Actividad13_Regresion_no_lineal

Rodolfo Jesús Cruz Rebollar

2024-09-12

El objetivo es encontrar el mejor modelo que relacione la velocidad de los automóviles y las distancias necesarias para detenerse en autos de modelos existentes en 1920 (base de datos car). La ecuación encontrada no sólo deberá ser el mejor modelo obtenido sino también deberá ser el más económico en terminos de la complejidad del modelo.

Parte 1: Análisis de normalidad

```
# Librería car con base de datos

library(car)

## Loading required package: carData

# Acceder a la base de datos de cars en R

carros = cars

# Verificar la importación correcta de los archivos

head(carros)

##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

Pruebas de normalidad univariada para distancia

Prueba de hipótesis para normalidad:

H_0: la variable distancia sigue una distribución normal.

H_1: la variable distancia no sigue una distribución normal.

Prueba de normalidad Shapiro Wilk

```r
# Realizar prueba de shapiro wilk para distancia

shapiro.test(carros$dist)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  carros$dist
## W = 0.95144, p-value = 0.0391
```

Prueba de normalidad de Jarque Bera

```r
# Realizar prueba de Jarque Bera para distancia

library(moments)

jarque.test(carros$dist)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  carros$dist
## JB = 5.2305, p-value = 0.07315
## alternative hypothesis: greater
```

Pruebas de normalidad univariada para velocidad

Prueba de normalidad Shapiro Wilk

```r
# Realizar prueba de shapiro wilk para velocidad

shapiro.test(carros$speed)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  carros$speed
## W = 0.97765, p-value = 0.4576
```

Prueba de normalidad de Jarque Bera

```r
# Realizar test de Jarque Bera para velocidad

jarque.test(carros$speed)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  carros$speed
## JB = 0.80217, p-value = 0.6696
## alternative hypothesis: greater
```

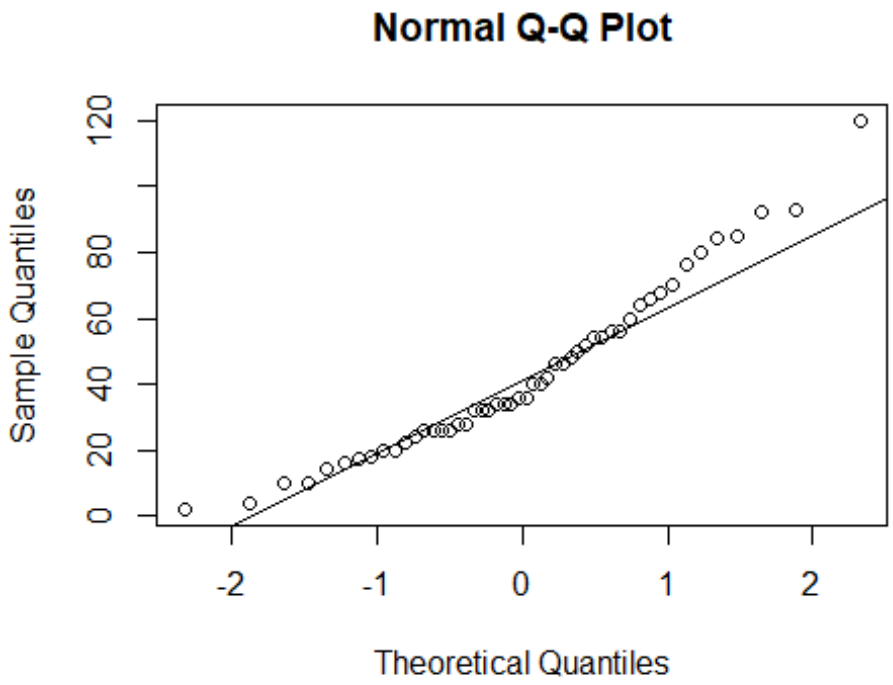
En los tests anteriores se observa que para el caso de la variable distancia, el p valor del test de normalidad de Shapiro Wilk es igual a 0.0391, lo cual al ser inferior a 0.05, se tiene evidencia estadística para rechazar H_0, por lo cual, se afirma que la variable distancia no sigue una distribución normal de acuerdo con Shapiro Wilk, además, el p valor del test de Jarque Bera es igual a 0.07315, lo cual al ser mayor a 0.05, no se rechaza H_0, por lo que se tiene evidencia estadística para rechazar H_0, por lo que se afirma que la distancia sigue una distribución normal de acuerdo con el test de Jarque Bera.

Por otro lado, para el caso de la variable velocidad, el p valor del test de Shapiro Wilk es igual a 0.4576, lo cual al ser mayor a 0.05, se tiene evidencia estadística para no rechazar H_0, motivo por el cual, se puede afirmar que según el test de Shapiro Wilk, la variable velocidad sigue una distribución normal. Por otro lado, el p valor del test de Jarque Bera para la velocidad es igual a 0.6696, lo cual al ser superior a 0.05, se tiene evidencia estadística para no rechazar H_0, por lo cual se afirma que en base al test de Jarque Bera, la variable velocidad sigue una distribución normal. En resumen, la variable velocidad sí sigue una distribución normal en base a ambos tests realizados, sin embargo la variable distancia sigue distribución normal según Jarque Bera, pero no de acuerdo a Shapiro Wilk, por lo que nos quedaremos con la prueba de Jarque Bera, ya que es la prueba de normalidad más exigente de todas para verificar normalidad, aún más que la de Shapiro Wilk, por lo que tomando esto en cuenta, podemos afirmar que la variable distancia también sigue una distribución aproximadamente normal.

Gráficos para identificar posibles alejamientos de normalidad

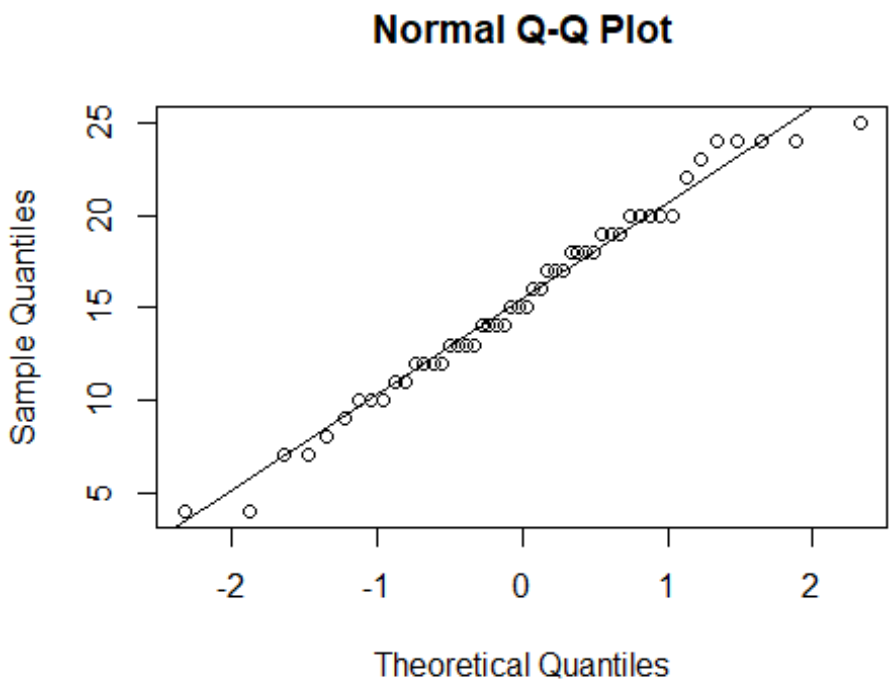
QQ-plot de los datos

```
# Gráfico de qq-plot de los datos de distancia

qqnorm(carros$dist)

qqline(carros$dist)
```



En el qq-plot de la variable distancia se puede observar que los puntos del gráfico se encuentran en su mayoría sobre la recta de normalidad, o al menos muy cerca de la misma, sin embargo también hay otros puntos que se alejan de la recta ideal de normalidad, sin embargo dado que la gran mayoría de puntos del qq-plot se ubican sobre la recta ideal de normalidad, mientras que solo unos pocos se alejan de la recta principal (sobretodo en los extremos de dicha recta), se puede afirmar que de acuerdo al qq-plot y a los tests de normalidad para la variable distancia, los datos de dicha variable siguen una distribución normal.

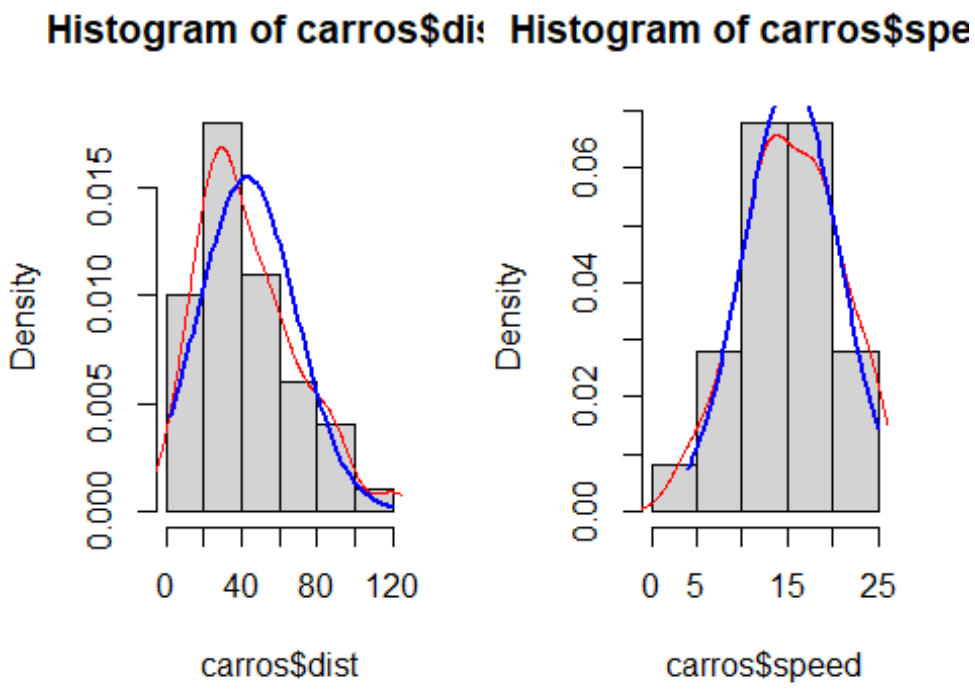
```
# QQplot de los datos de velocidad

qqnorm(carros$speed)

qqline(carros$speed)
```



En cuanto al qq-plot de la variable velocidad, se aprecia que los puntos en el gráfico se encuentran en su gran mayoría ubicados sobre la recta ideal de normalidad, sin embargo, nuevamente se presentan ligeros alejamientos de la normalidad, en los extremos de la recta, ya que algunos de los puntos del gráfico se desvían de la recta principal de normalidad, no obstante las desviaciones de la normalidad son mínimas, por lo que se concluye que la velocidad sigue una distribución normal.

Histograma y su distribución teórica de probabilidad

