Actividad 5. Transformaciones

Daniela Jiménez Téllez

2024-08-14

Trabaja con el set de datos Mc Donalds menu Download Mc Donalds menu, que contiene diversas características del menú de alimentos de Mc Donalds.

Importación de librerías

```
library(MASS)
library(nortest)
library(moments)
```

```
library(car)
```

```
## Cargando paquete requerido: carData
```

Importación de datos

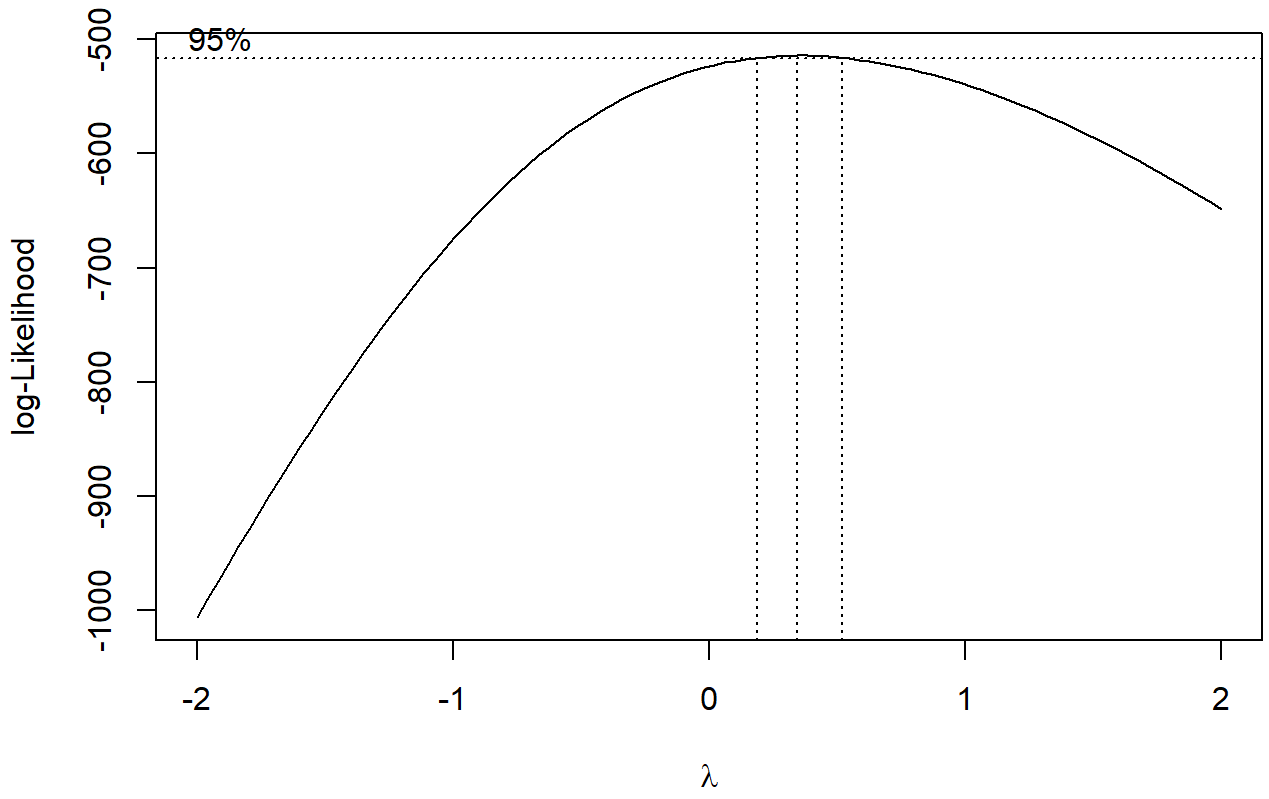
```
data <- read.csv("mc-donalds-menu.csv")

carbohydrates <- data$Carbohydrates
carbohydrates <- carbohydrates[carbohydrates != 0]
```

