

A1- Regresión no lineal

Eryk Elizondo González A01284899

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

1. Accede a los datos de cars en R (data = cars)

```
Data <- cars
Speed <- Data$speed
Dist <- Data$dist
```

1. Prueba normalidad univariada de la velocidad y distancia (prueba con dos de las pruebas vistas en clase)

Hipótesis

- H_0 : Los datos provienen de una población normal
- H_1 : Los datos no provienen de una población normal

Valor frontera

$$\alpha = 0.03$$

Regla de decisión

- Se rechaza H_0 si valor $p < \alpha$

```
ad.test(Speed)
```

```
##
## Anderson-Darling normality test
##
## data: Speed
## A = 0.26143, p-value = 0.6927
```

```
jarque.test(Speed)
```

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

```

ad.test(Dist)

##
## Anderson-Darling normality test
##
## data: Dist
## A = 0.74067, p-value = 0.05021

jarque.test(Dist)

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

```

Análisis

No se rechaza la hipótesis nula (H_0). Esto indica que los datos de velocidad y distancia provienen de una población normal.

2. Realiza gráficos que te ayuden a identificar posibles alejamientos de normalidad:

Velocidad

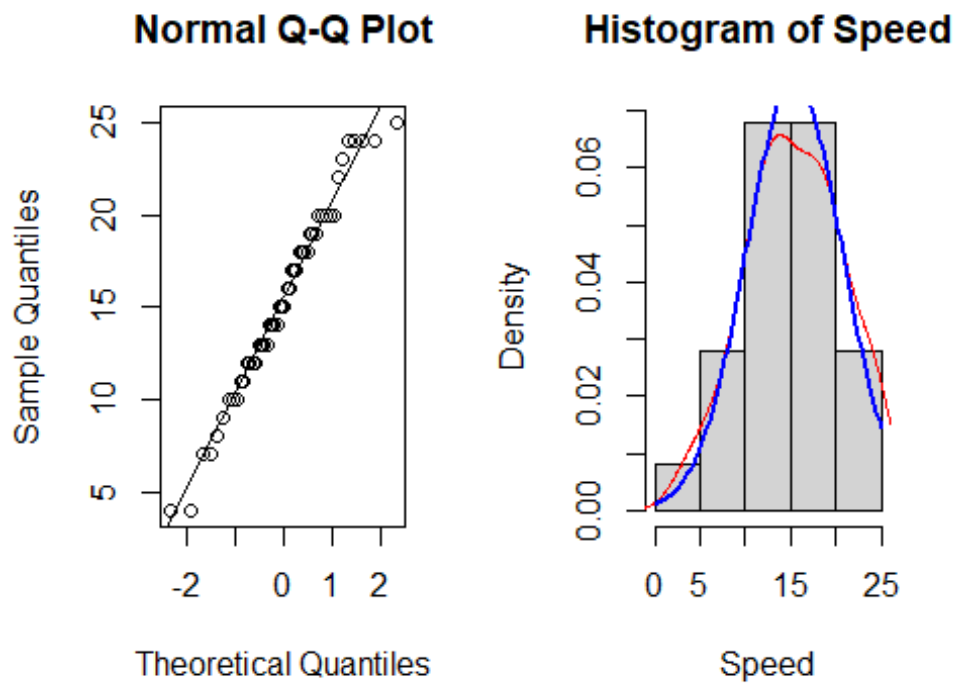
```

par(mfrow=c(1,2))

qqnorm(Speed)
qqline(Speed)

hist(Speed, freq=FALSE)
lines(density(Speed), col="red")
curve(dnorm(x, mean=mean(Speed), sd=sd(Speed)), add=TRUE, col="blue", lwd=2)

```



Análisis

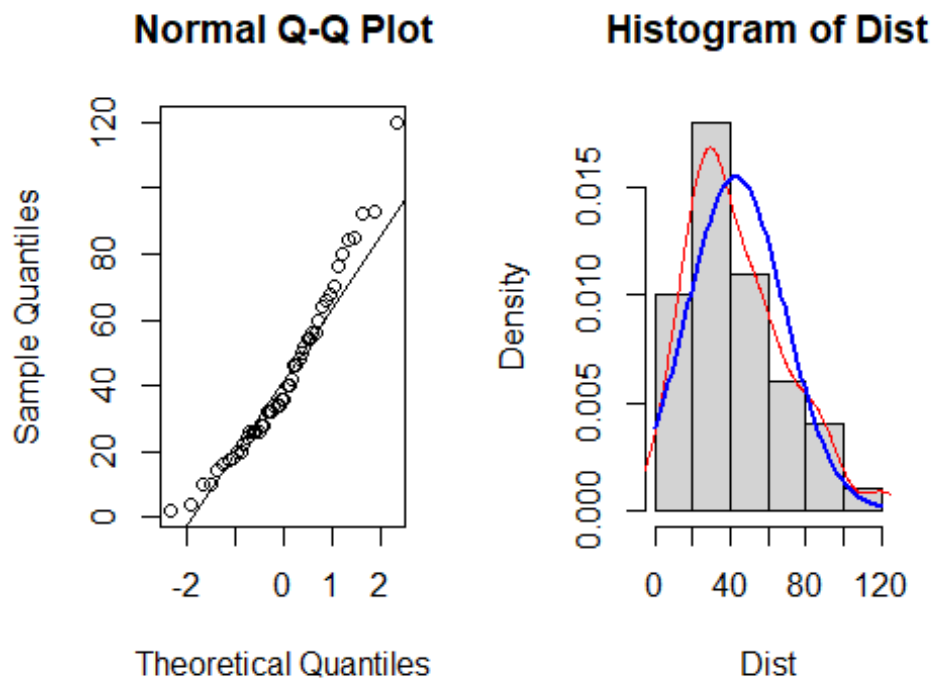
El qqplot de la velocidad muestra un comportamiento normal y no se observa ningún sesgo aparente. El histograma muestra simetría en su distribución.

Distancia

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

qqnorm(Dist)
qqline(Dist)

hist(Dist, freq=FALSE)
lines(density(Dist), col="red")
curve(dnorm(x, mean=mean(Dist), sd=sd(Dist)), add=TRUE, col="blue", lwd=2)
```



Análisis

El qqplot así como el histograma de la distancia muestra un sesgo a la derecha.

3. Calcula el coeficiente de sesgo y el coeficiente de curtosis (sugerencia: usar la librería e1071, usar: `skeness` y `kurtosis`) para cada variable.

```
cat("El coeficiente de sesgo de la velocidad es de", skewness(Speed), "y el  
coeficiente de curtosis es", kurtosis(Speed), "\n")
```

```
## El coeficiente de sesgo de la velocidad es de -0.1105533 y el coeficiente  
de curtosis es -0.6730924
```

```
cat("El coeficiente de sesgo de la distancia es de", skewness(Dist), "y el  
coeficiente de curtosis es", kurtosis(Dist))
```

```
## El coeficiente de sesgo de la distancia es de 0.7591268 y el coeficiente  
de curtosis es 0.1193971
```

2. Emite una conclusión final sobre la normalidad de los datos. Argumenta basándote en todos los análisis realizados en esta parte. Incluye posibles motivos de alejamiento de normalidad.

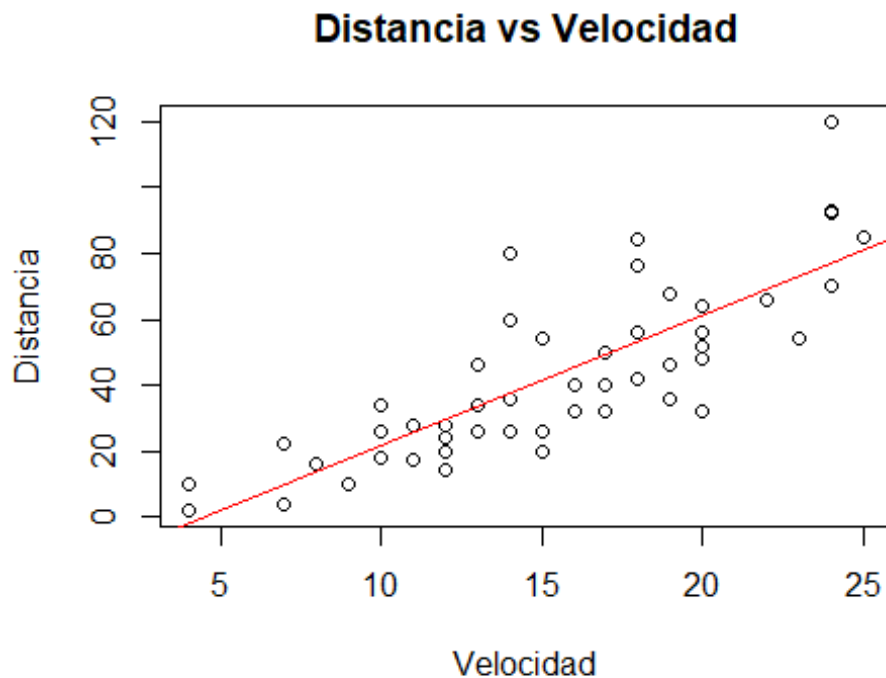
Se observa que, aunque ambas variables se comportan como una normal, la velocidad posee un acercamiento a la normal mayor a la distancia, evidente debido a los diferentes valores de p y el sesgo en sus histogramas. Ya que la distancia es la variable dependiente, tiene un rango limitado de valores posibles cercanos al cero debido a que se busca

minimizar la distancia de frenado, mientras que la variable independiente de la velocidad no tiene esta limitación y permite tener suficientes observaciones para comportarse como una normal más simétrica.

Parte 2: Regresión lineal

1. Prueba regresión lineal simple entre distancia y velocidad. Usa $\text{lm}(y \sim x)$.