```
par(mfrow = c(1, 2))

hist(carros$dist, freq = FALSE)

lines(density(carros$dist),col = "red")

curve(dnorm(x, mean = mean(carros$dist), sd = sd(carros$dist)),
      from = min(carros$dist), to = max(carros$dist), add = TRUE, col =
"blue",lwd = 2)

hist(carros$speed, freq = FALSE)

lines(density(carros$speed),col = "red")

curve(dnorm(x, mean = mean(carros$speed), sd = sd(carros$speed)),
      from = min(carros$speed), to = max(carros$speed), add = TRUE, col =
"blue",lwd = 2)
```



En cuanto a los histogramas y distribuciones teóricas y empíricas de probabilidad para cada variable, se puede observar que en el caso de la variable distancia, la distribución empírica de los datos en color rojo se encuentra ligeraente desviada hacia la izquierda con respecto a la distribución teórica de dichos datos en color azul, lo cual indica que la distribución empírica tiene un ligero sesgo positivo con respecto a la distribución teórica de los datos de distancia, por lo cual, dado que el alejamiento entre ambas distribuciones es prácticamente reducido, se puede afirmar que los datos de distancia sí siguen una distribución normal. Además respecto a la velocidad, la distribución empírica de los datos en color rojo igualmente muestra un ligero alejamiento hacia abajo con respecto a la distribucíon teórica de los datos en color azul, lo cual indica que la distribución empírica de velocidad posee una curtosis ligeramente menor que la distribución teórica de velocidad, motivo por el cual dado que para velocidad, ambas distribuciones poseen un alejamiento mayormente reducido, entonces se puede afirmar que la variable velocidad en efecto sigue una distribución normal.

Sesgo y curtosis por variable

Sesgo y curtosis de distancia

```
# Calcular sesgo de la distancia

cat("Sesgo de distancia: ", skewness(carros$dist), "\n")

## Sesgo de distancia:  0.7824835
```

```r
# Calcular curtosis de la distancia

cat("Curtosis de distancia: ", kurtosis(carros$dist))
```

```
## Curtosis de distancia:  3.248019
```

En cuanto al sesgo y la curtosis para la distancia, se aprecia que el sesgo es de 0.78, mientras que la curtosis es de 3.24, lo cual señala que la variable distancia posee un sesgo ligeramente positivo junto con una curtosis muy alta, por lo cual, se puede afirmar que de acuerdo al sesgo y la curtosis, la variable distancia sigue una distribución que es aproximadamente normal, dado que al tener alta curtosis, dicha distribución no es completamente normal todavía, por lo que en realidad según la curtosis, la distribución es de tipo leptocúrtica, ya que se tiene una alta curtosis provocando que la distribución de los datos tenga un pico muy marcado hacia arriba, no obstante la distribución de la distancia sí se puede considerar como aproximadamente normal.

Sesgo y curtosis de velocidad

```r
# Calcular sesgo de la velocidad

cat("Sesgo de velocidad: ", skewness(carros$speed), "\n")
```

```
## Sesgo de velocidad:  -0.1139548
```

```r
# Calcular curtosis de la velocidad

cat("Curtosis de velocidad: ", kurtosis(carros$speed))
```

```
## Curtosis de velocidad:  2.422853
```

En cuanto a la variable velocidad, se puede apreciar que el sesgo tiene un valor de -0.11, mientras que la curtosis tiene un valor de 2.42, lo cual indica que la distribuión de los datos de velocidad tiene un sesgo ligeramente negativo además de alta curtosis, indicando que la distribución posee un pico hacia arriba, es decir es de tipo leptocúrtica, no obstante, dado que el sesgo es pequeño y la curtosis es alta pero no sobrepasa por mucho a la de una distribución normal ideal, se concluye que la variable velocidad también sigue una distribución aproximadamente normal.

Conclusión final parte 1

En resumen, tomando en cuenta todos los análisis anteriores, es posible concluir que ambas variables, tanto la distancia como la velocidad, siguen una distribución que es aproximadamente normal, esto a pesar de que según el análisis del sesgo y la curtosis de las distribuciones, éstas mismas resultan tener una curtosis alta, por lo que se consideran leptocúrticas, sin embargo, el exceso de curtosis en las distribuciones no ocasiona que éstas mismas se alejen demasiado de la distribución normal ideal, además de lo anterior, considerando también que los gráficos de las distribuciones empíricas y teóricas de los datos por cada variable no muestran signos de que las distribuciones empíricas (obtenidas a partir de los datos analizados) se alejen

considerablemente de sus distribuciones teóricas correspondientes, entonces en base a todo lo anteriormente mencionado se puede concluir finalmente que ambas variables (distancia y velocidad) siguen una distribución que resulta ser aproximadamente normal. Además, los alejamientos de normalidad evidenciados en los gráficos previos se pueden deber a que posiblemente exista presencia de datos atípicos o influyentes en el dataset, además dichos alejamientos también se deben al hecho de que para ambas variables, la curtosis tiene un valor considerablemente alto, provocando que la distribución de ambas variables se aleje hacia arriba con respecto a la distribución normal ideal, además de que el sesgo presente en dichas distribuciones contribuye a desplazar horizontalmente las distribuciones, ocasionando también un alejamiento de la normalidad pero en horizontal, mientras que la curtosis lo hace en forma vertical.

Parte 2: Regresión lineal

Prueba de regresión lineal entre distancia y velocidad

```
# Generar un modelo de regresión lineal entre distancia y velocidad

linear.model = lm(dist ~ speed, data = carros)

linear.model

##
## Call:
## lm(formula = dist ~ speed, data = carros)
##
## Coefficients:
## (Intercept)        speed
##     -17.579        3.932
```

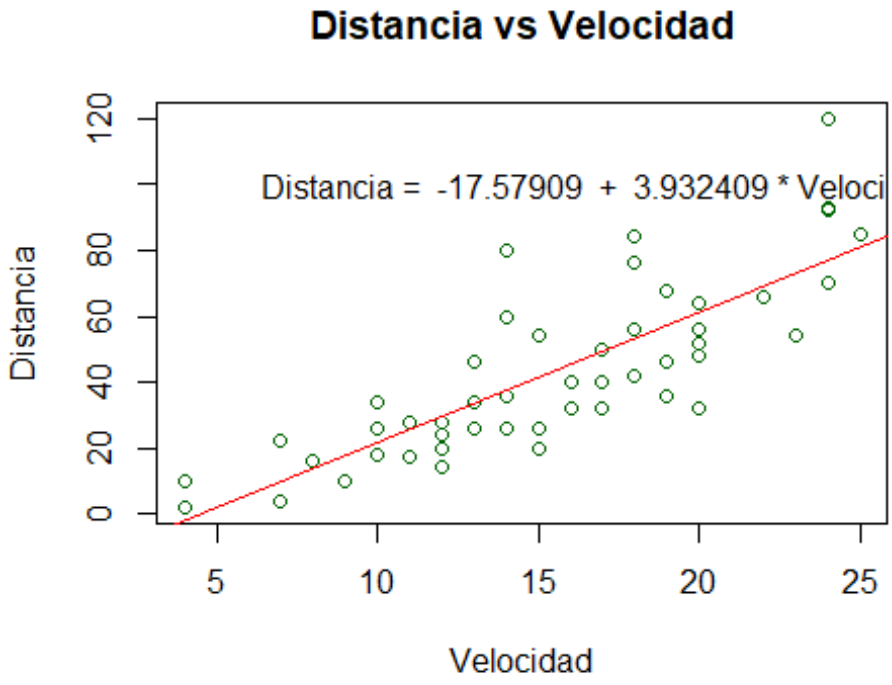
Escribe el modelo lineal obtenido

```
# Modelo lineal de velocidad en función de distancia

cat("Distancia = ", linear.model$coefficients[1], " + ",
linear.model$coefficients[2],
    "* Velocidad")

## Distancia =  -17.57909  +  3.932409 * Velocidad
```

Grafica los datos y el modelo (ecuación) que obtuviste

```
# Gráfico de los datos de velocidad vs distancia

plot(carros$speed, carros$dist, xlab = "Velocidad", ylab = "Distancia",
     main = "Distancia vs Velocidad", col = "darkgreen")

# Modelo obtenido gráficamente
```

```r
abline(linear.model, col = "red")

# Ecuación del modelo en el gráfico

text(17, 100, "Distancia =  -17.57909  +  3.932409 * Velocidad")
```



Analiza significancia del modelo: individual, conjunta y coeficiente de determinación

Prueba de hipótesis para significancia de β_i:

H_0: $\beta_i = 0$ (el coeficiente β_i no es estadísticamente significativo)

H_1: $\beta_i \neq 0$ (el coeficiente β_i sí es estadísticamente significativo)

Prueba de hipótesis para la significancia del modelo:

H_0: el modelo es estadísticamente significativo.

H_1: el modelo no es estadísticamente significativo.

```r
# Desplegar el summary del modelo para analizar la significancia por
variable, del modelo
# en general y el coeficiente de determinación del modelo

summary(linear.model)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = carros)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

En el summary del modelo lineal se puede observar que el p valor del intercepto del modelo β_0 es igual a 0.0123, lo cual es menor a 0.05, por lo que se tiene evidencia estadística para rechazar H_0, por lo que se concluye que β_0 sí es estadísticamente significativo, además, también se observa que el p valor para el coeficiente β_1 de la variable velocidad es igual a 1.49e-12, lo cual al ser menor que 0.05 indica que se tiene evidencia estadística para rechazar H_0, indicando así que el coeficiente β_1 asociado a la variable predictora de velocidad sí es estadísticamente significativo. Adicionalmente, en cuanto a la significancia conjunta del modelo, se aprecia que en el resumen del mismo, el p valor para todo el modelo en general es igual a 1.49e-12, lo cual al ser menor a 0.05, se tiene evidencia estadística para rechazar H_0 de la hipótesis asociada a la significancia del modelo, por lo cual, se puede concluir que en general, el modelo lineal obtenido sí es estadísticamente significativo, indicando que el modelo sí resulta útil desde la perspectiva estadística para explicar el patrón existente detrás de los datos originales. Por otro lado, en cuanto al coeficiente de determinación del modelo obtenido, se puede apreciar que éste mismo tiene un valor de 0.6511 (se usa el R^2 ajustado porque solamente se tiene 1 variable independiente, por lo que no es necesario ajustar el modelo a otras variables), lo cual significa que el modelo lineal obtenido es capaz de explicar el 65.11% de la variabilidad total presente en los datos originales, siendo un porcentaje mayormente considerable de variabilidad explicada por el modelo, por lo cual, es posible concluir que el modelo junto con sus variables son estadísticamente significativos, además de que el modelo explica un porcentaje mayormente elevado de la variabilidad de los datos originales.