Preguntas

1. Utiliza la transformación Box-Cox. Utiliza el modelo exacto y el aproximado de acuerdo con las sugerencias de Box y Cox para la transformación

```
# BoxCox
boxcox_model <- boxcox(lm((carbohydrates + 1) ~ 1))
```




```
# Exacta
lambda_exact <- boxcox_model$x[which.max(boxcox_model$y)]
print(lambda_exact)
```

```
## [1] 0.3434343
```

```
# Resultados
data_boxcox_exact <- (carbohydrates^lambda_exact - 1) / lambda_exact
```

2. Escribe las ecuaciones de los modelos encontrados.

Para el modelo de BoxCox exacto se utilizó la ecuación

$$Y(\lambda) = \frac{X^{\lambda} - 1}{\lambda}, \quad \text{donde } \lambda = 0.3434$$

3. Analiza la normalidad de las transformaciones obtenidas con los datos originales. Utiliza como argumento de normalidad:

> 3.1 Compara las medidas: Mínimo, máximo, media, mediana, cuartil 1 y cuartil 3, sesgo y curtosis.

> 3.2 Obten el histograma de los 2 modelos obtenidos (exacto y aproximado) y los datos originales.

> 3.3 Realiza la prueba de normalidad de Anderson-Darling o de Jarque Bera para los datos transformados y los originales

```
# 3.1

measures <- function(x){
  list(
    Min = min(x),
    Max = max(x),
    Mean = mean(x),
    Median = median(x),
    Q1 = quantile(x, 0.25),
    Q3 = quantile(x, 0.75),
    Skewness = skewness(x),
    Kurtosis = kurtosis(x)
  )}

original_measures <- measures(carbohydrates)
exact_measures <- measures(data_boxcox_exact)

cat("Medidas para los datos originales: \n")
```

```
## Medidas para los datos originales:
```

```
print(original_measures)
```

```
## $Min
## [1] 4
##
## $Max
## [1] 141
##
## $Mean
## [1] 50.45082
##
## $Median
## [1] 46
##
## $Q1
## 25%
##  34
##
## $Q3
## 75%
##  61
##
## $Skewness
## [1] 1.221876
##
## $Kurtosis
## [1] 4.802472
```

```
cat("Medidas para los datos exactos: \n")
```