```
M1 <- lm(Dist ~ Speed)
plot(Speed, Dist, main = "Distancia vs Velocidad", ylab = "Distancia", xlab =
"Velocidad")
abline(M1, col = "red", lw = 1.5)
```



$\text{Distancia} = -17.5791 + 3.9324 * \text{Velocidad}$

2. Analiza significancia del modelo: individual, conjunta y coeficiente de determinación. Usa $\text{summary}(\text{Modelo})$

```
summary(M1)

##
## Call:
## lm(formula = Dist ~ Speed)
##
## 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
```

1. Modelo

Hipótesis

- H_0 : El modelo no explica significativamente la variabilidad en las distancias ($\beta_1 = \beta_2 = 0$)
- H_1 : Al menos uno de los coeficientes de los predictores es significativamente diferente de cero.

Valor frontera

$$\alpha = 0.03$$

Regla de decisión

- Se rechaza H_0 si valor $p < \alpha$

Conclusión

El valor p del F-estadístico para el modelo completo es 1.49e-12 lo cual es menor que el valor frontera. Rechazamos la hipótesis nula (H_0) a favor de la alternativa (H_1). Esto significa que el modelo es significativamente diferente de un modelo sin predictores y que al menos uno de los predictores tiene un efecto significativo sobre la distancia.

2. Coeficientes

Hipótesis 1

- $H_0: \beta_0 = 0$
- $H_1: \beta_0 \neq 0$

Hipótesis 2

- $H_0: \beta_1 = 0$
- $H_1: \beta_1 \neq 0$

Valor frontera

$$\alpha = 0.03$$

Regla de decisión

- Se rechaza H_0 si valor $p < \alpha$

Conclusión

Se rechaza la hipótesis nula (H_0) de las hipótesis 1 y 2 ya que son significativamente diferentes de cero.

3. Coeficiente de determinación

El porcentaje de variación explicada por el modelo es de 65.11%

3. Analiza validez del modelo.

1. Normalidad de los residuos

Hipótesis

- H_0 : Los datos provienen de una población normal
- H_1 : Los datos no provienen de una población normal

Valor frontera

$$\alpha = 0.03$$

Regla de decisión

- Se rechaza H_0 si valor $p < \alpha$

```
ad.test(M1$residuals)
```

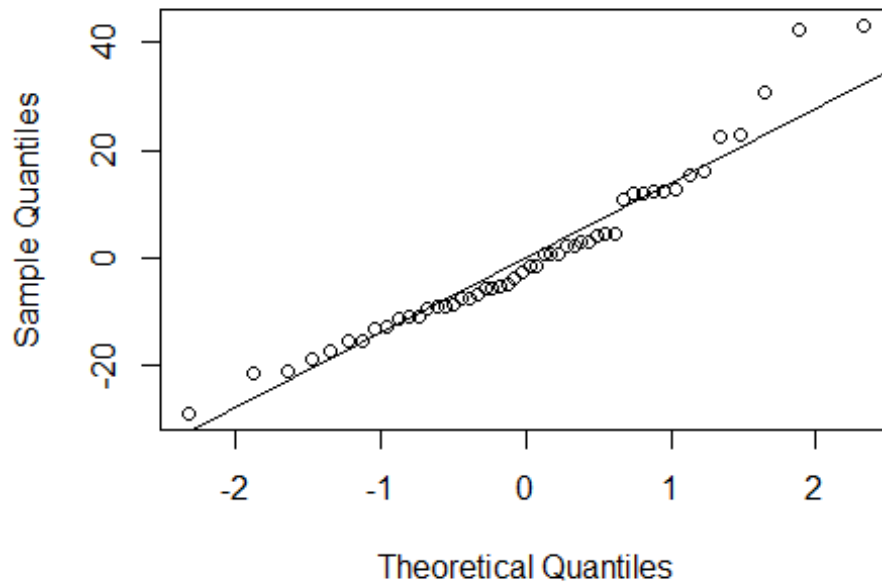
```
##  
## Anderson-Darling normality test  
##  
## data:  M1$residuals  
## A = 0.79406, p-value = 0.0369
```

```
# Gráficas
```

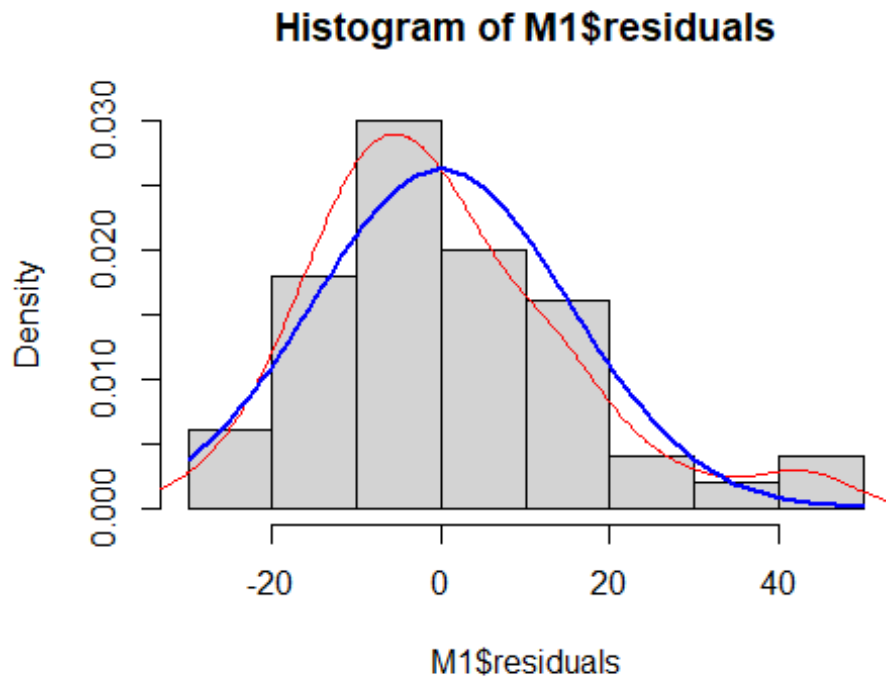
```
qqnorm(M1$residuals)
```

```
qqline(M1$residuals)
```

Normal Q-Q Plot



```
# Histograma  
hist(M1$residuals, freq=FALSE)  
lines(density(M1$residual), col="red")  
curve(dnorm(x, mean=mean(M1$residuals), sd=sd(M1$residuals)), add=TRUE,  
col="blue", lwd=2)
```

Conclusión

No se rechaza la hipótesis nula (H_0). Esto indica que no hay evidencia significativa para afirmar que los datos no provienen de una población normal.

2. Verificación de media cero

Hipótesis

- $H_0: \mu_e = 0$
- $H_1: \mu_e \neq 0$

Valor frontera

$$\alpha = 0.03$$

Regla de decisión

- Se rechaza H_0 si valor $p < \alpha$