Analiza validez del modelo

Residuos con media cero

Prueba de hipótesis:

$H_0: \mu_R = 0$ (media de los residuos es cero)

$H_1: \mu_R \neq 0$ (media de los residuos no es cero)

```
# Verificar que la media de los residuos del modelo sea igual a cero

t.test(linear.model$residuals, mu = 0, alternative = "two.sided",
conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  linear.model$residuals
## t = 1.0315e-16, df = 49, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -4.326  4.326
## sample estimates:
##    mean of x
## 2.220446e-16
```

Dado que el p valor del test es igual a 1, entonces no se rechaza H_0, ya que 1 es mayor a 0.05, por lo cual, eso significa que la media de los residuos es igual a cero.

Normalidad de residuos

Prueba de hipótesis:

H_0 los residuos siguen una distribución normal.

H_1: los residuos no siguen una distribución normal.

```
# Realizar prueba de normalidad de Jarque Bera de los residuos del modelo
lineal

jarque.test(linear.model$residuals)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  linear.model$residuals
## JB = 8.1888, p-value = 0.01667
## alternative hypothesis: greater
```

Dado que se observa que el p valor del test es igual a 0.01667, lo cual es menor a 0.05, se concluye que se rechaza H_0, lo cual implica que los residuos del modelo lineal no siguen una distribución normal.

Homocedasticidad, independencia, linealidad

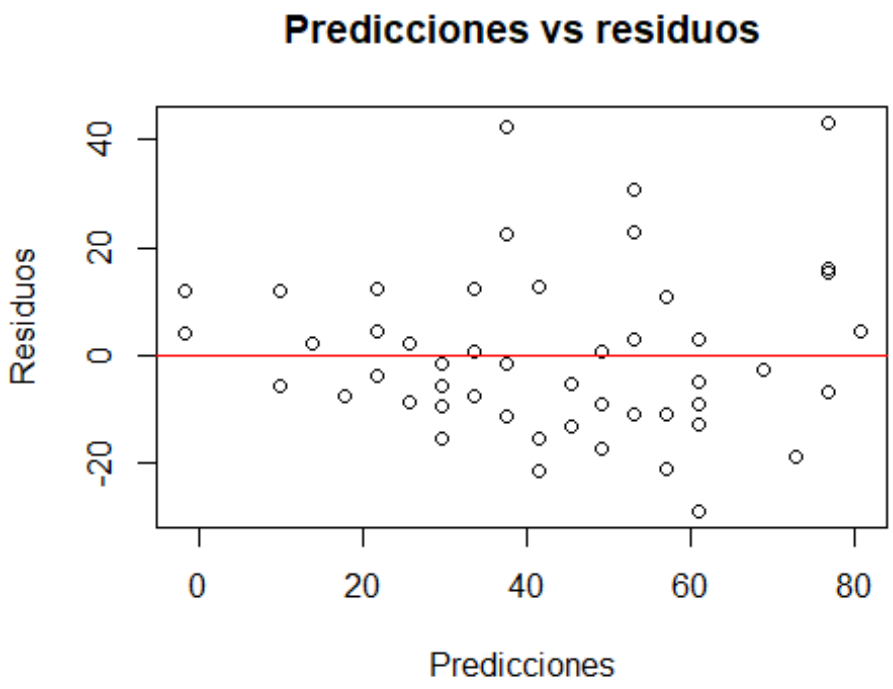
Homocedasticidad

Prueba de hipótesis:

H_0: la varianza de los residuos es constante (sí hay homocedasticidad).

H_1: la varianza de los residuos no es constante (no hay homocedasticidad).

```
# Realizar prueba de Breusch-Pagan para determinar homocedasticidad de
residuos

library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

# Prueba de homocedasticidad de Breusch-Pagan

bptest(linear.model)

##
##  studentized Breusch-Pagan test
##
## data:  linear.model
## BP = 3.2149, df = 1, p-value = 0.07297

# Graficar predicciones del modelo vs residuos del mismo

plot(linear.model$fitted.values, linear.model$residuals,
     main = "Predicciones vs residuos", xlab = "Predicciones", ylab =
"Residuos")

abline(h = 0, col = "red")
```



Dado que en el gráfico de predicciones vs residuos se observa que los residuos graficados se encuentran dispersos a lo larg del gráfico sin evidenciar algún patrón o tendencia en particular, además de que en la prueba de normalidad de Jarque Bera, el p valor fue de 0.01667, con lo cual, se rechaza la normalidad presente en los residuos del modelo, por lo que en general el modelo lineal no cumple con el supuesto de normalidad.

Independencia

Prueba de hipótesis:

H_0: los residuos no están autocorrelacionados entre sí (sí son independientes).

H_1: los residuos sí están autocorrelacionados entre sí (no son independientes).

```
# Realizar prueba de Durbin Watson para independencia

dwtest(linear.model)

##
##  Durbin-Watson test
##
## data:  linear.model
## DW = 1.6762, p-value = 0.09522
## alternative hypothesis: true autocorrelation is greater than 0
```

Dado que en el gráfico previo se observa que los residuos del modelo no siguen algún patrón o tendencia en particular, además de que también se aprecia que en el test de independencia de Durbin Watson el p valor resultante es de 0.09522, lo cual es mayor a 0.05, por lo que no se rechaza H_0 y por tanto, teniendo en cuenta todo lo anterior, se puede concluir que los residuos del modelo no presentan autocorrelación, es decir que son independientes.

Linealidad

Prueba de hipótesis:

H_0: no hay términos omitidos que indican linealidad.

H_1: hay una especificación errónea en el modelo que indica no linealidad.

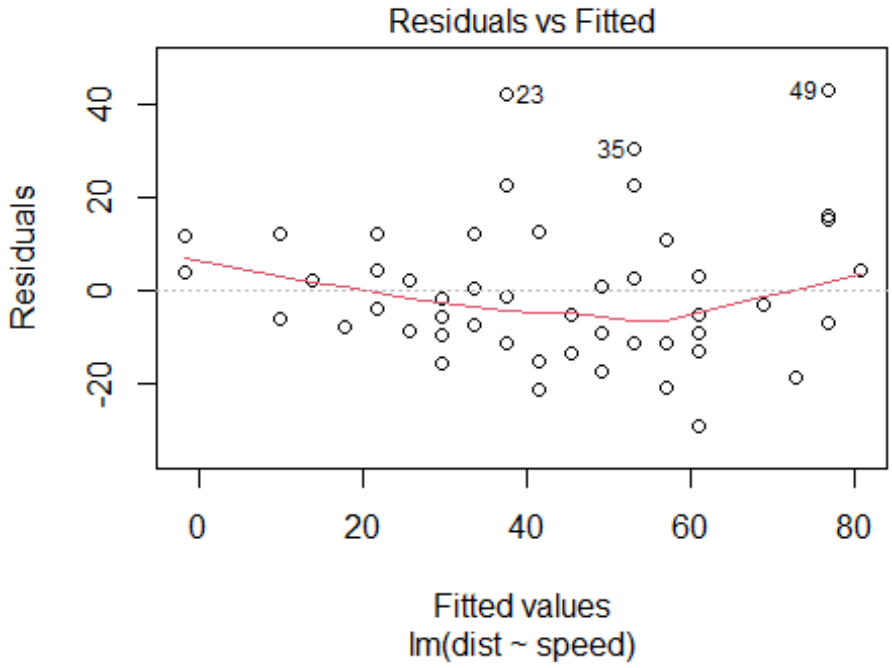
```
# Realizar test de RESET de Ramsey para detectar linealidad

resettest(linear.model)
```

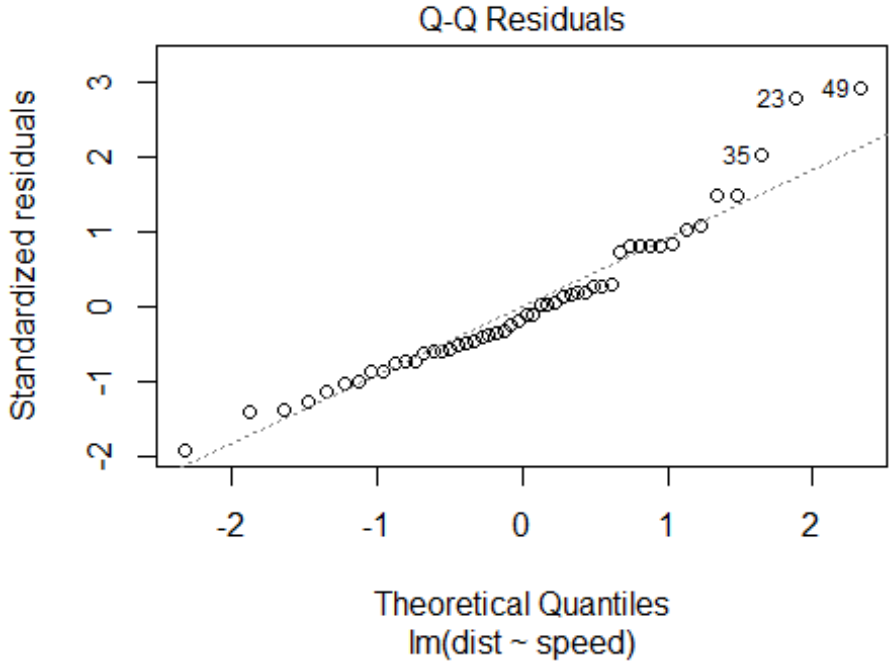
```
##
##  RESET test
##
## data:  linear.model
## RESET = 1.5554, df1 = 2, df2 = 46, p-value = 0.222
```

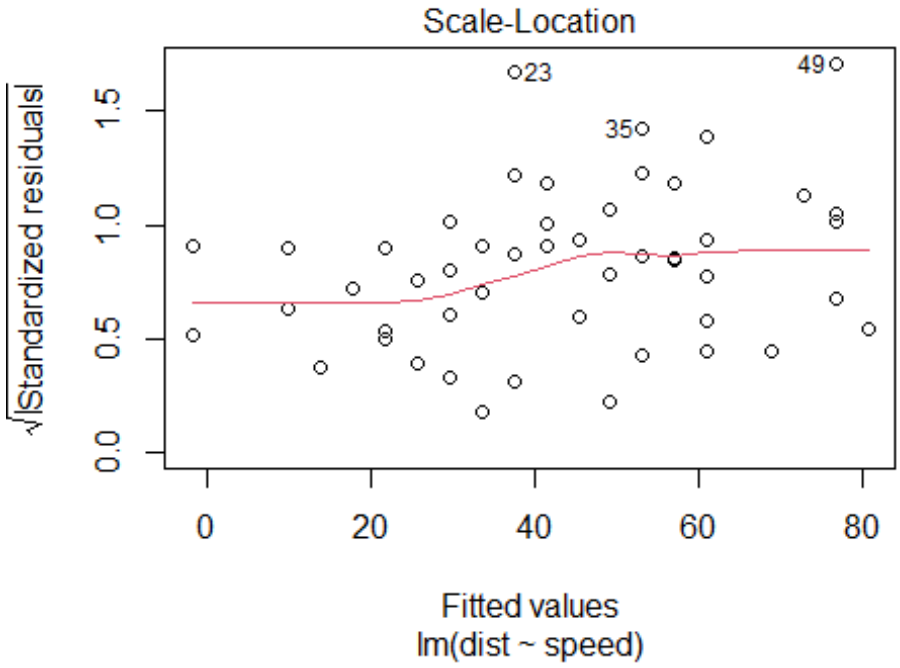
Dado que se observa que el p valor del test RESET de Ramsey es igual a 0.222, esto resulta ser mayor que 0.05, por lo que se tiene evidencia estadística para no rechazar H_0, por lo cual, se puede concluir que el modelo lineal obtenido en efecto cumple con el supuesto de linealidad.

