Histograma de Sodium Original")
```








```r
# Pruebas de Anderson-Darling
print("/*---Pruebas Anderson-Darling---*/")
```

```
## [1] "/*---Pruebas Anderson-Darling---*/"
```

```r
ad.test(sodium_data)
```

```
##
##  Anderson-Darling normality test
##
## data:  sodium_data
## A = 21.406, p-value < 2.2e-16
```

```r
print("*--------------------------------*")
```

```
## [1] "*--------------------------------*"
```

```r
ad.test(sodium_transformed_exact)
```

```
##
##  Anderson-Darling normality test
##
## data:  sodium_transformed_exact
## A = 4.0199, p-value = 4.98e-10
```

```r
print("*--------------------------------*")
```

```
## [1] "*--------------------------------*"
```

```r
ad.test(sodium_transformed_approx)
```

```
##
##  Anderson-Darling normality test
##
## data:  sodium_transformed_approx
## A = 9.791, p-value < 2.2e-16
```

```r
print("/*--------------------------------*/")
```

```
## [1] "/*--------------------------------*/"
```

- La prueba de Anderson-Darling en los sodio original indican que el valor A es muy grande y el valor P es mucho mas chico que 0.5, haciendo que esta no siga una distribución normalidad
- La transformacion de Box-Cox en ambas versiones mejoran los resultados de Anderson-Darling, haciendo que tengan mas normalidad, no obstante aun con esto, no se puede decir que sigan una distribución normal ## Detección y corrección de anomalías

```r
# Escalamiento robusto
sodium_robust <- (sodium_data - median(sodium_data)) / IQR(sodium_data)

# Eliminar valores atípicos usando el método de Tukey
outliers <- boxplot.stats(sodium_robust)$out
sodium_clean <- sodium_robust[!sodium_robust %in% outliers]
```

Transformación de Yeo-Johnson

```r
# Función objetivo para optimizar
objective_function <- function(lambda) {
  transformed_data <- VGAM::yeo.johnson(sodium_clean, lambda = lambda)
  -shapiro.test(transformed_data)$statistic
}

# Encontrar el valor óptimo de lambda para Yeo-Johnson
optimal_lambda_result <- optimize(objective_function, interval = c(-2, 2))
lambda_optimal_yj <- optimal_lambda_result$minimum

# Transformación de Yeo-Johnson con el lambda óptimo
sodium_yj_transformed <- VGAM::yeo.johnson(sodium_clean, lambda = lambda_optimal_yj)

# Análisis de normalidad
print("/*------------Medidas Yeo-Johnson------------*/")
```

```
## [1] "/*------------Medidas Yeo-Johnson------------*/"
```

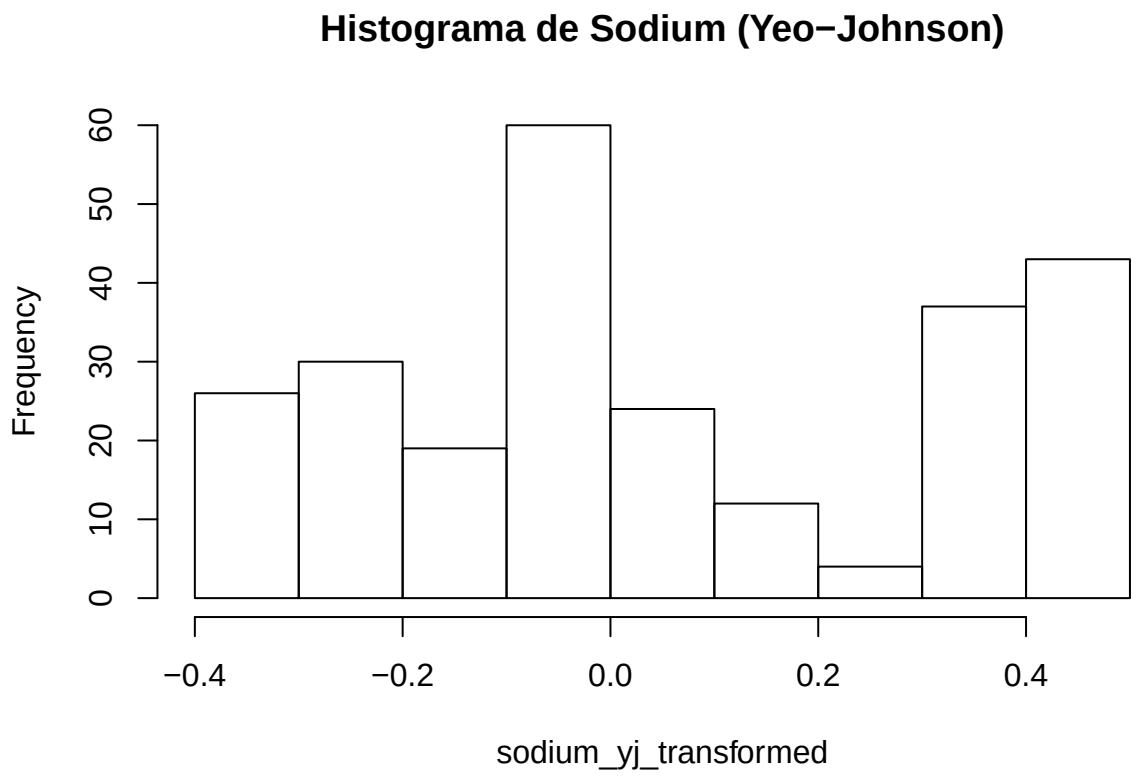
```r
summary(sodium_yj_transformed)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.36196 -0.15112  0.00000  0.05367  0.35308  0.44882
```

```r
print("/*-----------------------------------------*/")
```

```
## [1] "/*-----------------------------------------*/"
```

```r
# Histograma
hist(sodium_yj_transformed, col = 0, main = "Histograma de Sodium (Yeo-Johnson)")
```




```r
# Prueba de Anderson-Darling
print("/*---Prueba Anderson-Darling Yeo-Johnson---*/")
```

```
## [1] "/*---Prueba Anderson-Darling Yeo-Johnson---*/"
```

```r
ad.test(sodium_yj_transformed)
```

```
##
##  Anderson-Darling normality test
##
## data:  sodium_yj_transformed
## A = 7.587, p-value < 2.2e-16
```

```r
print("/*----------------------------------------*/")
```

```
## [1] "/*----------------------------------------*/"
```

- Los valores arrojados apuntan a que pese a que en la prueba de Anderson-Darling la Transformación de Yeo-Johnson tenga mejores resultados que el set original, aun no podemos decir que los resultados representen una distribucion normal

Observaciones

- La transformaciones, tanto las de Box-Cox, como la de Yeo-Johnson ayudan a mejorar la normalidad de los datos, esto puede ser util para tratar los datos, dependiendo de lo que uno requiera utilizar

Diferencias entre tratamiento y Escalamiento de los datos

- La transformacion modifica la distribución de los datos, mientras que el escalamiento modifica la escala de los datosn sin modificar la forma de estos ultimos
- La meta de la transformación es poder lograr una distribución especifica como la normal mientras que el escalamiento se utiliza para hacer que los datos se encuentren en un rango en especifico
- Mientras que el trataimiento involucra correcciones de consistencias, modificaciones o darles formato para posteriormente poder hacer un analisis estadistico adecuado, el escalamiento se refiere esencialmente a clasificar los datos en distintas escalas para aquellos analisis que requieran un rango especifico en los datos.