```
t.test(M1$residuals, mu = 0)
```

```
##
## One Sample t-test
##
## data: M1$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
```

Conclusión

No se rechaza la hipótesis nula (H_0). Esto indica que no hay evidencia significativa para afirmar que la media de los residuos es diferente de cero.

3. Homocedasticidad

Hipótesis

- H_0 : La varianza de los errores es constante (homocedasticidad)
- H_1 : La varianza de los errores no es constante (heterocedasticidad)

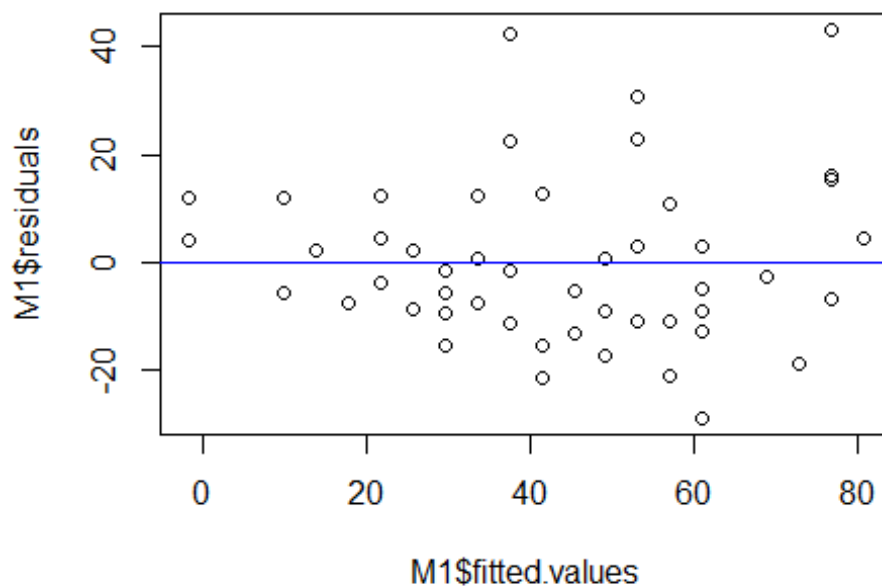
Valor frontera

$$\alpha = 0.03$$

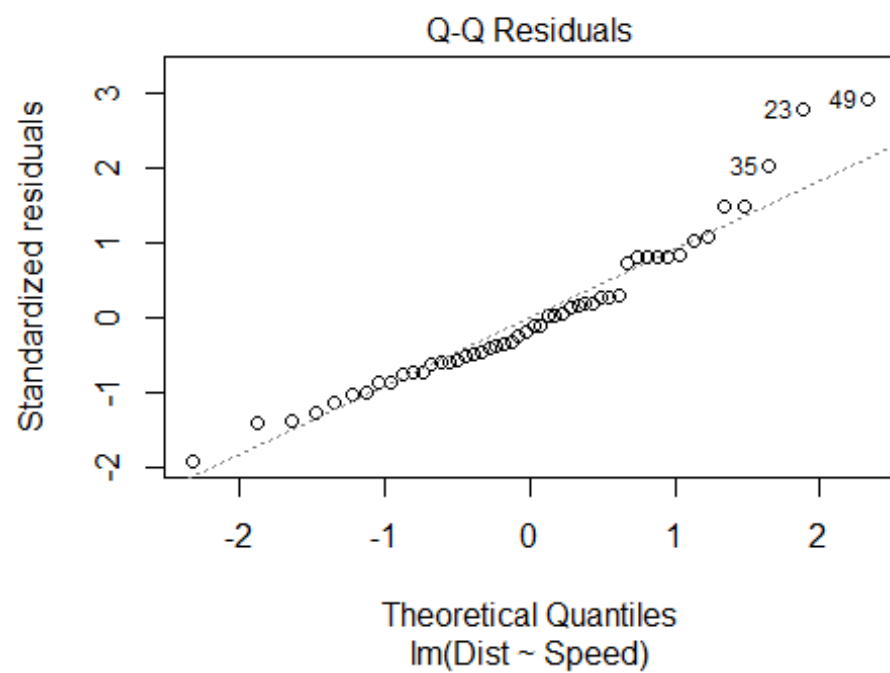
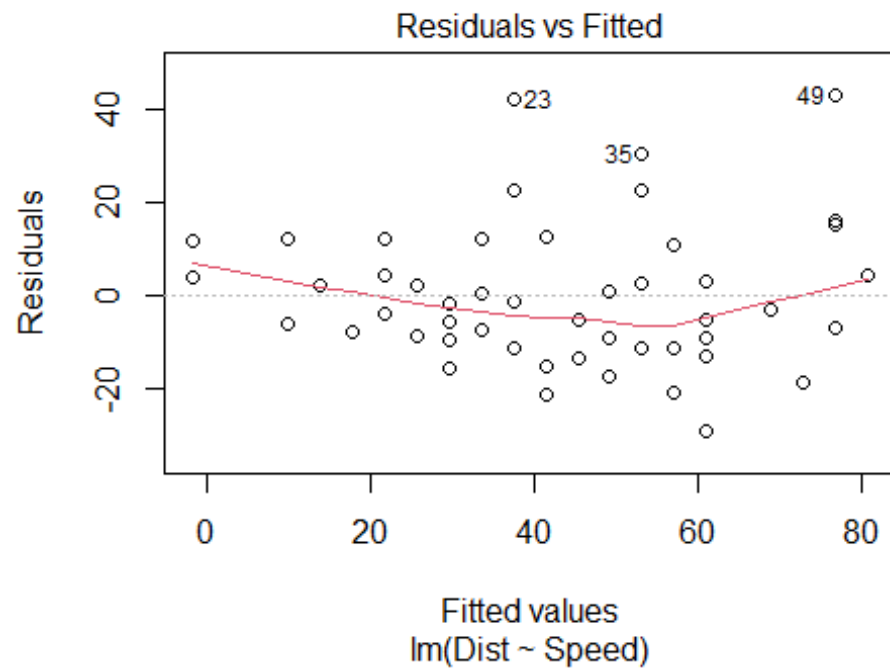
Regla de decisión

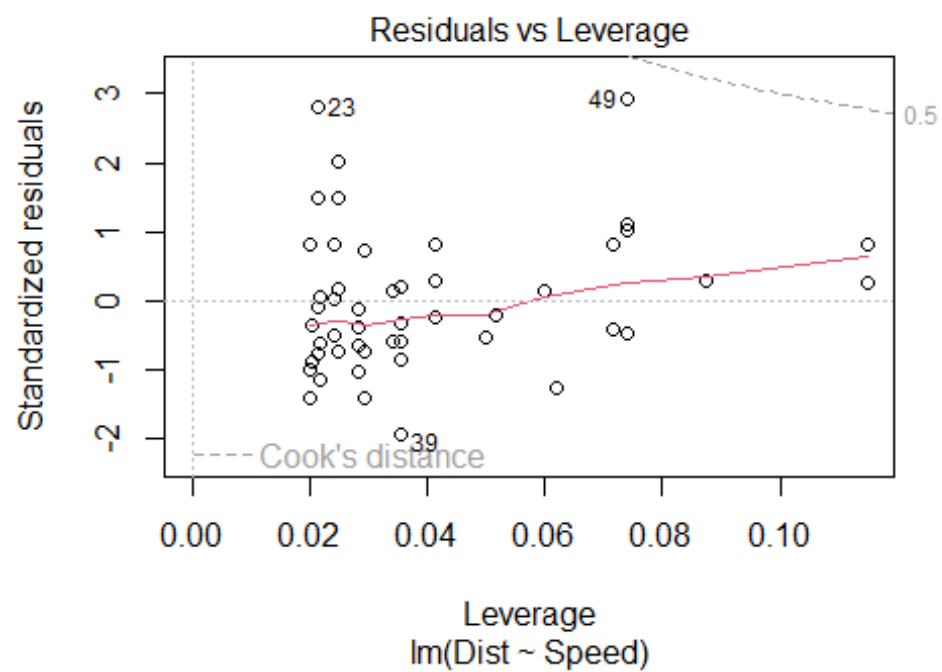
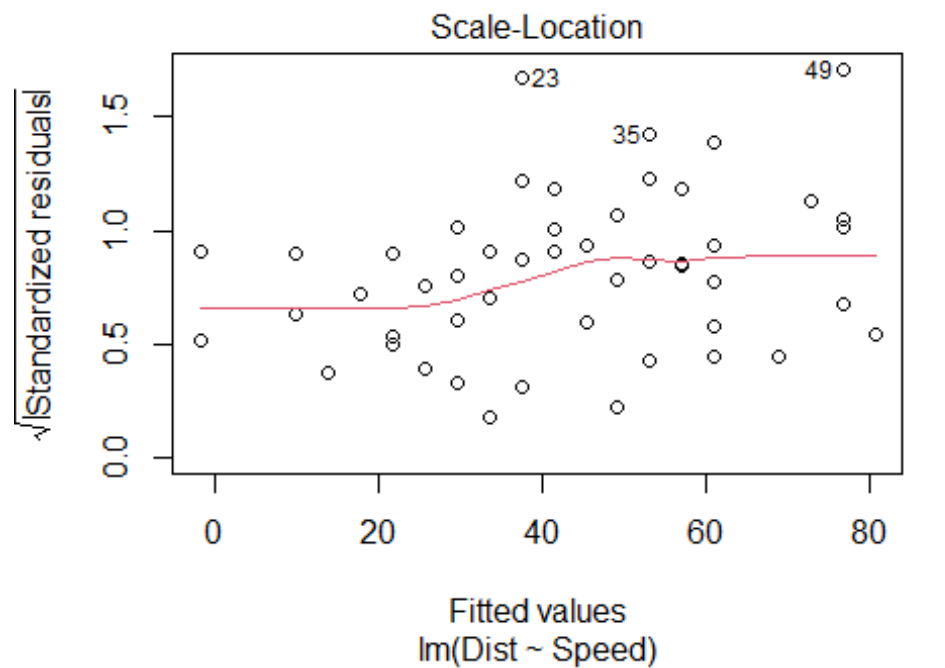
- Se rechaza H_0 si valor $p < \alpha$