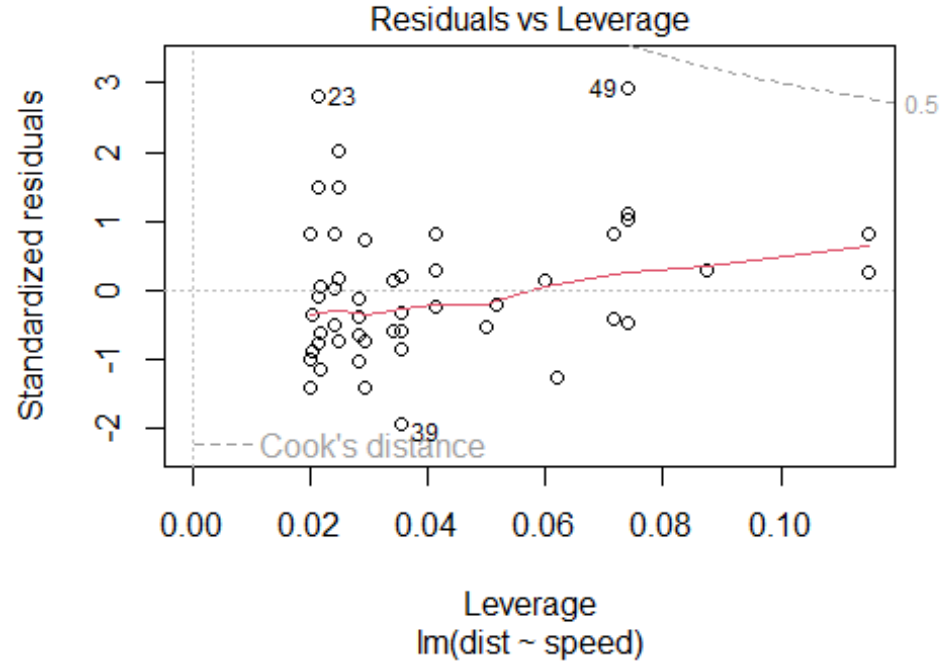
Plot(modelo) para los gráficos del modelo

```
# Realizar gráficas para los diferentes atributos del modelo lineal
generado

plot(linear.model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(dist ~ speed)
23
35
49
40
20
0
-20
0
20
40
60
80


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(dist ~ speed)
23
35
49
3
2
1
0
-1
-2
-2
-1
0
1
2

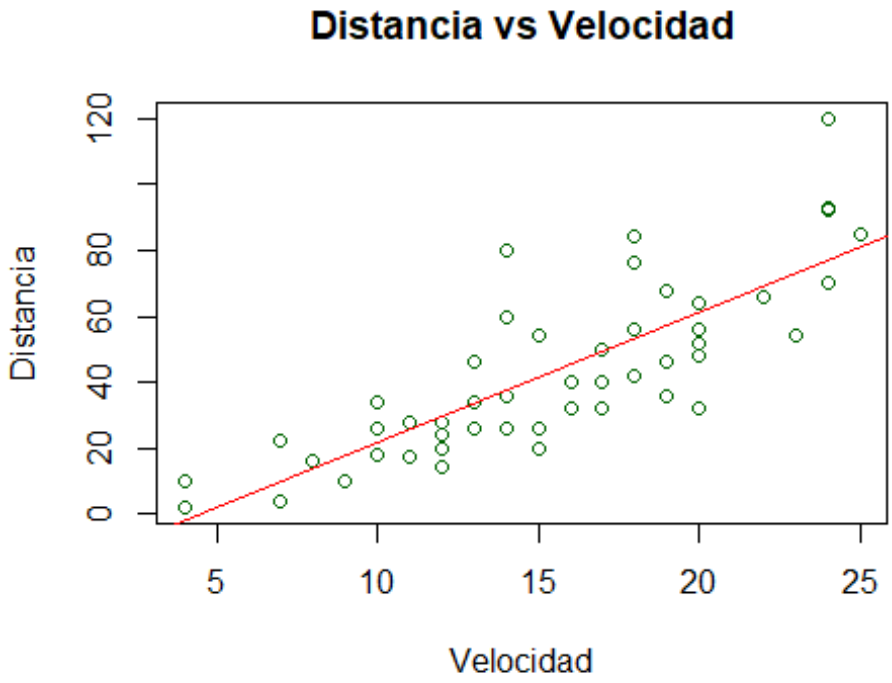




Grafica los datos y el modelo de distancia en función de la velocidad

```
# Gráfico de los datos de velocidad vs distancia

plot(carros$speed, carros$dist, xlab = "Velocidad", ylab = "Distancia",
```

```
     main = "Distancia vs Velocidad", col = "darkgreen")

# Modelo obtenido gráficamente

abline(linear.model, col = "red")
```



Comenta sobre la idoneidad del modelo en función de significancia y validez

En resumen, se puede concluir que el modelo lineal obtenido resulta ser estadísticamente significativo en general además de que todos sus coeficientes son significativos, lo cual indica que todas las variables predictoras del modelo son estadísticamente significativas también, por lo que no hay varables en el modelo que no aporten información para predecir la distancia en función de la velocidad, además de que también el modelo lineal obtenido satisface todos los supuestos de una regresión lineal, a excepción del supuesto de normalidad, por lo que será necesario realizar transformaciones posteriormente para intentar que los residuos del modelo cumplan dicho supuesto, por lo que el modelo lineal obtenido aún no es el ideal para predecir distancia en función de velocidad.

Parte 3: Regresión no lineal

Dado que el test de normalidad de Jarque Bera es de los más robustos, nos basaremos en este criterio para decidir a cuál de ambas variables aplicar la transformación de

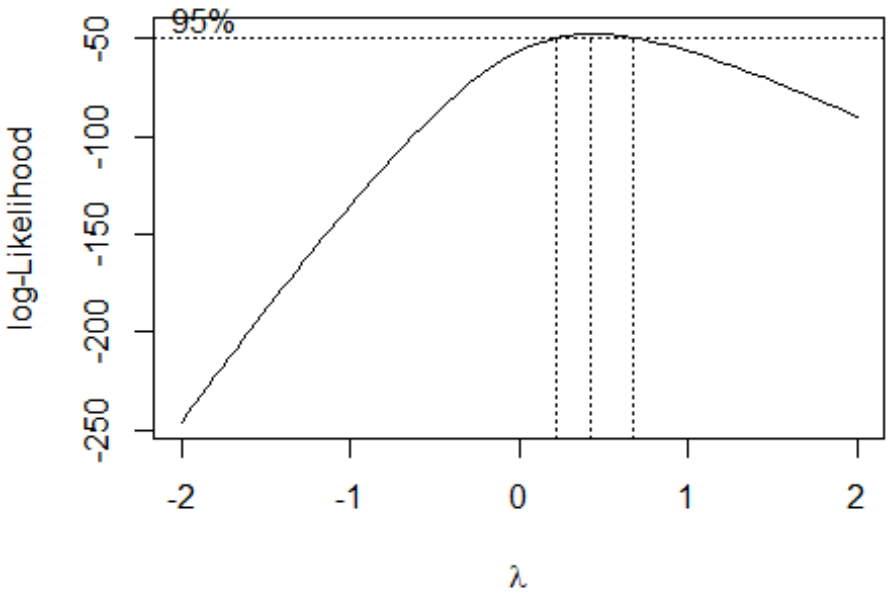
normalidad, por lo que aquella que resulte tener el menor p valor en el test de Jarque Bera, será aquella que tendrá el mayor alejamiento de la normalidad y será por tanto la que tendremos que normalizar:

```
# Jarque Bera para distancia

jarque.test(carros$dist)

##
##  Jarque-Bera Normality Test
##
## data:  carros$dist
## JB = 5.2305, p-value = 0.07315
## alternative hypothesis: greater

# Jarque Bera para velocidad

jarque.test(carros$speed)

##
##  Jarque-Bera Normality Test
##
## data:  carros$speed
## JB = 0.80217, p-value = 0.6696
## alternative hypothesis: greater
```

En los test de Jarque Bera se aprecia que el p valor menor pertenece a la variable de distancia, por lo cual a pesar de que dicho p valor sea mayor que 0.05 y cumpla normalidad ligeramente, aún así será necesario normalizar los datos de distancia para que dicha variable cumpla con mayor fuerza el supuesto de normalidad.

Valor óptimo de λ en la transformación Box Cox para el modelo lineal

```
# Librería MASS para usar el comando boxcox()

library(MASS)

# Utilizar la función Box Cox para hallar el valor óptimo de lambda tal
que
# se maximize la función de verosimilitud para el modelo lineal
encontrado

boxcox_modelo = boxcox(linear.model)
```




```
# Buscar el valor óptimo de lambda tal que se maximize la función de
verosimilitud

lambda_mejor = boxcox_modelo$x[which(boxcox_modelo$y ==
max(boxcox_modelo$y))]

cat("Lambda óptimo = ", lambda_mejor)

## Lambda óptimo =  0.4242424
```

Transformación exacta y aproximada y sus ecuaciones

Transformación aproximada

Dado que el valor de λ óptimo es bastante más cercano a 0.5 que a 0, se utilizará la siguiente función como transformación aproximada de Box Cox:

$$y = \sqrt{(x)}$$

Donde y son los valores de distancia normalizados y x son los valores de distancia originales.

Nota: no se realiza ninguna traslación en los datos de distancia antes de aplicar la transformación de Box Cox puesto que no hay datos negativos o que sean iguales a 0:

```r
# Comprobar que no existan datos negativos o iguales a 0

summary(carros$dist)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   26.00   36.00   42.98   56.00  120.00
```

```r
# Aplicar la función anterior a los datos originales de distancia para normalizarlos

dist_aprox = sqrt(carros$dist)

# Mostrar los primeros valores de distancia ya transformados

head(dist_aprox)
```

```
## [1] 1.414214 3.162278 2.000000 4.690416 4.000000 3.162278
```

Transformación exacta

Para implementar la transformación de Box Cox exacta, se tomará en cuenta la función:

$$f(x,\lambda) = \frac{x^{\lambda} - 1}{\lambda}$$

Donde x son los datos de distancia originales sin transformar.

