Act 13. Regresión no lineal

Oscar Gutierrez

2024-09-10

```
data = cars
```

Análisis de normalidad

h0: Los datos siguen una distribucion normal h1: Los datos no siguen una distribucion normal

```
library(nortest)
library(moments)

ad.test(data$speed)
```

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

```
jarque.test(data$speed)
```

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

```
ad.test(data$dist)
```

```
##
##  Anderson-Darling normality test
##
## data:  data$dist
## A = 0.74067, p-value = 0.05021
```

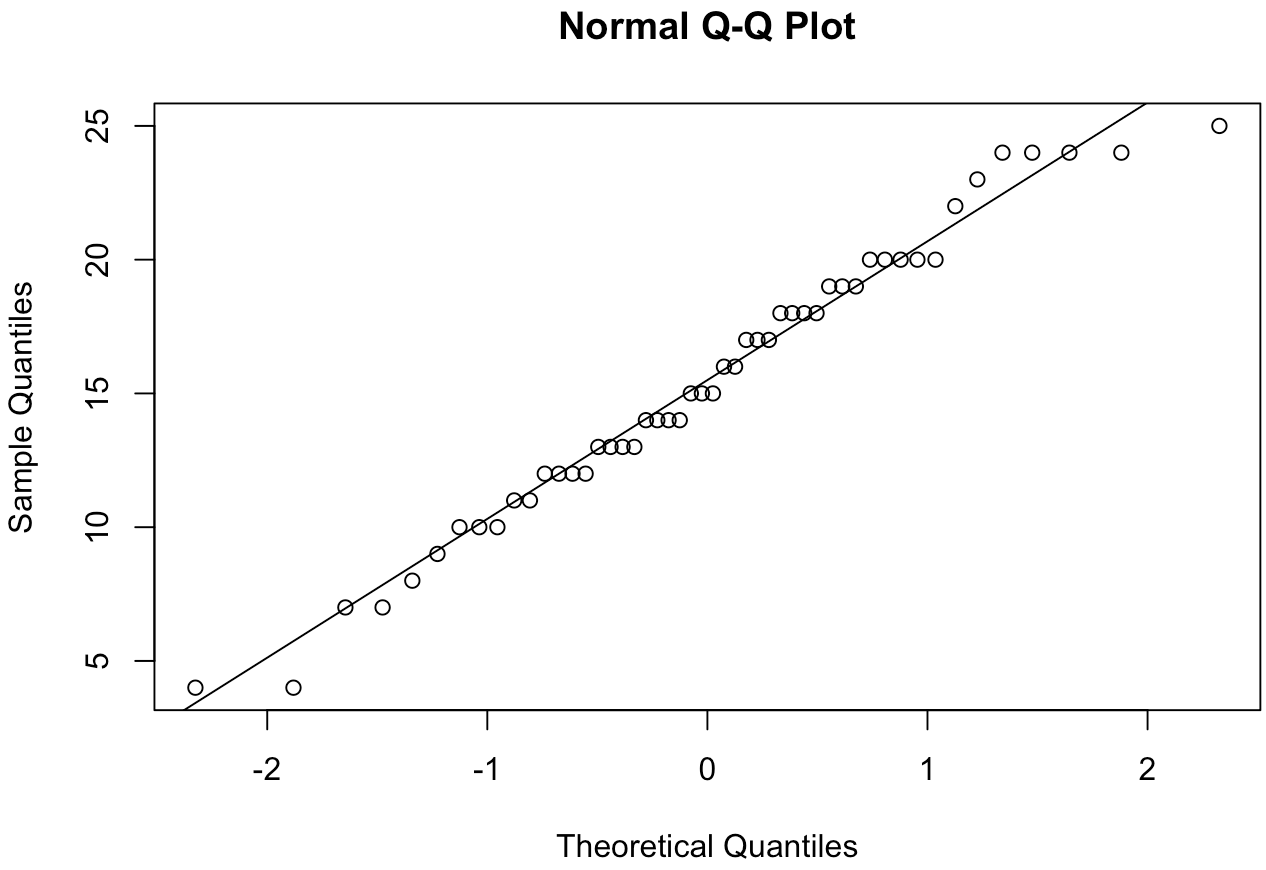
```
jarque.test(data$dist)
```

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

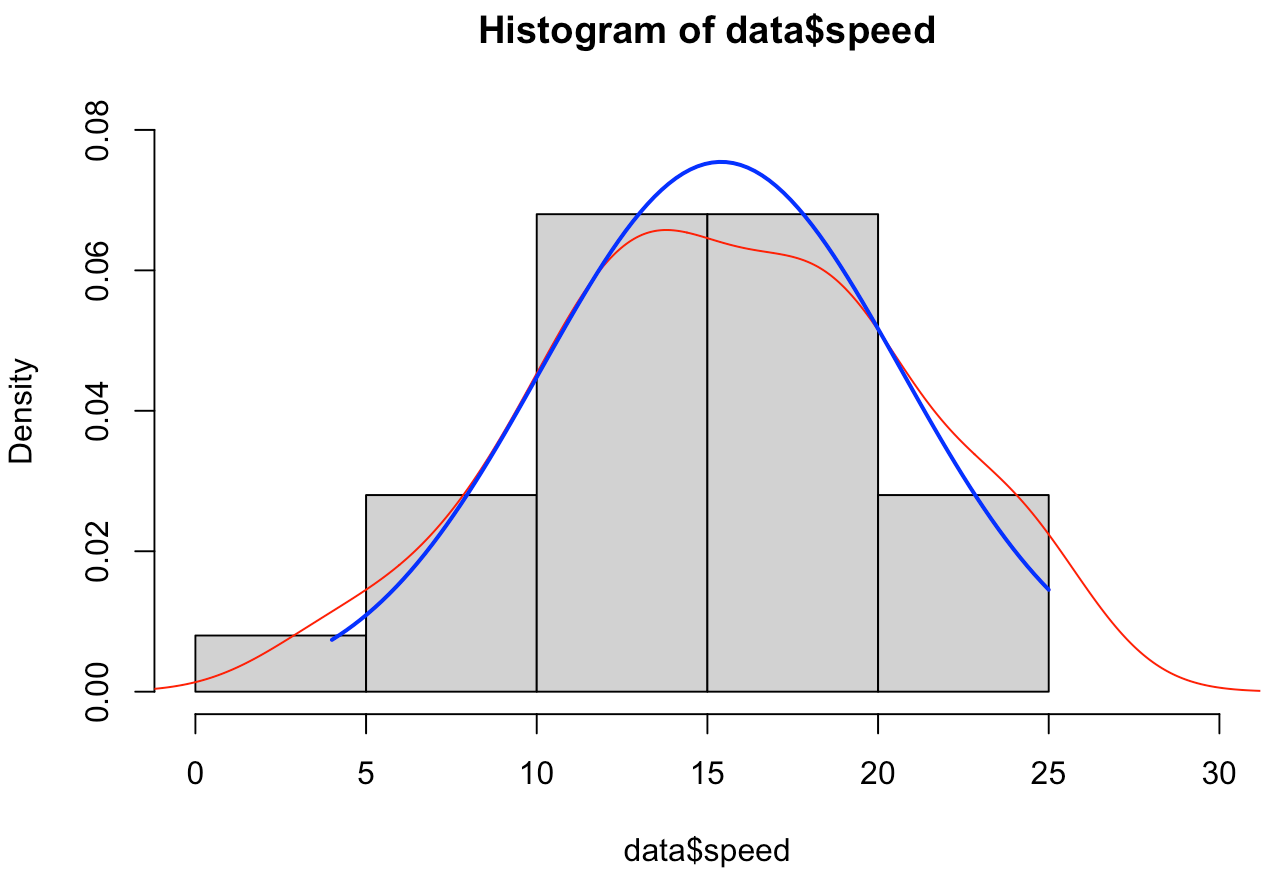
Con un α de 0.05, ambas variables siguen una distribución normal de acuerdo con las pruebas de Anderson-Darling y Jarque-Bera.