```
plot(M1$fitted.values, M1$residuals)
abline(h=0, col = 'blue')
```



```
plot(M1)
```





```
bptest(M1)
```

```
##
```

```
## studentized Breusch-Pagan test
```

```
##  
## data: M1  
## BP = 3.2149, df = 1, p-value = 0.07297
```

Conclusión

No se rechaza la hipótesis nula (H_0). Esto indica que no hay evidencia significativa para afirmar que la varianza de los errores no es constante (heterocedasticidad)

4. Linealidad

Hipótesis

- H_0 : No hay términos omitidos que indican linealidad
- H_1 : Hay una especificación errónea en el modelo que indica no linealidad

Valor frontera

$$\alpha = 0.04$$

Regla de decisión

- Se rechaza H_0 si valor $p < \alpha$

```
resettest(M1)  
##  
## RESET test  
##  
## data: M1  
## RESET = 1.5554, df1 = 2, df2 = 46, p-value = 0.222
```

Conclusión

No se rechaza la hipótesis nula (H_0). Esto indica que no hay evidencia significativa para afirmar que hay una especificación errónea en el modelo que indica no linealidad.

5. Independencia

Hipótesis

- H_0 : Los errores no están autocorrelacionados
- H_1 : Los errores están autocorrelacionados

Valor frontera

$$\alpha = 0.04$$

Regla de decisión

- Se rechaza H_0 si valor $p < \alpha$

```
bgtest(M1)
```

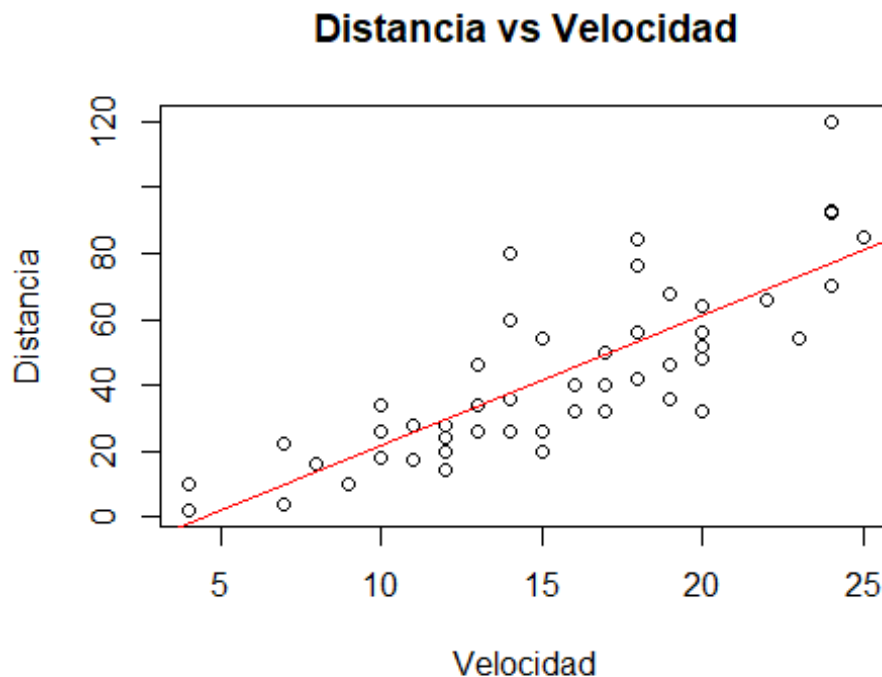
```
##  
## Breusch-Godfrey test for serial correlation of order up to 1  
##  
## data: M1  
## LM test = 1.2908, df = 1, p-value = 0.2559
```

Conclusión

No se rechaza la hipótesis nula (H_0). Esto indica que no hay evidencia significativa para afirmar que los errores están autocorrelacionados.

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

```
plot(Speed, Dist, main = "Distancia vs Velocidad", ylab = "Distancia", xlab =  
"Velocidad")  
abline(M1, col = "red", lw = 1.5)
```



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

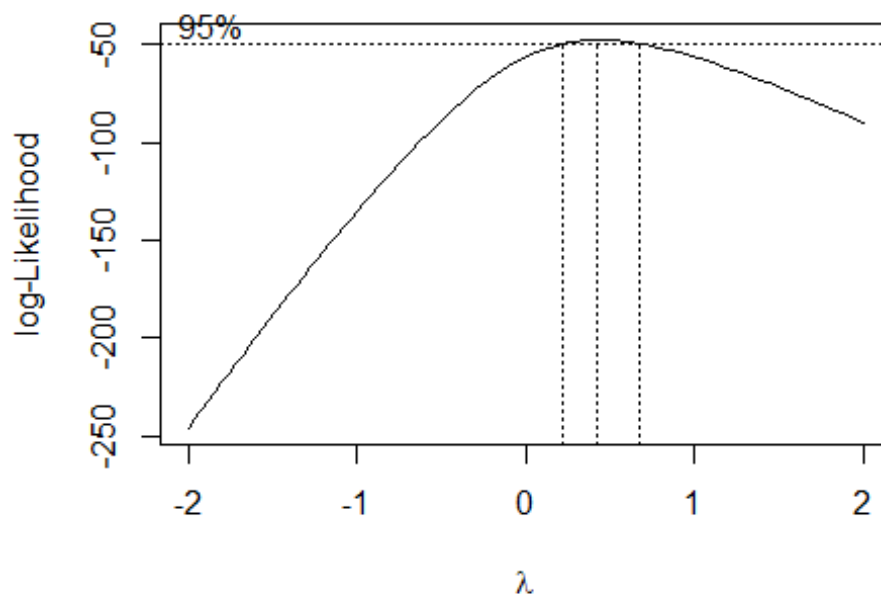
Considerando la significancia, el modelo y sus coeficientes son significativos y por ende apropiados para la predicción así como pasó todas las pruebas de validez, la única observación importante es el coeficiente de determinación que es relativamente bajo con un 0.6511.

Parte 3: Regresión no lineal

1. Con el objetivo de probar un modelo no lineal que explique la relación entre la distancia y la velocidad, haz una transformación con la base de datos car que te garantice normalidad en ambas variables (ojo: concéntrate solo en la variable que tiene más alejamiento de normalidad).

1. Encuentra el valor de λ en la transformación Box-Cox para el modelo lineal donde Y sea la distancia y X la velocidad.

```
bc <- boxcox(lm(Dist ~ Speed))
```



```
lambda <- bc$x[which.max(bc$y)]  
cat("Lambda de transformación Box-Cox:", lambda)  
## Lambda de transformación Box-Cox: 0.4242424
```

2. Define la transformación exacta y el aproximada de acuerdo con el valor de λ que encontraste en la transformación de Box y Cox. Escribe las ecuaciones de las dos transformaciones encontradas.