```r
# Transformar los datos originales de distancia usando la transformación exacta de BoxCox

dist_exacta = ((carros$dist ^ lambda_mejor) - 1) / lambda_mejor

# Mostrar los primeros datos de distancia transformados con la transformación exacta

head(dist_exacta)
```

```
## [1] 0.805831 3.903635 1.887150 6.390651 5.285168 3.903635
```

Ecuaciones de las transformaciones encontradas

Transformación aproximada:

$$D_{tr} = \sqrt{D}$$

Transformación exacta:

$$D_{tr} = \frac{D^{0.4242} - 1}{0.4242}$$

Donde D_{tr} son los valores de distancia transformados y D son los valores de distancia originales (sin transformar).

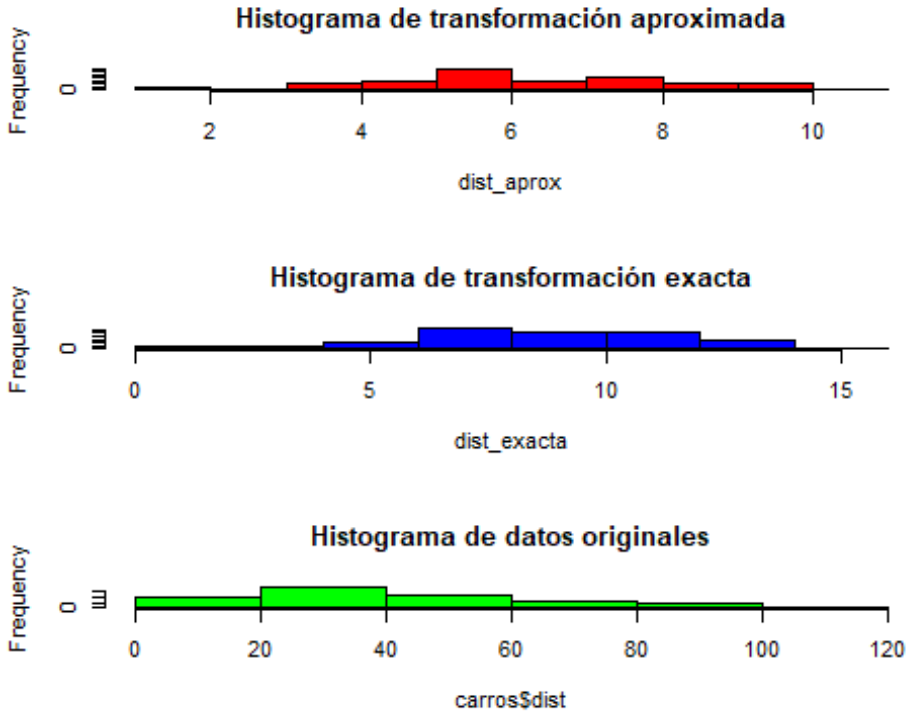
Análisis de normalidad de las transformaciones obtenidas

Comparación de sesgo y curtosis

```
# Sesgo y curtosis de la transformación aproximada de distancia

cat("Sesgo de distancia con tr aproximada: ", skewness(dist_aprox), "\n")

## Sesgo de distancia con tr aproximada:  -0.0196131

cat("Curtosis de distancia con tr aproximada: ", kurtosis(dist_aprox))

## Curtosis de distancia con tr aproximada:  2.796264

# Sesgo y curtosis de la transformación exacta de distancia

cat("Sesgo de distancia con tr exacta: ", skewness(dist_exacta), "\n")

## Sesgo de distancia con tr exacta:  -0.1753974

cat("Curtosis de distancia con tr exacta: ", kurtosis(dist_exacta))

## Curtosis de distancia con tr exacta:  2.929109
```

Histograma de transformaciones y datos originales

```
par(mfrow = c(3, 1))

hist(dist_aprox, col = "red", main = "Histograma de transformación
aproximada")

hist(dist_exacta, col = "blue", main = "Histograma de transformación
exacta")

hist(carros$dist, col = "green", main = "Histograma de datos originales")
```



Pruebas de normalidad para datos transformados

Recordando la prueba de hipótesis para verificar normalidad:

H_0: los datos transformados siguen una distribución normal.

H_1: los datos transformados no siguen una distribución normal.

Prueba de Jarque Bera para transformación aproximada

```
# Test de normalidad de Jarque Bera para distancia normalizada con modelo aproximado

jarque.test(dist_aprox)

##
##  Jarque-Bera Normality Test
##
## data:  dist_aprox
## JB = 0.089682, p-value = 0.9561
## alternative hypothesis: greater

# Test de normalidad de Jarque Bera para distancia normalizada con modelo exacto

jarque.test(dist_exacta)

##
##  Jarque-Bera Normality Test
```

```
##
## data:  dist_exacta
## JB = 0.26684, p-value = 0.8751
## alternative hypothesis: greater
```

De acuerdo al sesgo y la curtosis, además de los histogramas de las trnsformaciones vs los datos originales y los resultados de los tests de Jarque bera de los datos transformados, es posible observar que los datos de distancia transformados tanto con el modelo exacto como con el aproximado poseen un grado bajo de sesgo (-0.17 y -0.01 respectivamente), mientras que su grado de curtosis es mayormente alto (2.92 y 2.79 respectivamente), lo cual indica que la distribución de los datos transformados mediante ambos métodos posee un pico hacia arriba altamente evidente mientras que tiene un sesgo muy bajo casi nulo, por lo cual las distribuciones de los datos transformados poseen alejamiento vertical con respecto a la distribución normal ideal, no obstante dicho alejamiento no es mayormente significativo, por lo que se puede afirmar que las distribuciones son aproximadamente normales. Por otro lado, tomando en cuenta los histogramas de las transformaciones contra los datos originales, se observa que la distribución de los datos transformados tanto con la transformación aproximada como exacta es aproximadamente simétrica, además de que también se reduce el sesgo que tienen los nuevos datos transformados con respecto a los datos iniciales, provocando que las nuevas distribuciones de los datos transformados se acerquen más a la normalidad que la distribución de los datos de distancia originales, además, en los test de normalidad de Jarque Bera realizados para los datos transformados con ambos métodos (exacto y aproximado), se obtuvo un p valor de 0.8751 y 0.9561, los cuales son mayores que 0.05, por lo cual, no se rechaza H_0 de la hipótesis de normalidad y por tanto, se concluye que en efecto, los datos transformados usando ambas transformaciones de Box Cox siguen una distribución normal.

Conclusión sobre las 2 transformaciones realizadas

En términos generales, después de normalizar la distancia con ambas transformaciones, se puede concluir que la mejor transformación de ambas resulta ser la transformación de Box Cox aproximada, debido a que es aquella que posee un valor p en la prueba de normalidad de Jarque Bera igual a 0.9561, contra el p valor de 0.8751 de la transformación exacta, lo cual indica que la transformación aproximada cumple con más fuerza el supuesto de normalidad, además de que también con la transformación aproximada se garantiza un menor grado tanto de sesgo como de curtosis en la distribución de los datos transformados, dado que el sesgo y curtosis asociados a la transformación aproximada son -0.019 y 2.796 respectivamente, mientras que el sesgo y curtosis asociados a la transformación exacta son -0.17 y 2.92 respectivamente, además ambas transformaciones toman en cuenta la economía del modelo ya que solamente se considera una sola variable independiente en el mismo, por lo que se explica el mayor porcentaje posible de variabilidad de los datos originales con el nmínimo de variables independientes, motivo por el cual, tomando en cuenta todo lo descrito con anterioridad, es posible concluir que se escogerá a la transformación aproximada de Box Cox como la mejor de ambas transformaciones.

Regresión lineal simple entre la mejor transformación y la velocidad

Modelo lineal para la transformación

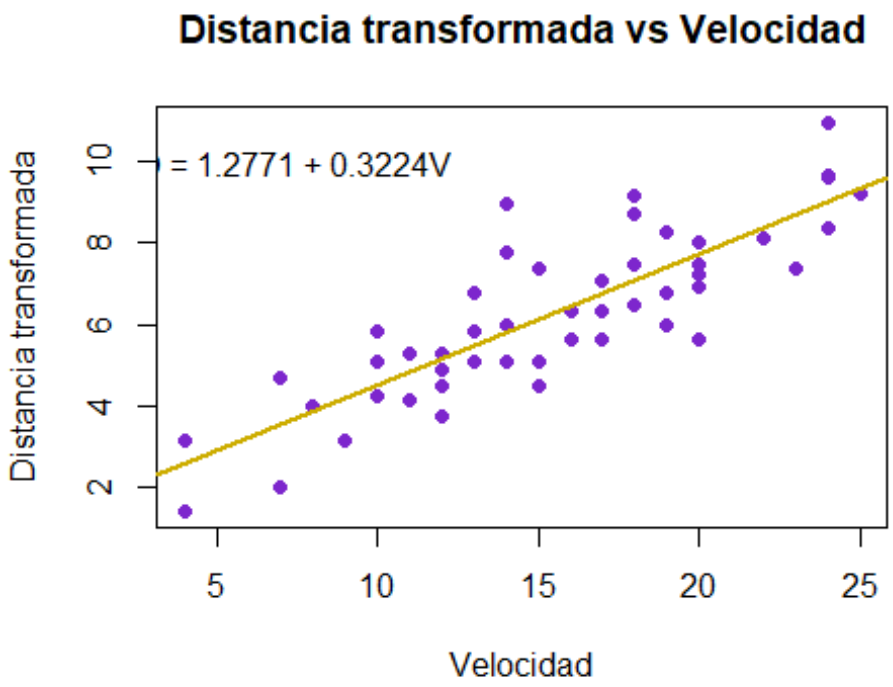
$$D_{tr} = \sqrt{D}$$

Donde D_{tr} se refiere a la distancia normalizada y D hace referencia a la distancia original sin normalizar.