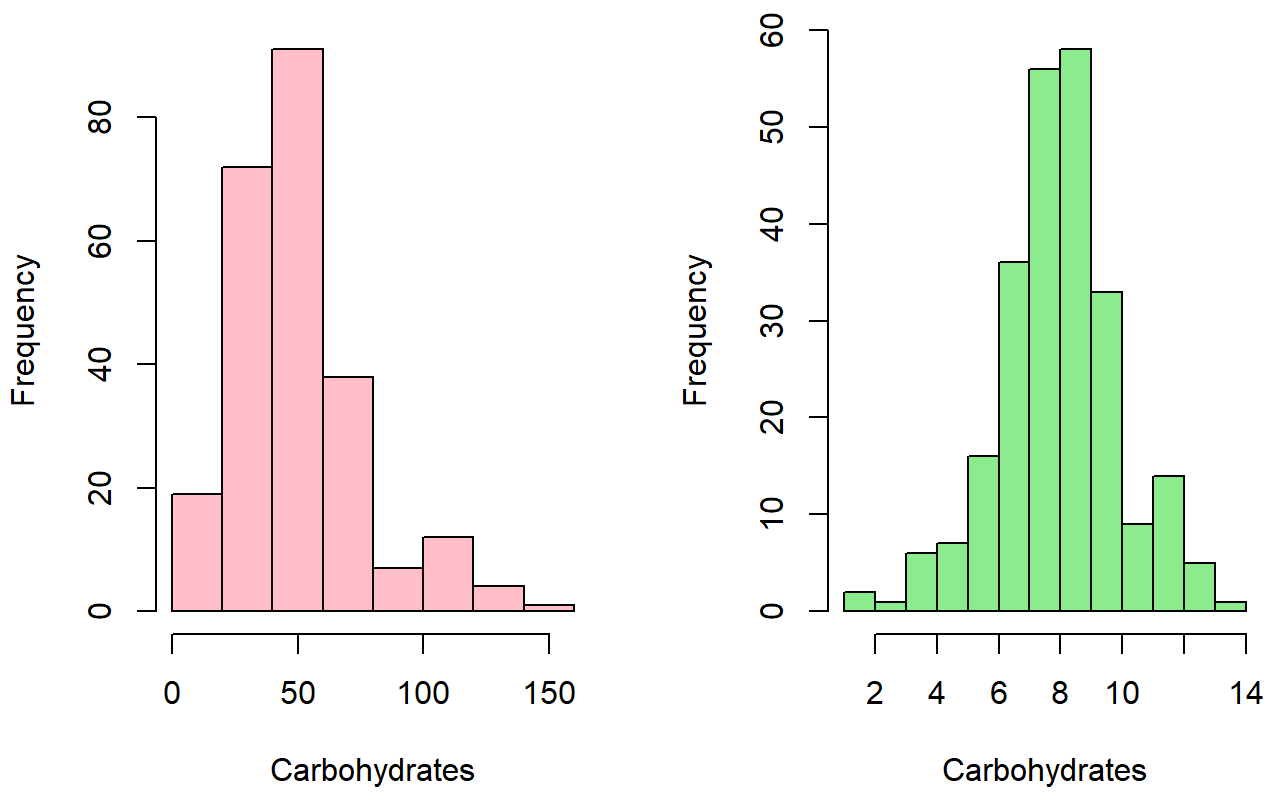
```
## Medidas para los datos exactos:
```

```
print(exact_measures)
```

```
## $Min
## [1] 1.775553
##
## $Max
## [1] 13.0203
##
## $Mean
## [1] 7.94662
##
## $Median
## [1] 7.932608
##
## $Q1
##     25%
## 6.86328
##
## $Q3
##      75%
## 9.036363
##
## $Skewness
## [1] -0.02247246
##
## $Kurtosis
## [1] 3.722577
```

```
# Histogramas

par(mfrow = c(1, 2))
hist(carbohydrates, main = "Histograma de Datos Originales", col = "pink", xlab = "Carbohydrate
s")
hist(data_boxcox_exact, main = "Histograma de Transformación Box-Cox Exacta", col = "lightgree
n", xlab="Carbohydrates")
```




```
# Anderson-Darling

ad_original <- ad.test(carbohydrates)
ad_exact <- ad.test(data_boxcox_exact)

cat("Prueba de normalidad de Anderson-Darling para los datos originales: \n")
```

```
## Prueba de normalidad de Anderson-Darling para los datos originales:
```

```
print(ad_original)
```

```
##
##  Anderson-Darling normality test
##
## data:  carbohydrates
## A = 5.9462, p-value = 1.149e-14
```

```
cat("Prueba de normalidad de Anderson-Darling para los datos exactos: \n")
```