```
bce <- (((Dist)^lambda) - 1) / (lambda) # Box-Cox exacto  
bca <- sqrt(Dist) # Box-Cox aproximado sqrt(x)
```

3. Analiza la normalidad de las transformaciones obtenidas.

1. Prueba de normalidad con Anderson-Darling

```
ta0 <- ad.test(Dist) # Original
ta1 <- ad.test(bca) # Box-Cox aproximado
ta2 <- ad.test(bce) # Box-Cox exacto
```

2. Prueba de normalidad con Jarque Bera

```
tj0 <- jarque.test(Dist) # Original
tj1 <- jarque.test(bca) # Box-Cox aproximado
tj2 <- jarque.test(bce) # Box-Cox exacto
```

3. Resumen de medidas

```
m0 = round(c(as.numeric(summary(Dist)), kurtosis(Dist), skewness(Dist),
ta0$p.value, tj0$p.value), 3)
m1 = round(c(as.numeric(summary(bca)), kurtosis(bca), skewness(bca),
ta1$p.value, tj1$p.value), 3)
m2 = round(c(as.numeric(summary(bce)), kurtosis(bce), skewness(bce),
ta2$p.value, tj2$p.value), 3)
```

4. Tabla

```
m <- as.data.frame(rbind(m0, m1, m2))
row.names(m)=c("Original", "Box-Cot Aproximado", "Box-Cot Exacto")
names(m)=c("Minimo", "Q1", "Mediana", "Media", "Q3", "Máximo", "Curtosis", "Sesgo", "
Valor p (Anderson)", "Valor p (Jarque)")
```

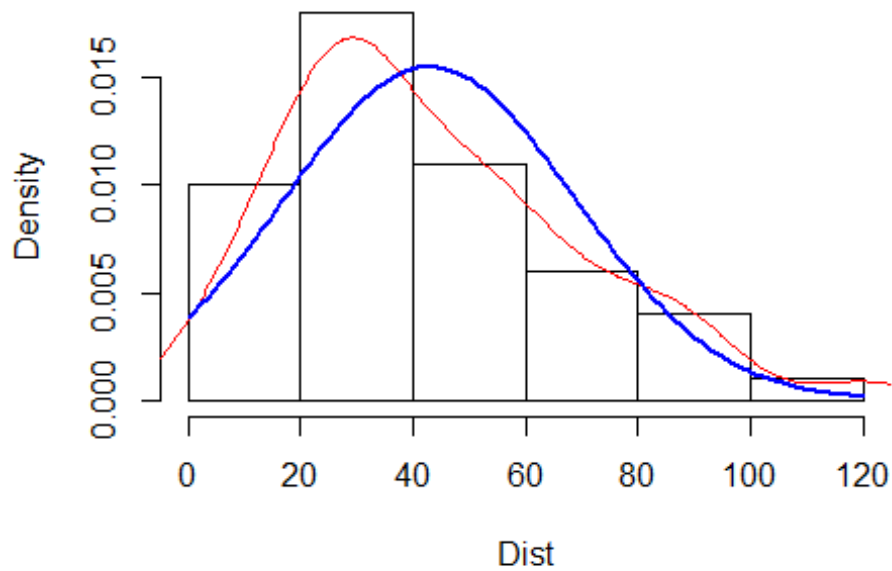
```
m
```

##	Minimo	Q1	Mediana	Media	Q3	Máximo	Curtosis	Sesgo
## Original	2.000	26.000	36.000	42.980	56.000	120.000	0.119	0.759
## Box-Cot Aproximado	1.414	5.099	6.000	6.242	7.483	10.954	-0.314	-0.019
## Box-Cot Exacto	0.806	7.033	8.423	8.712	10.646	15.609	-0.187	-0.170
##	Valor p (Anderson)				Valor p (Jarque)			
## Original			0.050			0.073		
## Box-Cot Aproximado			0.973			0.956		
## Box-Cot Exacto			0.972			0.875		

5. Histogramas

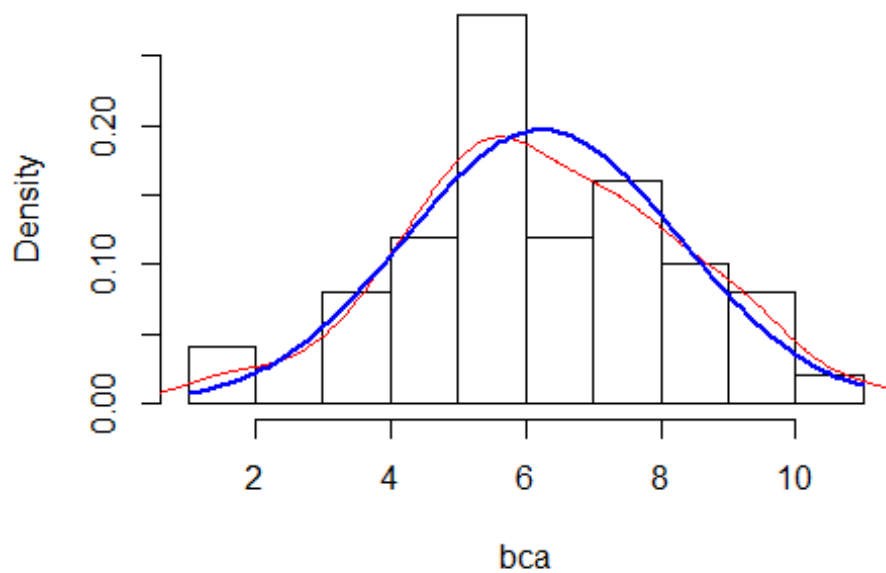
```
hist(Dist, prob = TRUE, col = 0, freq = FALSE, main = "Datos originales")
lines(density(Dist), col="red")
curve(dnorm(x, mean = mean(Dist), sd=sd(Dist)), add=TRUE, col="blue", lwd=2)
```


Datos originales

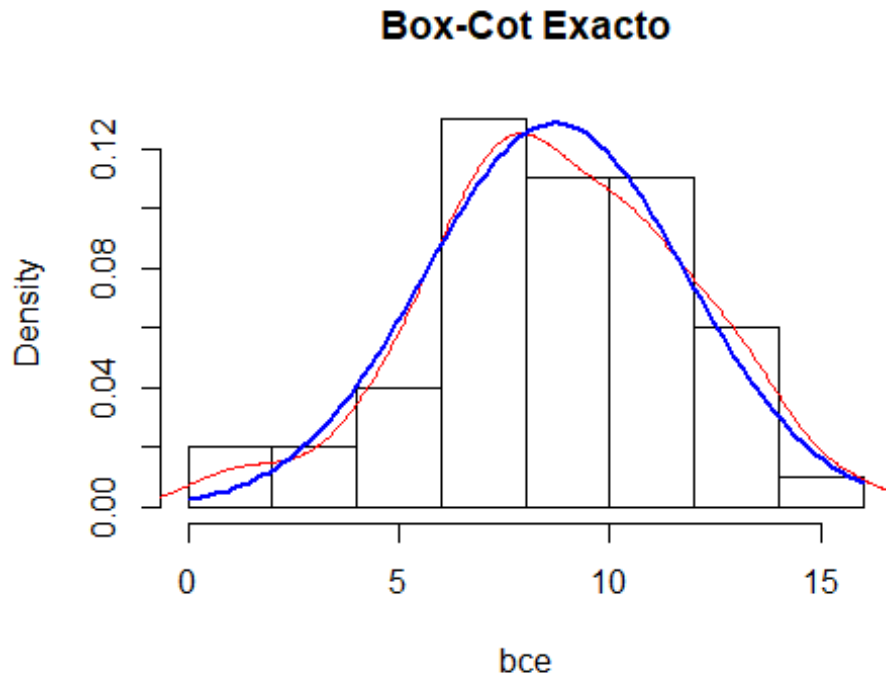


```
hist(bca, prob = TRUE, col = 0, freq = FALSE, main = "Box-Cot Aproximado")  
lines(density(bca), col="red")  
curve(dnorm(x, mean = mean(bca), sd=sd(bca)), add=TRUE, col="blue", lwd=2)
```

Box-Cot Aproximado