```
# Crear un modelo de regresión que contemple la distancia normalizada en
función de la
# velocidad

reg_tr = lm(dist_aprox ~ carros$speed)
```

Gráfica de datos y del modelo lineal de la transformación elegida vs velocidad

```
# Graficar los datos de velocidad originales contra la distancia
transformada

plot(carros$speed, dist_aprox, pch = 19, col = "purple3", xlab =
"Velocidad",
     ylab = "Distancia transformada", main = "Distancia transformada vs
Velocidad")

# Graficar recta del modelo lineal de la transformación elegida vs
velocidad

abline(reg_tr, col = "gold3", lwd = 2)

# Colocar ecuación del modelo de la transformación elegida en el gráfico
# D: distancia transformada, V: velocidad original

text(7.5, 10, "D = 1.2771 + 0.3224V")
```



Análisis de significancia del modelo con distancia normalizada

Prueba de hipótesis para significancia del modelo:

H_0: el modelo no es estadísticamente significativo.

H_1: el modelo sí es estadísticamente significativo.

Prueba de hipótesis para significancia de coeficientes β:

$H_0: \beta_i = 0$ (β_i no es estadísticamente significativo)

$H_1: \beta_i \neq 0$ (β_i sí es estadísticamente significativo)

```
# Mostrar resumen del modelo para analizar la significancia en general
del mismo, así como
# de sus variables

summary(reg_tr)
```

```
##
## Call:
## lm(formula = dist_aprox ~ carros$speed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0684 -0.6983 -0.1799  0.5909  3.1534
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.27705    0.48444   2.636   0.0113 *
## carros$speed  0.32241    0.02978  10.825 1.77e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.102 on 48 degrees of freedom
## Multiple R-squared:  0.7094, Adjusted R-squared:  0.7034
## F-statistic: 117.2 on 1 and 48 DF,  p-value: 1.773e-14
```

```
# Calcular coeficiente de correlación entre los datos de distancia
transformados y
# los datos originales de la velocidad

r = cor(carros$speed, dist_aprox)

cat("coeficiente de correlación entre velocidad y distancia transformada:
", r)
```

```
## coeficiente de correlación entre velocidad y distancia transformada:
0.8422666
```

En el resumen del nuevo modelo generado se observa que en cuanto a la signficancia conjunta del modelo en sí, el p valor del modelo en general es igual a 1.773e-14, lo cual es mucho menor que 0.05, por lo que se cuenta con evidencia estadística para rechazar H_0, motivo por el cual se concluye que en cuanto a la significancia conjunta del modelo, éste mismo sí resulta ser estadísticamente significativo, por lo que sí resulta útil estadísticamente hablando para explicar la relación existente entre la distancia y la velocidad. Por otro lado, en cuanto a la significancia de los coeficientes β_i, se observa que el coeficiente β_0 posee un p valor de 0.0113, lo cual al ser menor que 0.05 se tiene evidencia estadística para rechazar H_0, por lo cual, se afirma que el coeficiente β_0 sí es estadísticamente significativo, además en cuanto al coeficiente β_1 de la variable speed (velocidad), éste tiene un p valor de 1.77e-14, lo cual al ser mucho menor que 0.05, se tiene evidencia estadística para rechazar H_0, por lo tanto, se puede afirmar que el coeficiente β_1 sí es estadísticamente significativo y con ello, la variable speed (velocidad) asociada a dicho coeficiente también lo es.

Adicionalmente, en cuanto al coeficiente de correlación del nuevo modelo generado, se observa que éste coeficiente posee un valor de 0.8422, lo cual es aprentemente alto, no obstante, considerando que debido a la propia naturaleza y distribución de los datos vista en los gráficos anteriores, no resulta posible ajustar un modelo de regresión lineal a los nuevos datos en cuestión, por lo cual, la verdadera relación entre las variables velocidad y diatancia ya no es de carácter lineal, motivo por el cual, ya no es posible utilizar el coeficiente de correlación r como una métrica de evaluación de la relación entre las variables implicadas en el modelo, lo cual significa que es necesario explorar otro tipo de posibles relaciones entre dichas variables, particularmente aquellas de naturaleza no lineal para utilizar otro tipo de métricas de evaluación del

modelo y ver si mejora el desempeño del mismo para predecir la distancia en función de la velocidad.

Análisis de la validez del nuevo modelo

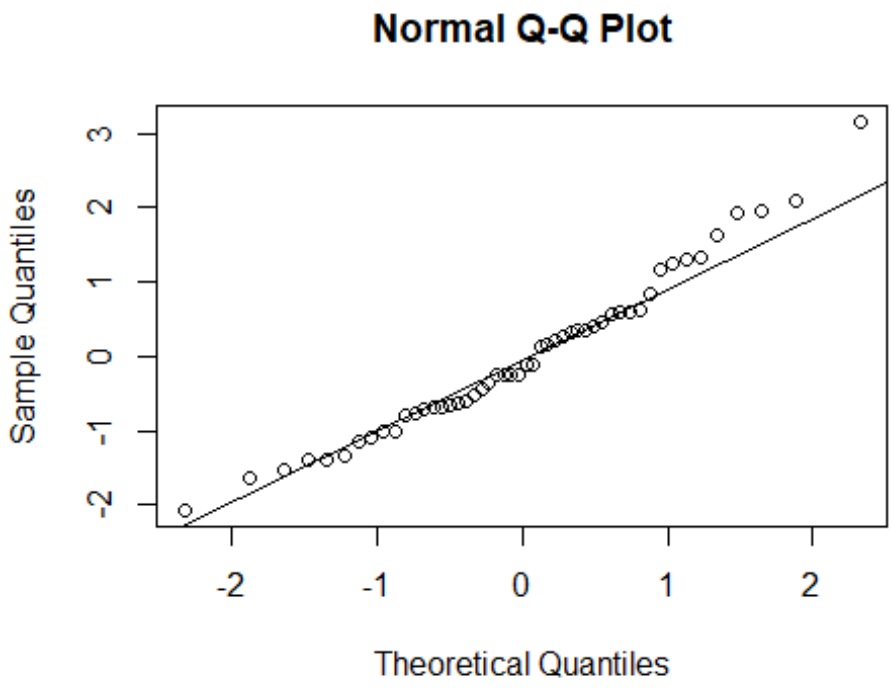
Normalidad de los residuos

Prueba de hipótesis:

H_0: los residuos del modelo siguen una distribución normal.

H_1: los residuos del modelo no siguen una distribución normal.

```
# Realizar el QQ-plot de los residuos del nuevo modelo obtenido

qqnorm(reg_tr$residuals)

qqline(reg_tr$residuals)
```




```
# Realizar test de normalidad de Jarque Bera para los residuos del nuevo
modelo
# de distancia normalizada vs velocidad

jarque.test(reg_tr$residuals)

##
##  Jarque-Bera Normality Test
##
```

```
## data:  reg_tr$residuals
## JB = 2.862, p-value = 0.2391
## alternative hypothesis: greater
```

En cuanto a la normalidad de los residuos del modelo, se observa que en el test de Jarque Bera de los residuos del modelo, el valor p de la prueba es igual a 0.2391, lo cual es considerablemente mayor a 0.05, por lo que se tiene evidencia estadística para no rechazar H_0, motivo por el cual, es posible concluir que los residuos del modelo siguen una distribución normal. Además de que también en el qq-plot del modelo, se aprecia que los percentiles graficados se ubican en su mayoría sobre la recta de normalidad, sin embargo hay otros percentiles que por el contrario, se desvían o alejan de la recta de normalidad, lo cual básicamente es indicativo de que aquellos datos que se alejan de la recta de normalidad en el qq-plot, son candidatos a ser datos atípicos o influyentes en el modelo, los cuales ocasionan que la distribución de los datos presente un determinado grado de sesgo, ya sea positivo, o negativo, lo cual a su vez propicia que dicha distribución no sea completamente normal.

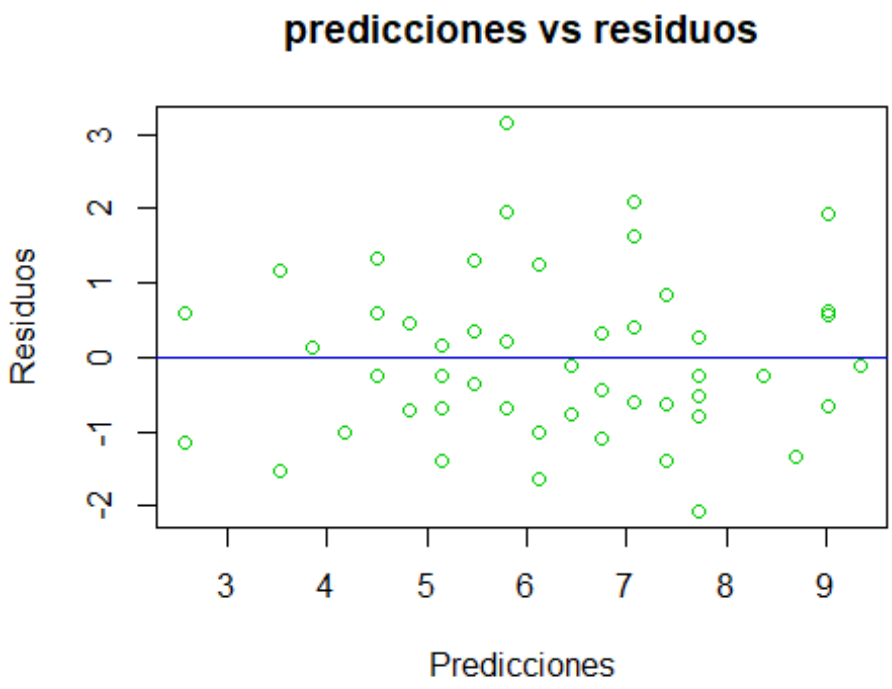
Homocedasticidad de residuos

Prueba de hipótesis:

H_0: los residuos del modelo tienen varianza constante (sí hay homocedasticidad).

H_1: los residuos del modelo no tienen varianza constante (no hay homocedasticidad)

```
# Realizar una gráfica de predicciones del modelo vs residuos para
comprobar que la
# varianza sea constante

plot(reg_tr$fitted.values, reg_tr$residuals, main = "predicciones vs
residuos",
     xlab = "Predicciones", ylab = "Residuos", col = "green3")

# Graficar la media cero de los residuos

abline(h = 0, col = "blue3")
```




```
# Realizar prueba de Breusch-Pagan para verificar homocedasticidad de
residuos

bptest(reg_tr)