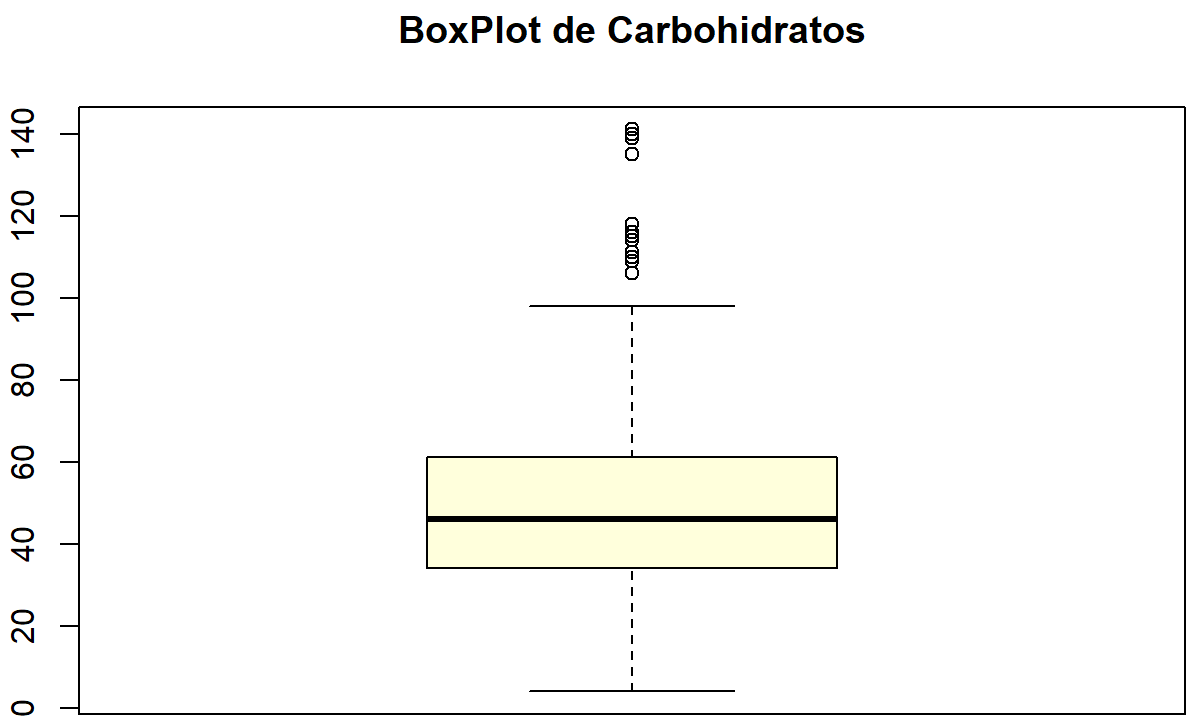
```
## Prueba de normalidad de Anderson-Darling para los datos exactos:
```

```
print(ad_exact)
```

```
##
##  Anderson-Darling normality test
##
## data:  data_boxcox_exact
## A = 1.4393, p-value = 0.0009978
```

4. Detecta anomalías y corrige tu base de datos (datos atípicos, ceros anámalos, etc).

```
boxplot(carbohydrates, col = "lightyellow", main = "BoxPlot de Carbohidratos")
```




```
Q1 <- quantile(carbohydrates, 0.25)
Q3 <- quantile(carbohydrates, 0.75)
IQR <- Q3 - Q1

outliers <- carbohydrates[carbohydrates < (Q1 - 1.5 * IQR) | carbohydrates > (Q3 + 1.5 * IQR)]
carbo_clean <- carbohydrates[!(carbohydrates %in% outliers)]
```

5. Utiliza la transformación de Yeo Johnson y encuentra el valor de lambda que maximiza el valor p de la prueba de normalidad que hayas utilizado (Anderson-Darling o Jarque Bera).

```
yeojohnson_result <- powerTransform(carbo_clean ~ 1)
lambda_yeojohnson <- yeojohnson_result$lambda
print(lambda_yeojohnson)
```

```
##        Y1
## 0.7633124
```

```
carbohydrates_yeojohnson <- bcPower(carbo_clean, lambda_yeojohnson)
```

6. Escribe la ecuación del modelo encontrado.

Para la transformación de Yeo Johnson se encontró la ecuación

$$Y(\lambda) = \begin{cases} \frac{(X+1)^{\lambda}-1}{\lambda}, & \text{si } X \geq 0 \\ \frac{(1-X)^{2-\lambda}-1}{2-\lambda}, & \text{si } X < 0 \end{cases} \quad \text{con } \lambda = 0.7633$$

7. Analiza la normalidad de las transformaciones obtenidas con los datos originales. Utiliza como argumento de normalidad:

7.1 Compara las medidas: Mínimo, máximo, media, mediana, cuartil 1 y cuartil 3, sesgo y curtosis.

7.2 Obten el histograma de los 2 modelos obtenidos (exacto y aproximado) y los datos originales.

7.3 Realiza la prueba de normalidad de Anderson-Darling para los datos transformados y los originales.

```
# Medidas

measures <- function(x){
  list(
    Min = min(x),
    Max = max(x),
    Mean = mean(x),
    Median = median(x),
    Q1 = quantile(x, 0.25),
    Q3 = quantile(x, 0.75),
    Skewness = skewness(x),
    Kurtosis = kurtosis(x)
  )}

original_measures <- measures(carbohydrates)
exact_measures <- measures(data_boxcox_exact)

cat("Medidas para los datos originales: \n")
```

```
## Medidas para los datos originales:
```

```
print(original_measures)
```

```
## $Min
## [1] 4
##
## $Max
## [1] 141
##
## $Mean
## [1] 50.45082
##
## $Median
## [1] 46
##
## $Q1
## 25%
##  34
##
## $Q3
## 75%
##  61
##
## $Skewness
## [1] 1.221876
##
## $Kurtosis
## [1] 4.802472
```

```
cat("Medidas para los datos exactos: \n")
```

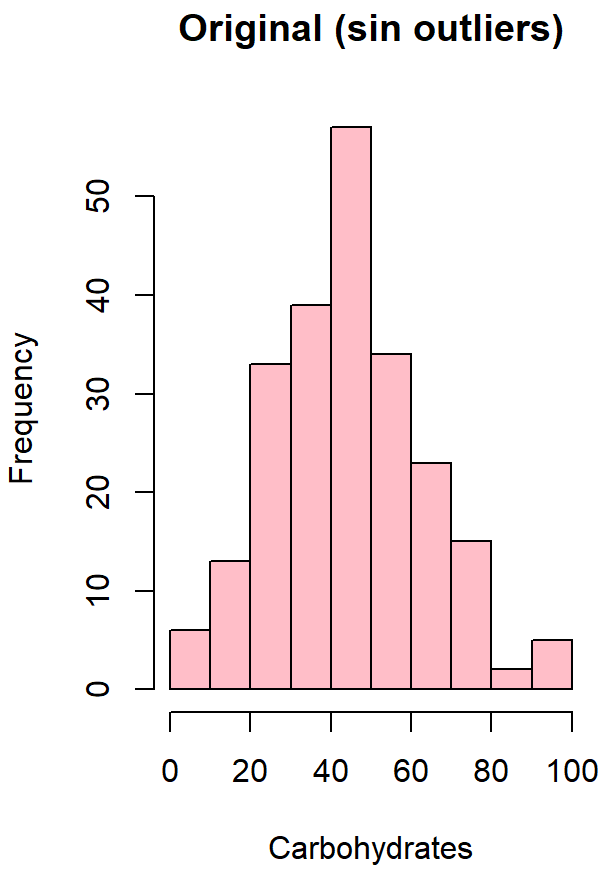
```
## Medidas para los datos exactos:
```