```
hist(bce, prob = TRUE, col = 0, freq = FALSE, main = "Box-Cot Exacto")
lines(density(bce), col="red")
curve(dnorm(x, mean = mean(bce), sd=sd(bce)), add=TRUE, col="blue", lwd=2)
```



6. Análisis

Hipótesis

- H_0 : Los datos provienen de una población normal
- H_1 : Los datos no provienen de una población normal

Valor frontera

$$\alpha = 0.03$$

Regla de decisión

- Se rechaza H_0 si valor $p < \alpha$

Análisis

No se rechaza la hipótesis nula (H_0) para ninguna de las transformaciones. Esto indica que los datos de distancia provienen de una población normal en los 3 casos.

3. Detecta anomalías y corrige tu base de datos transformado (datos atípicos, ceros anómalos, etc): solo en caso de no tener normalidad en las transformaciones. En caso de corrección de los datos por anomalías, vuelve a buscar la lambda para tus nuevos datos.

Dado que todos los datos originales y en sus 2 transformaciones son normales, no se corregirá la base de datos.

2. Concluye sobre las dos transformaciones realizadas: Define la mejor transformación de los datos de acuerdo a las características de las dos transformaciones encontradas (exacta o aproximada). Toman en cuenta la normalidad de los datos y la economía del modelo.

De acuerdo a las pruebas y gráficas, se puede observar que ambas transformaciones son extremadamente similares en terminos de mejora, sin embargo, considerando la economía del modelo, es mejor irse por la transformación aproximada ya que la ecuación es mucho más sencilla y no hay una diferencia significativa entre la mediana y media a comparación de la transformación exacta, o del sesgo o pruebas de normalidad.

3. Con la mejor transformación (punto 2), realiza la regresión lineal simple entre la mejor transformación (exacta o aproximada) y la variable velocidad:

1. Escribe el modelo lineal para la transformación.

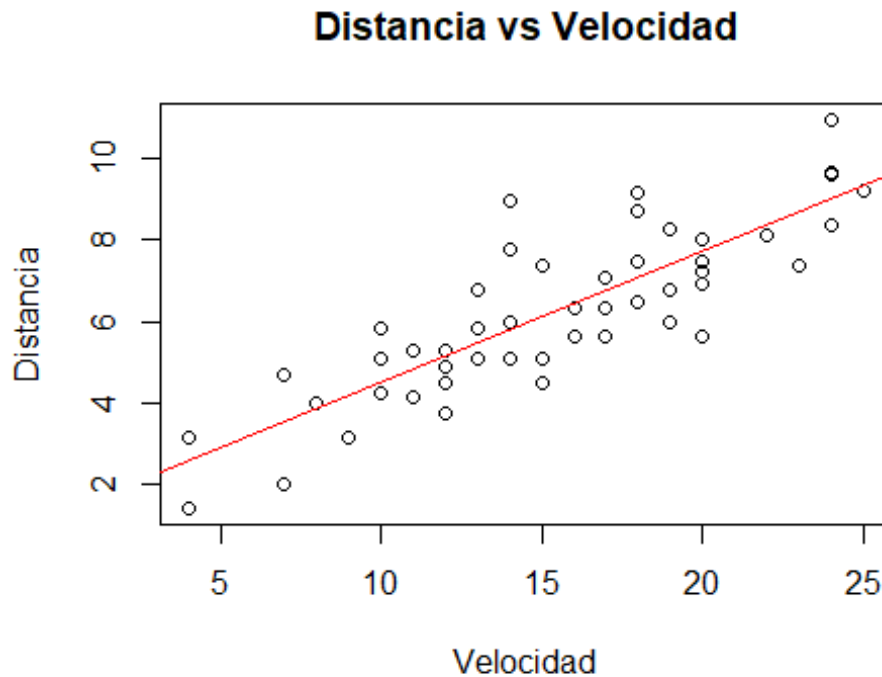
```
M2 <- lm(bca ~ Speed)
summary(M2)
```

```
##
## Call:
## lm(formula = bca ~ 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 *
## 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
```

$\text{Distancia(Transformada)} = 1.27705 + 0.32241 * \text{Velocidad}$

2. Grafica los datos y el modelo lineal (ecuación) de la transformación elegida vs velocidad.

```
plot(Speed, bca, main = "Distancia vs Velocidad", ylab = "Distancia", xlab =  
"Velocidad")  
abline(M2, col = "red", lw = 1.5)
```



3. Analiza significancia del modelo (individual, conjunta y coeficiente de correlación)

1. Modelo

Hipótesis

- H_0 : El modelo no explica significativamente la variabilidad en las distancias ($\beta_1 = \beta_2 = 0$)
- H_1 : Al menos uno de los coeficientes de los predictores es significativamente diferente de cero.

Valor frontera

$$\alpha = 0.03$$

Regla de decisión

- Se rechaza H_0 si valor $p < \alpha$

Conclusión

El valor p del F-estadístico para el modelo completo es 1.773e-14 lo cual es menor que el valor frontera. Rechazamos la hipótesis nula (H_0) a favor de la alternativa (H_1). Esto

significa que el modelo es significativamente diferente de un modelo sin predictores y que al menos uno de los predictores tiene un efecto significativo sobre la distancia.

2. Coeficientes

Hipótesis 1

- $H_0: \beta_0 = 0$
- $H_1: \beta_0 \neq 0$

Hipótesis 2

- $H_0: \beta_1 = 0$
- $H_1: \beta_1 \neq 0$

Valor frontera

$$\alpha = 0.03$$

Regla de decisión

- Se rechaza H_0 si valor $p < \alpha$

Conclusión

Se rechaza la hipótesis nula (H_0) de las hipótesis 1 y 2 ya que son significativamente diferentes de cero.

3. Coeficiente de determinación

El porcentaje de variación explicada por el modelo es de 70.94%

4. Analiza validez del modelo: normalidad de los residuos, homocedasticidad e independencia. Indica si hay candidatos a datos atípicos o influyentes en la regresión. Usa `plot(Modelo)` para los gráficos y añade pruebas de hipótesis.

1. Normalidad de los residuos

Hipótesis

- H_0 : Los datos provienen de una población normal
- H_1 : Los datos no provienen de una población normal

Valor frontera

$$\alpha = 0.03$$

Regla de decisión

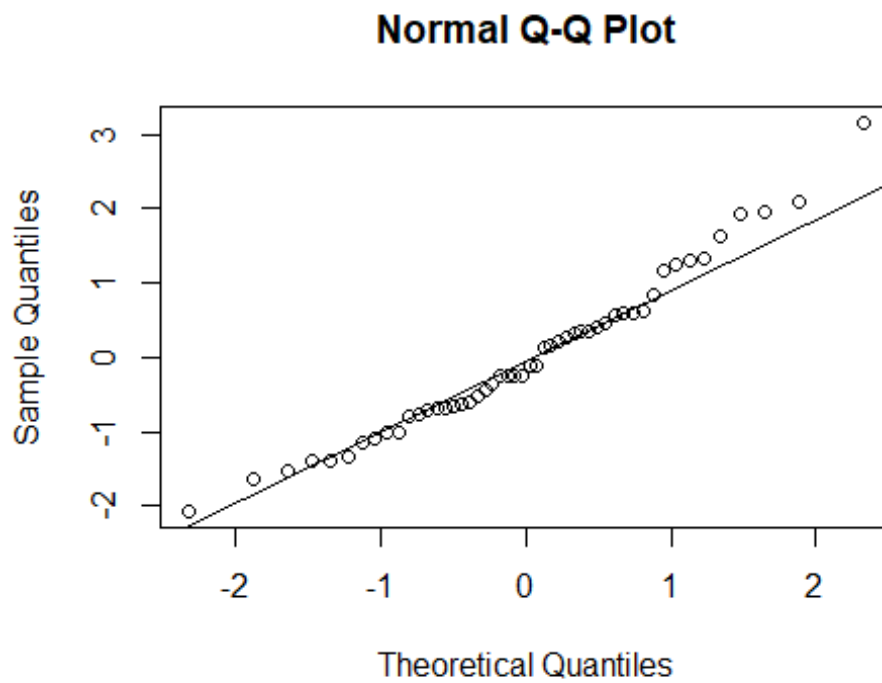
- Se rechaza H_0 si valor $p < \alpha$

```
ad.test(M2$residuals)
```

```
##  
## Anderson-Darling normality test  
##  
## data: M2$residuals  
## A = 0.39752, p-value = 0.3551
```

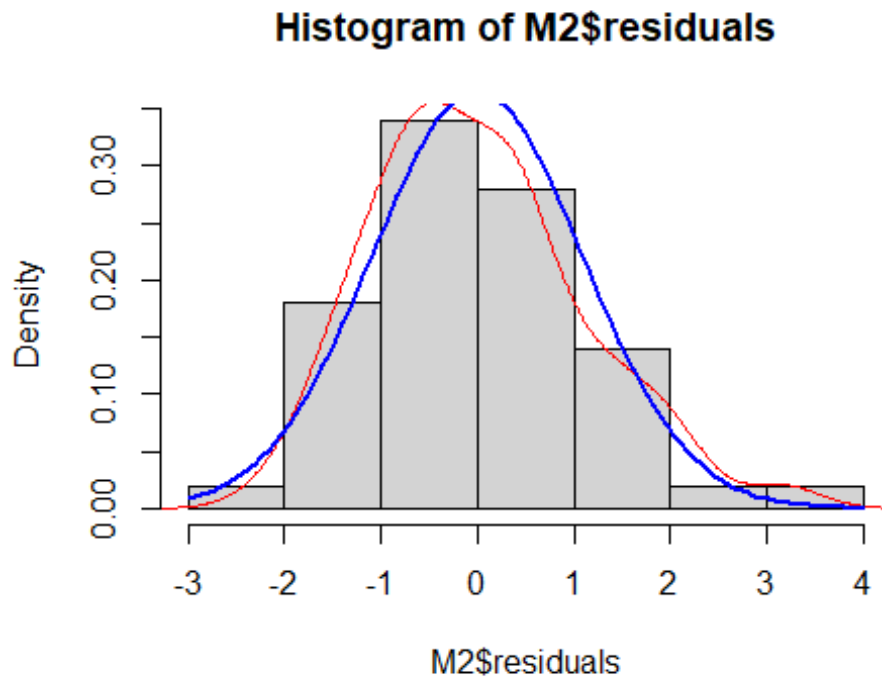
```
# Gráficas
```