##
##  studentized Breusch-Pagan test
##
## data:  reg_tr
## BP = 0.011192, df = 1, p-value = 0.9157
```

En el test de Breusch-Pagan se aprecia que el p valor del test es igual a 0.9157, lo cual al ser mucho mayor que 0.05, se tiene suficiente evidencia estadística para no rechazar H_0, motivo por el cual se afirma que los residuos del nuevo modelo tienen varianza constante, es decir, que los residuos presentan homocedasticidad. Además, en el gráfico de predicciones del modelo vs residuos, se observa que los residuos graficados (puntos del gráfico) no repsentan ninguna tendencia o patrón específico conforme varían las predicciones, por lo que se puede afirmar que los residuos del modelo poseen una varianza constante como se había visto en el test de homocedasticidad, además dado que los residuos no siguen ningún patrón o tendencia particular, la hipótesis nula H_0 no se rechazó, concluyendo así que la homocedasticidad se encuentra presente entre los residuos del modelo.

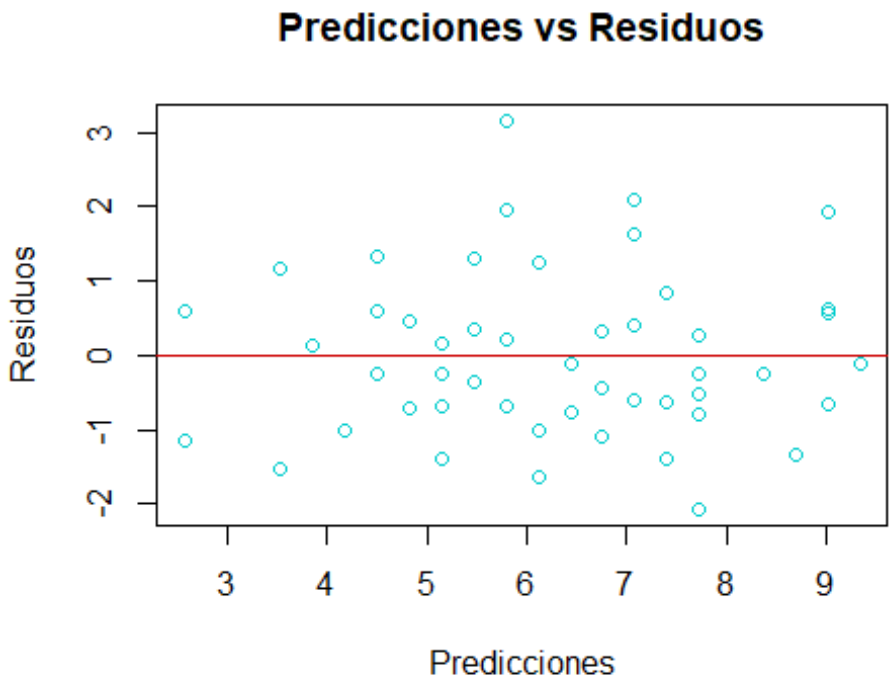
Independencia de residuos

Prueba de hipótesis:

H_0: los residuos del modelo no presentan autocorrelación (sí son independientes).

H_1: los residuos del modelo sí presentan autocorrelación (no son independientes).

```
# Graficar nuevamente predicciones del modelo vs residuos del mismo

plot(reg_tr$fitted.values, reg_tr$residuals, col = "cyan3", xlab =
"Predicciones",
     ylab = "Residuos", main = "Predicciones vs Residuos")

# Media cero de los residuos

abline(h = 0, col = "red3")
```

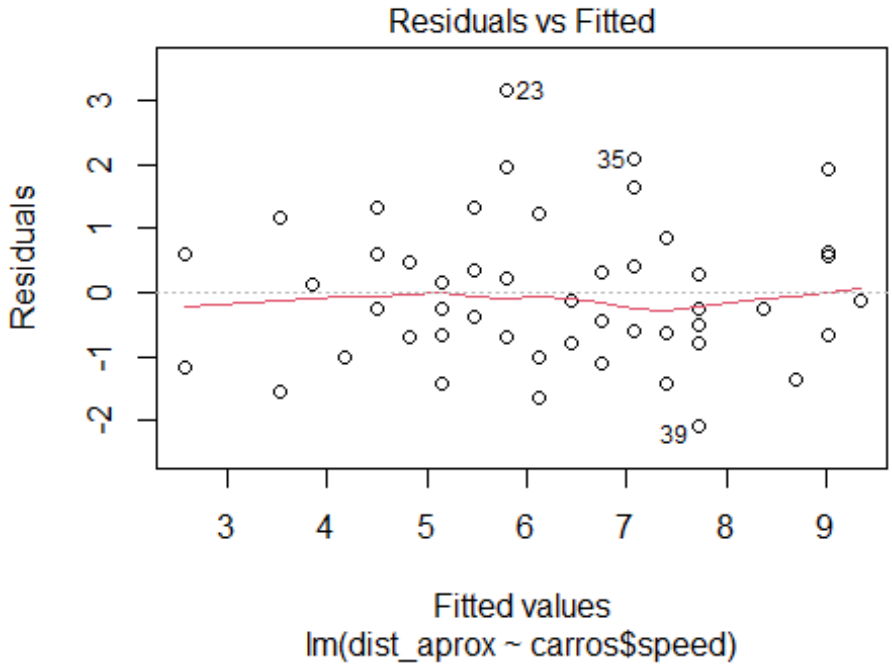


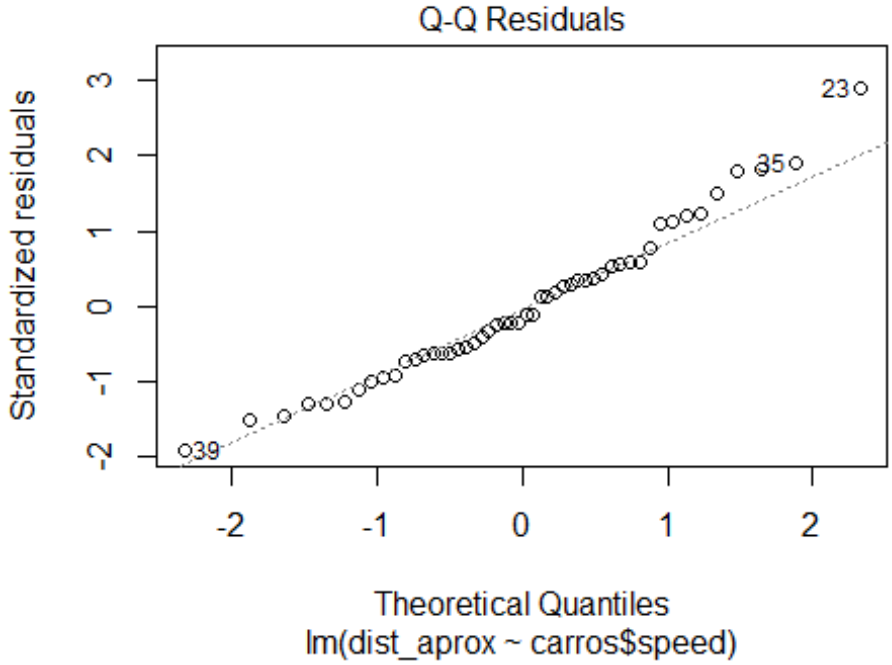

```
# Realizar test de independencia de Durbin Watson para los residuos del
modelo

dwtest(reg_tr)

##
##  Durbin-Watson test
##
## data:  reg_tr
## DW = 1.9417, p-value = 0.3609
## alternative hypothesis: true autocorrelation is greater than 0
```

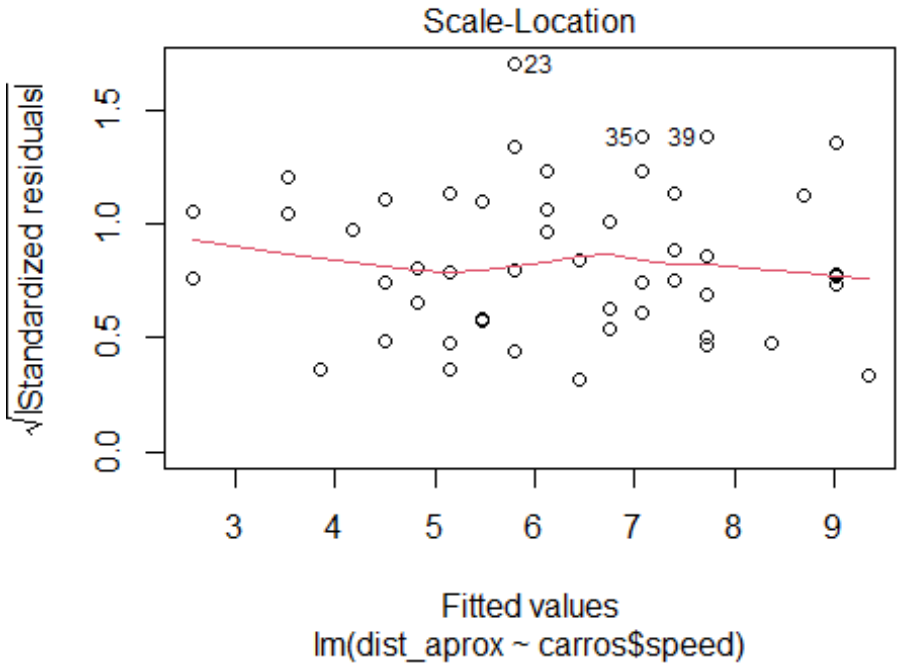
En el test de Durbin Watson para independencia realizado anteriormente, se observa que el p valor del test es igual a 0.3609, lo cual al ser mucho mayor que 0.05, se tiene suficiente evidencia estadística para no rechazar H_0, motivo por el cual, es posible concluir que los residuos del modelo no presentan autocorrelación, es decir que sí son independientes entre sí. Además, en el gráfico de predicciones vs residuos se aprecia que los residuos no siguen tendencias o patrones en particular, lo cual indica que la dispersión de los mismos se atribuye prácticamente al azar y no a algún otro factor subyacente en los datos, por lo cual, debido a ese motivo, no se rechazó H_0 en la prueba de hipótesis, llevando a concluir que los residuos del modelo son en efecto independientes.

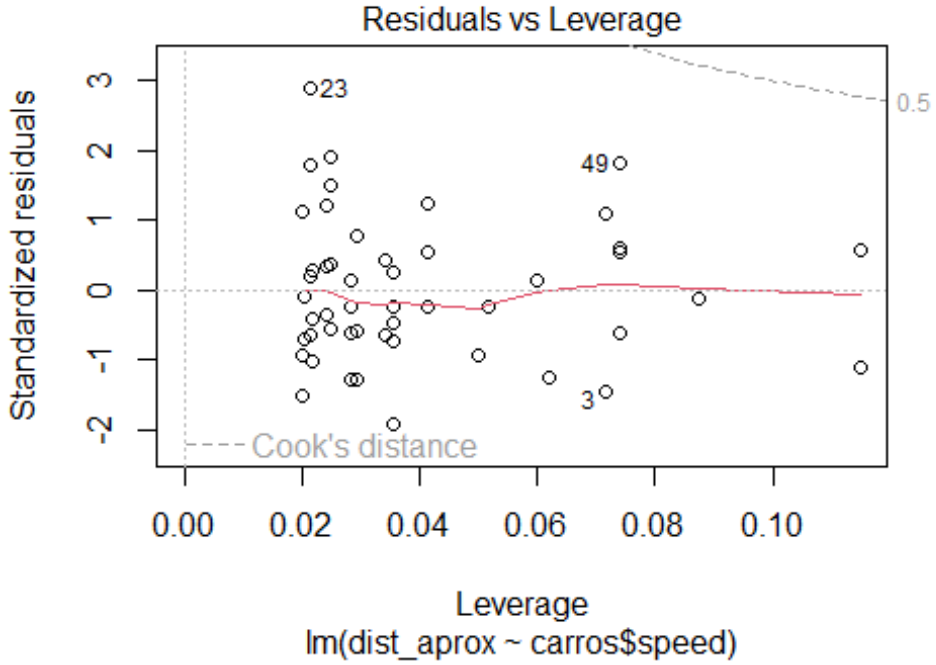
Gráficos del modelo general