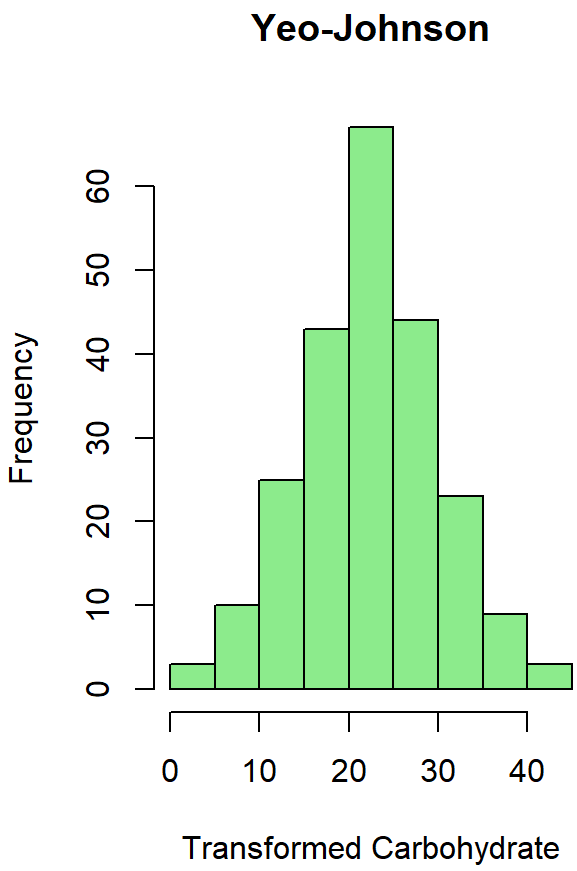
```
print(exact_measures)
```

```
## $Min
## [1] 1.775553
##
## $Max
## [1] 13.0203
##
## $Mean
## [1] 7.94662
##
## $Median
## [1] 7.932608
##
## $Q1
##     25%
## 6.86328
##
## $Q3
##      75%
## 9.036363
##
## $Skewness
## [1] -0.02247246
##
## $Kurtosis
## [1] 3.722577
```

```
# Histogramas

par(mfrow = c(1, 2))
hist(carbo_clean, main = "Original (sin outliers)", xlab = "Carbohydrates", col = "pink")
hist(carbohydrates_yeojohnson, main = "Yeo-Johnson", xlab = "Transformed Carbohydrate", col = "l
ightgreen")
```






```
# Anderson-Darling

ad_original <- ad.test(carbo_clean)
ad_exact <- ad.test(carbohydrates_yeojohnson)

cat("Prueba de normalidad de Anderson-Darling para los datos originales: \n")
```

```
## Prueba de normalidad de Anderson-Darling para los datos originales:
```

```
print(ad_original)
```

```
##
##  Anderson-Darling normality test
##
## data:  carbo_clean
## A = 0.39639, p-value = 0.367
```

```
cat("Prueba de normalidad de Anderson-Darling para los datos exactos: \n")
```

```
## Prueba de normalidad de Anderson-Darling para los datos exactos:
```

```
print(ad_exact)
```

```
##
##  Anderson-Darling normality test
##
## data:  carbohydrates_yeojohnson
## A = 0.21729, p-value = 0.8407
```

> 8. Define la mejor transformación de los datos de acuerdo a las características de los modelos que encontraste. Toma en cuenta los criterios del inciso anterior para analizar normalidad y la economía del modelo.

Habiendo hecho el análisis anterior, puedo decir que la mejor transformación fue la de Yeo Johnson ya que muestra una mejor normalidad en los datos. Igualmente, es más fácil ocupar esta transformación con bases de datos “sucias” ya que acepta diferentes tipos de valores, a diferencia que la transformación de BoxCox, la cual solo acepta valores positivos.

> 9. Concluye sobre las ventajas y desventajes de los modelos de Box Cox y de Yeo Johnson.

Como mencioné en la pregunta anterior, ambas transformaciones son buenas; sin embargo, creo que la de Yeo Johnson puede ser más flexible debido a que acepta valores de 0 o negativos, a diferencia de la transformación BoxCox. A pesar de esto, la transf. BoxCox tiene ventajas tales como que mejora la normalidad en datos ya limpios, y además de eso es mucho más conocida, por lo que podemos ocuparla de manera más fácil gracias a su presencia en diferentes librerías, etc, cosa que no pasa mucho con la de Yeo Johnson.

> 10. Analiza las diferencias entre la transformación y el escalamiento de los datos:

> > 10.1 Escribe al menos 3 diferencias entre lo que es la transformación y el escalamiento de los datos. 10.2 Indica cuándo es necesario utilizar cada uno.

Tres diferencias entre la transformación y el escalamiento de datos es que:

- La transformación se ocupa para que los datos sigan un comportamiento específico, como el de una distribución normal. Mientras que el escalamiento de datos es para poner diferentes tipos de datos en un mismo rango o escala, y así poder sacar información significativa de ellos.
- La transformación cambia la distribución y el escalamiento no altera los datos de manera que tengan que seguir una distribución.

- Obvio todo depende del problema, pero el escalamiento sirve para modelos específicos de Machine Learning como lo son las redes neurnales o algoritmos como KNN, y la transformación es más para modelos estadísticos que nos permiten ver el comportamiento de los datos.