```
qqnorm(M2$residuals)  
qqline(M2$residuals)
```



```
# Histograma
```

```
hist(M2$residuals, freq=FALSE)  
lines(density(M2$residual), col="red")  
curve(dnorm(x, mean=mean(M2$residuals), sd=sd(M2$residuals)), add=TRUE,  
col="blue", lwd=2)
```



Conclusión

No se rechaza la hipótesis nula (H_0). Esto indica que no hay evidencia significativa para afirmar que los datos no provienen de una población normal.

2. Verificación de media cero

Hipótesis

- $H_0: \mu_e = 0$
- $H_1: \mu_e \neq 0$

Valor frontera

$$\alpha = 0.03$$

Regla de decisión

- Se rechaza H_0 si valor $p < \alpha$

```
t.test(M2$residuals, mu = 0)
```

```
##
```

```
## One Sample t-test
```

```
##
```

```
## data: M2$residuals
```

```
## t = 3.9571e-16, df = 49, p-value = 1
```

```
## alternative hypothesis: true mean is not equal to 0
```

```
## 95 percent confidence interval:
## -0.3100858  0.3100858
## sample estimates:
## mean of x
## 6.105969e-17
```

Conclusión

No se rechaza la hipótesis nula (H_0). Esto indica que no hay evidencia significativa para afirmar que la media de los residuos es diferente de cero.

3. Homocedasticidad

Hipótesis

- H_0 : La varianza de los errores es constante (homocedasticidad)
- H_1 : La varianza de los errores no es constante (heterocedasticidad)

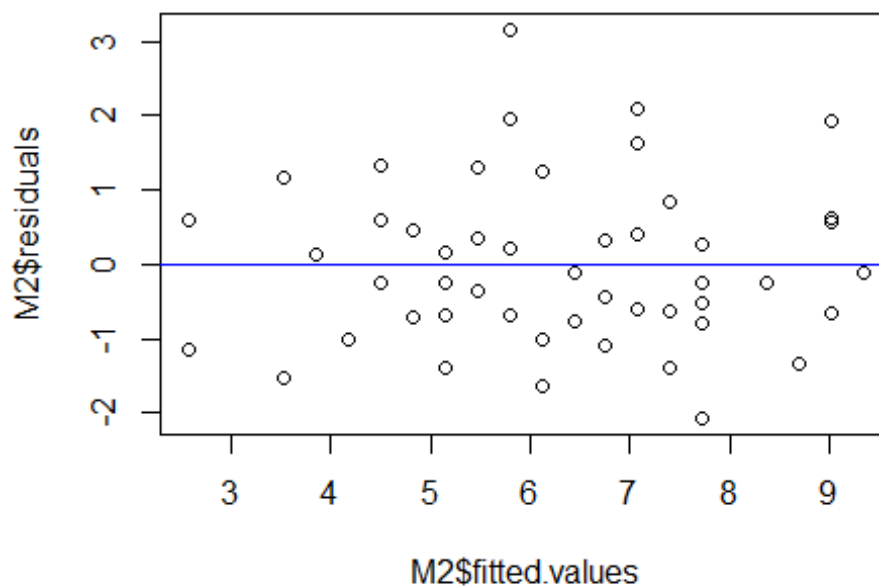
Valor frontera

$$\alpha = 0.03$$

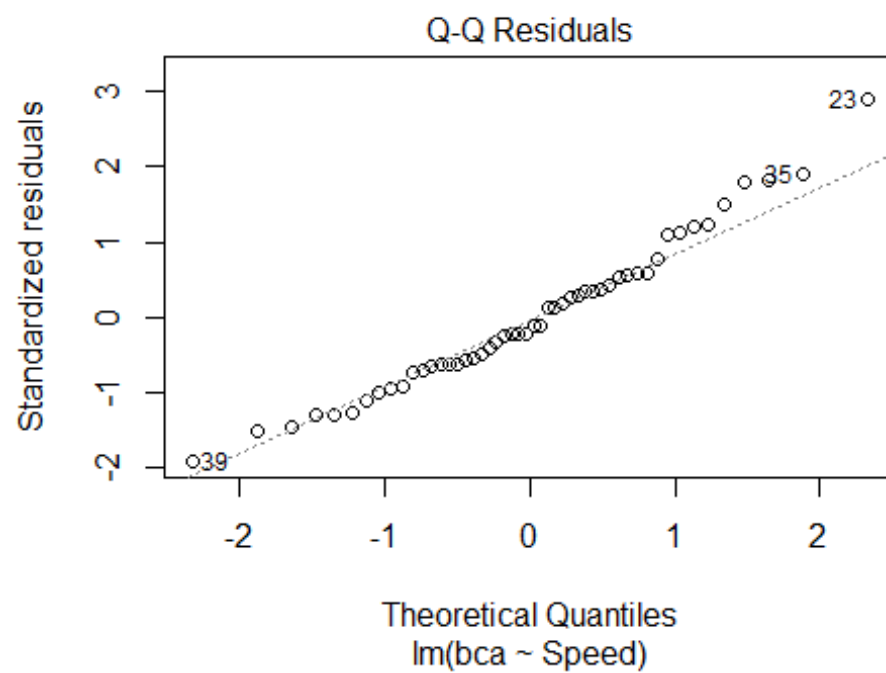
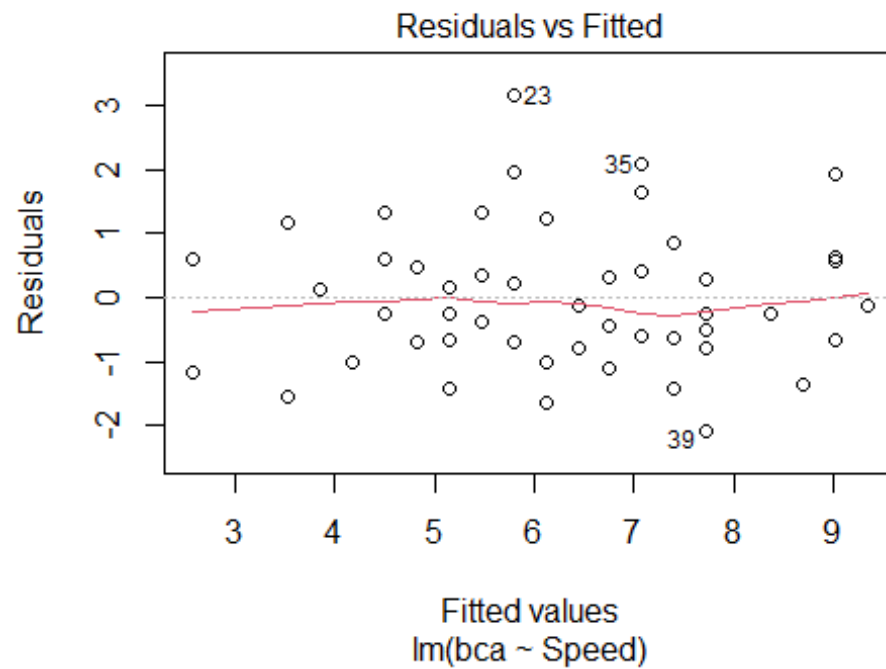
Regla de decisión

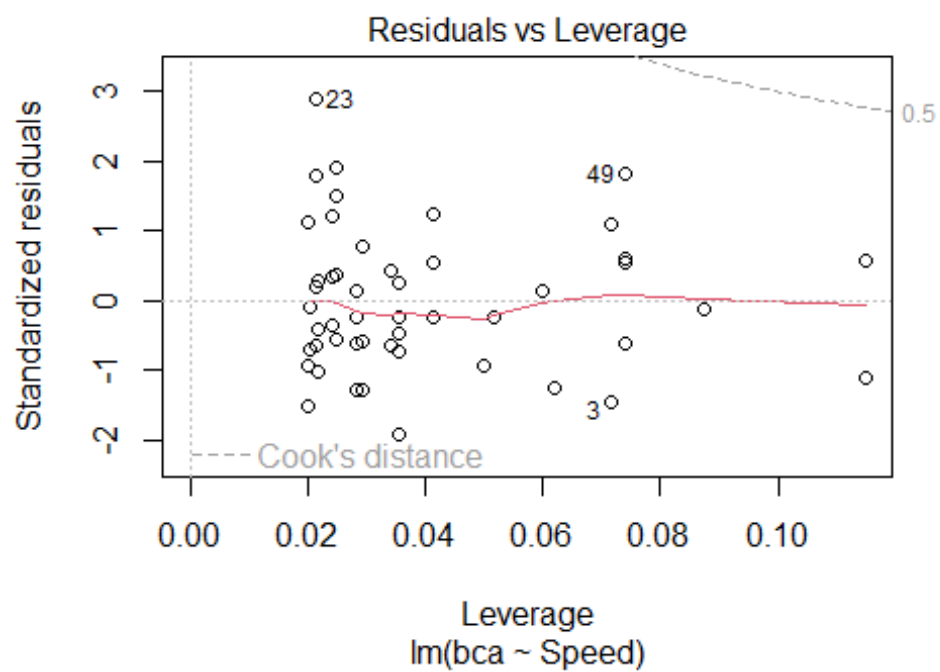
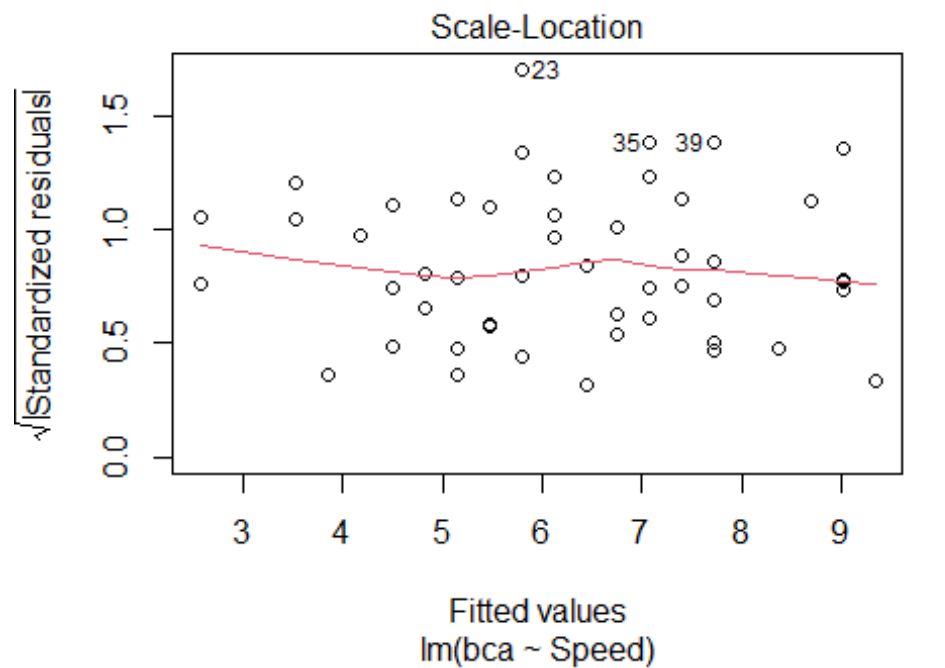
- Se rechaza H_0 si valor $p < \alpha$