```
# Realizar la serie de gráficos correspondiente al modelo en general con
plot(modelo)

plot(reg_tr)
```


Residuals vs Fitted
Residuals
Fitted values
lm(dist_aprox ~ carros$speed)
23
35
39


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(dist_aprox ~ carros$speed)
23
35
39





Despejar la distancia del modelo lineal entre la transformación y la velocidad

Siendo el modelo lineal entre la transformación y la velocidad:

$$\sqrt{D} = 1.2771 + 0.3224V$$

Donde $\sqrt{D}$ es la distancia normalizada mediante la transformación aproximada de Box Cox y V es la velocidad, ahora se procederá a despejar la distancia D del modelo anterior para obtener el modelo no lineal que relaciona la distancia con la velocidad directamente:

$$\left(\sqrt{D}\right)^2 = (1.2771 + 0.3224V)^2$$

$$D = (1.2771)^2 + 2(1.2771)(0.3224V) + (0.3224V)^2$$

Ahora, al resolver las operaciones previas obtenidas a partir del desarrollo del binomio cuadrado, el modelo no lineal resultante que relaciona la distancia con la velocidad es:

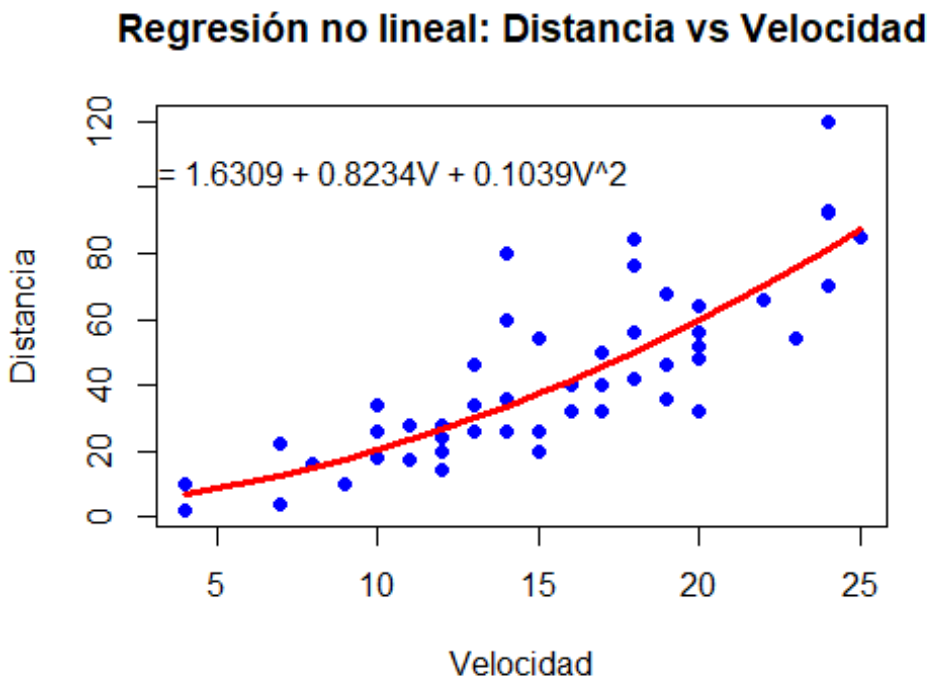
$$D = 1.6309 + 0.8234V + 0.1039V^2$$

```
# Definir una función para obtener la distancia correspondiente por cada
valor de velocidad
# empleando el modelo no lineal encontrado previamente

# x: velocidad

no.lineal = function(V){

  # Calcular distancia en función de la velocidad usando el modelo no
lineal

  # Unname se usa para quitar las etiquetas de nombres de los resultados

  unname(reg_tr$coefficients[1] ^ 2 + 2 * (reg_tr$coefficients[1]) *
    (reg_tr$coefficients[2] * V) + (reg_tr$coefficients[2] * V) ^ 2)

}
```

Gráfica de los datos y el modelo no lineal de distancia en función de velocidad

```
# Graficar los datos de distancia original vs velocidad original

plot(carros$speed, carros$dist, xlab = "Velocidad", ylab = "Distancia",
     main = "Regresión no lineal: Distancia vs Velocidad", col = "blue",
pch = 19)

# Graficar la curva correspondiente al modelo no lineal

lines(carros$speed, no.lineal(carros$speed), col = "red", lwd = 3)

# Colocar la ecuación del modelo no lineal en el gráfico
# D: distancia, V: velocidad
```

```
text(10, 105, "D = 1.6309 + 0.8234V + 0.1039V^2")
```



Adicionalmente, cabe mencionar que en cuanto a la idoneidad del modelo anterior (modelo lineal que relaciona la transformación de distancia con velocidad), es posible concluir que dicho modelo presenta un p valor de 1.773e-14, lo cual es mucho menor que 0.05, por lo que H_0 se rechazó, lo cual significa que dicho modelo sí es estadísticamente significativo, además de que también, ambos de sus coeficientes β poseen p valores menores a 0.05, específicamente 0.0113 para β_0 y 1.77e-14 para β_1 por lo que ambos resultan ser estadísticamente significativos, por lo cual es posible afirmar que el modelo lineal que relaciona la tranformación de distancia con la velocidad, o en otras palabras, el modelo no lineal de distancia en función de velocidad cumple con ser estadísticamente significativo y además es un modelo económico, puesto que se predicen los valores de la distancia en función de 1 sola variable independiente (mínimo posible de variables predictoras), además de que todos los coeficientes del modelo son estadísticamente significativos. Por otro lado, en cuanto a la validez del modelo no lineal, es posible afirmar que su modelo lineal asociado (el que relaciona la transformación de la distancia con la velocidad), sí cumple con el supuesto de normalidad de sus residuos, además de que también satisface los supuestos de homocedasticidad e independencia de una regresión lineal, por lo que debido a todo lo anterior, se concluye que el modelo no lineal generado sí resulta ser válido y significativo en el contexto estadístico para predecir la distancia recorrida por los automóvles en función de su velocidad.

Parte 4: Conclusión

Define cuál de los dos modelos analizados (Punto 1 o Punto 2) es el mejor modelo para describir la relación entre la distancia y la velocidad.

En conclusión, el mejor modelo para describir la relación existente entre la distancia y la velocidad es el modelo del punto 2 (modelo no lineal), principalmente debido a que dicho modelo sí cumple con el supuesto de normalidad mientras que el modelo lineal generado inicialmente no lo cumple, por lo cual a pesar de que ambos modelos cumplen con casi todos los supuestos, el modelo no lineal sí cumple normalidad mientras que el lineal no, por lo que las predicciones del modelo no lineal resultarán más confiables que las del modelo lineal, ya que el modelo no lineal satisface todos los supuestos y el lineal no, además de que también por otro lado, el modelo no lineal también exhibe una mejoría en cuanto al porcentaje de variabilidad de los datos explicado por el modelo (R^2), puesto que dicho modelo es capaz de explicar el 70.94% de la variabilidad total de los datos, mientras que el modelo lineal es capaz de explicar el 65.11% de la misma, por lo cual el modelo no lineal es capaz de explicar el mayor porcentaje de variabilidad de los datos, además, el modelo no lineal también tiene un p valor más pequeño que el modelo lineal, indicando que éste mismo (modelo no lineal) es más significativo o útil estadísticamente hablando para explicar la relación entre distancia y velocidad que el modelo lineal, aunado al hecho de que también en el modelo no lineal se aprecia que la variable de velocidad también posee un valor p más pequeño que en el modelo lineal, lo cual señala que en el modelo no lineal, la velocidad cobra más relevancia y por tanto aporta más información al modelo para realizar las predicciones de distancia a partir de velocidad, contribuyendo así a que el modelo no lineal sea estadísticamente más significativo que el modelo lineal, por lo que considerando todo lo anterior, se concluye que el mejor modelo para representar la relación entre distancia y velocidad es el modelo no lineal.

Comenta sobre posibles problemas del modelo elegido (datos atípicos, alejamiento de los supuestos, dificultad de cálculo o interpretación)

En términos generales, el modelo no lineal elegido presenta ciertos inconvenientes, los cuales radican principalmente en el hecho de que en el gráfico de los datos y del modelo no lineal se aprecia que existe presencia de ciertos datos atípicos, representados por aquellos puntos en el gráfico que se alejan considerablemente de la curva del modelo, de los cuales algunos pueden ser datos influyentes que probablemente estén ocasionando que la curva del modelo no lineal se incline más hacia un lado del gráfico en particular para buscar ajustarse a dichos datos influyentes, lo cual a su vez puede propiciar que la distribución de los datos involucrados en el modelo no lineal aún tenga cierto grado tanto de sesgo como de curtosis, lo cual estaría impidiendo que la distribución de los datos en cuestión sea completamente normal, no obstante, otro de los posibles inconvenientes del modelo no lineal escogido consiste en que será ligeramente más costoso en el sentido computacional, calcular las predicciones de distancia a partir del mismo, dado que conforme incrementa el grado de las funciones matemáticas, el modelo se torna más

complejo, lo cual a su vez implica que a los programas computacionales les tomará más tiempo calcular las predicciones deseadas, ya que el calcular funciones mayormente complejas tales como potencias superiores a 2, logaritmos, funciones exponenciales, entre otras, requiere de un mayor poder de cómputo y de una mayor capacidad de memoria computacional, lo cual podrían no tener los ordenadores básicos, ocasionando así diversos inconvenientes en el proceso de cálculo de las predicciones al usar dichos ordenadores con una capacidad de memoría y poder de cómputo reducidos.