```
plot(M2$fitted.values, M2$residuals)
abline(h=0, col = 'blue')
```




```
plot(M2)
```





```
bptest(M2)
```

```
##
```

```
## studentized Breusch-Pagan test
```

```
##
## data:  M2
## BP = 0.011192, df = 1, p-value = 0.9157
```

Conclusión

No se rechaza la hipótesis nula (H_0). Esto indica que no hay evidencia significativa para afirmar que la varianza de los errores no es constante (heterocedasticidad)

4. Linealidad

Hipótesis

- H_0 : No hay términos omitidos que indican linealidad
- H_1 : Hay una especificación errónea en el modelo que indica no linealidad

Valor frontera

$$\alpha = 0.04$$

Regla de decisión

- Se rechaza H_0 si valor $p < \alpha$

```
resettest(M2)
```

```
##
## RESET test
##
## data:  M2
## RESET = 0.47002, df1 = 2, df2 = 46, p-value = 0.628
```

Conclusión

No se rechaza la hipótesis nula (H_0). Esto indica que no hay evidencia significativa para afirmar que hay una especificación errónea en el modelo que indica no linealidad.

5. Independencia

Hipótesis

- H_0 : Los errores no están autocorrelacionados
- H_1 : Los errores están autocorrelacionados

Valor frontera

$$\alpha = 0.04$$

Regla de decisión

- Se rechaza H_0 si valor $p < \alpha$

```
bgtest(M2)
```

```
##
## Breusch-Godfrey test for serial correlation of order up to 1
##
## data: M2
## LM test = 0.015585, df = 1, p-value = 0.9007
```

Conclusión

No se rechaza la hipótesis nula (H_0). Esto indica que no hay evidencia significativa para afirmar que los errores están autocorrelacionados.

5. Despeja la distancia del modelo lineal obtenido entre la transformación y la velocidad. Obtendrás el modelo no lineal que relaciona la distancia con la velocidad directamente (y no con su transformación).

$$\text{Distancia(Transformada)} = \sqrt{\text{Distancia}}$$

$$\text{Distancia(Transformada)} = 1.27705 + 0.32241 * \text{Velocidad}$$

$$\sqrt{\text{Distancia}} = 1.27705 + 0.32241 * \text{Velocidad}$$

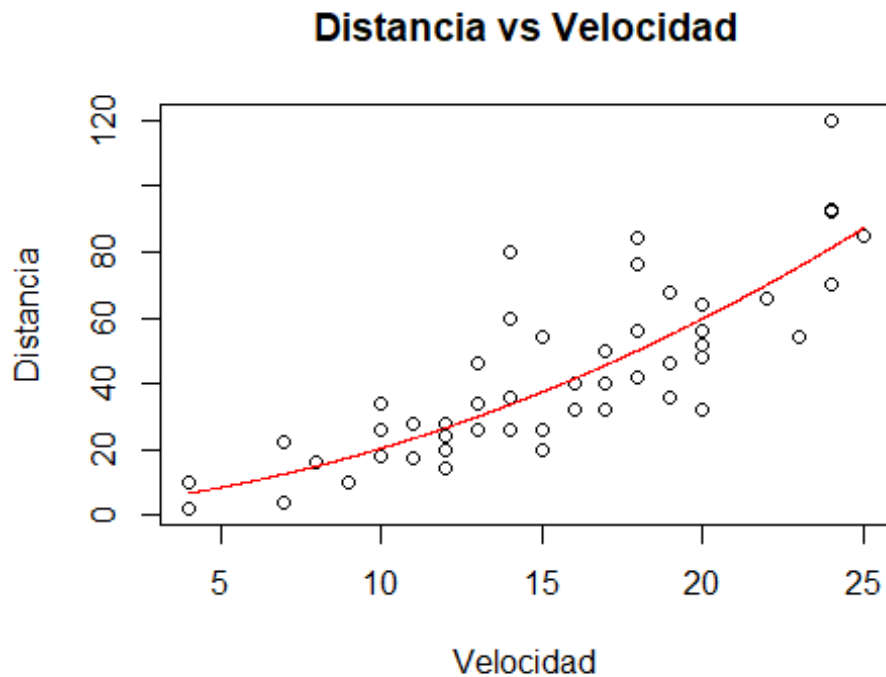
$$\text{Distancia} = (1.27705 + 0.32241 * \text{Velocidad})^2$$

```
b_0 <- M2$coefficients[1]
b_1 <- M2$coefficients[2]

m <- function(x){(b_0 + b_1 * x)^2}
```

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

```
plot(Speed, Dist, main = "Distancia vs Velocidad", ylab = "Distancia", xlab =
"Velocidad")
X <- seq(min(Speed), max(Speed), 0.01)
lines(X, m(X), col = "red", lw = 1.5)
```



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

Se observa que cumple con todas las pruebas de validez así como ser un modelo significativo al igual que de sus coeficientes. En este caso, el porcentaje de variación explicada por el modelo es de 70.94%

Parte 4: Conclusión

1. 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.

Aunque los 2 modelos son similares, el mejor modelo para describir la relación entre la distancia y la velocidad es el modelo 2 ya que, además de cumplir con la significancia y validez al igual que el modelo 1, el modelo 2 explica en un porcentaje de 70.94% la variación mientras que el modelo 1 solo explica el 65.11%

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

El problema principal con el modelo serían los datos atípicos, como se observa en la última gráfica, el modelo se ajusta bastante bien a los datos pero hay algunos datos por encima de la línea que se considerarían atípicos que harían parecer que el modelo no es tan preciso como realmente es, si se hiciera una limpieza aún más profunda (no se realizó porque los datos transformados se comportaban normalmente) se podrían tener un modelo

ligeramente más preciso. Es importante recalcar que computacionalmente es más económico la transformación aproximada a la exacta y hacer la transformación valió la pena para tener una mejora en el modelo a comparación de una lineal.