Visualización de gráficas de normalidad

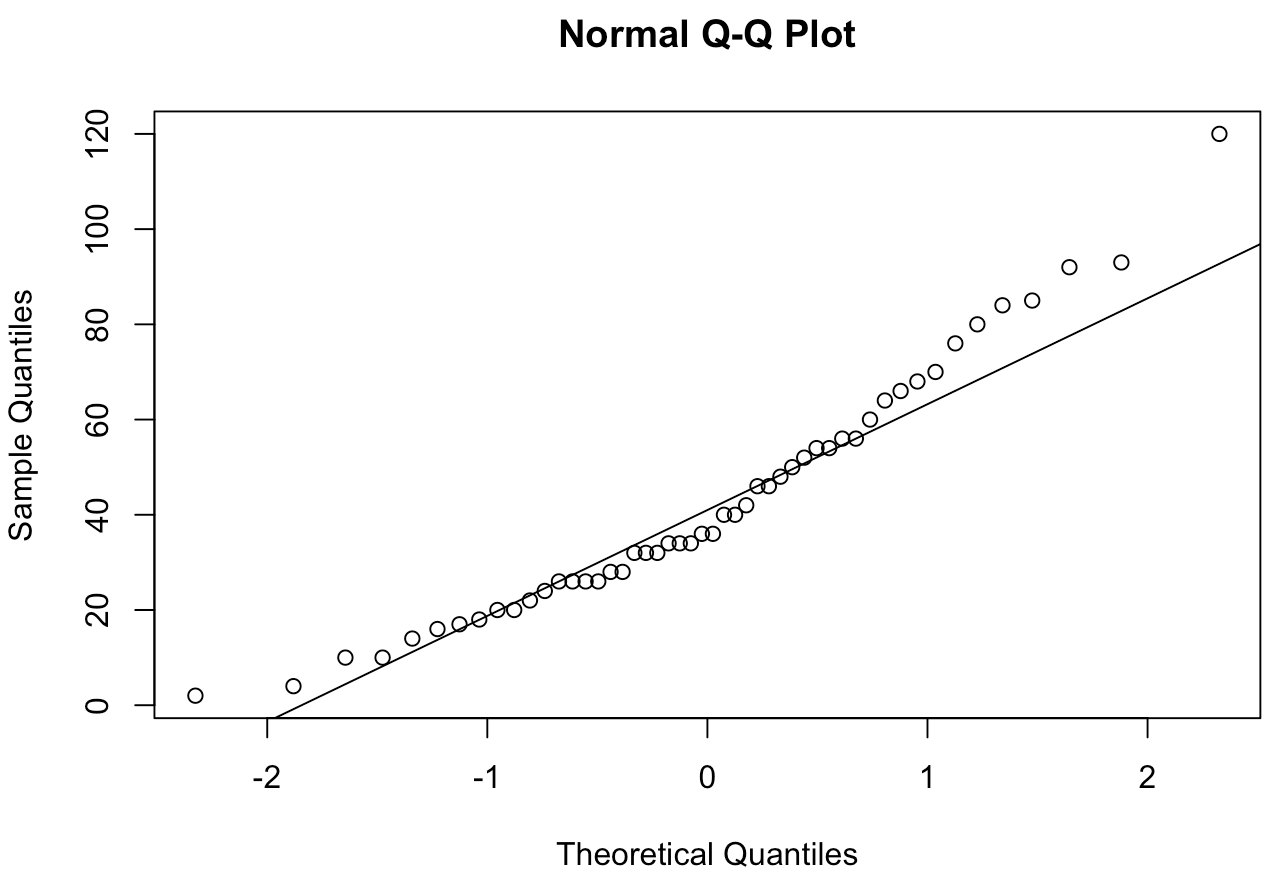
```
qqnorm(data$speed)
qqline(data$speed)
```



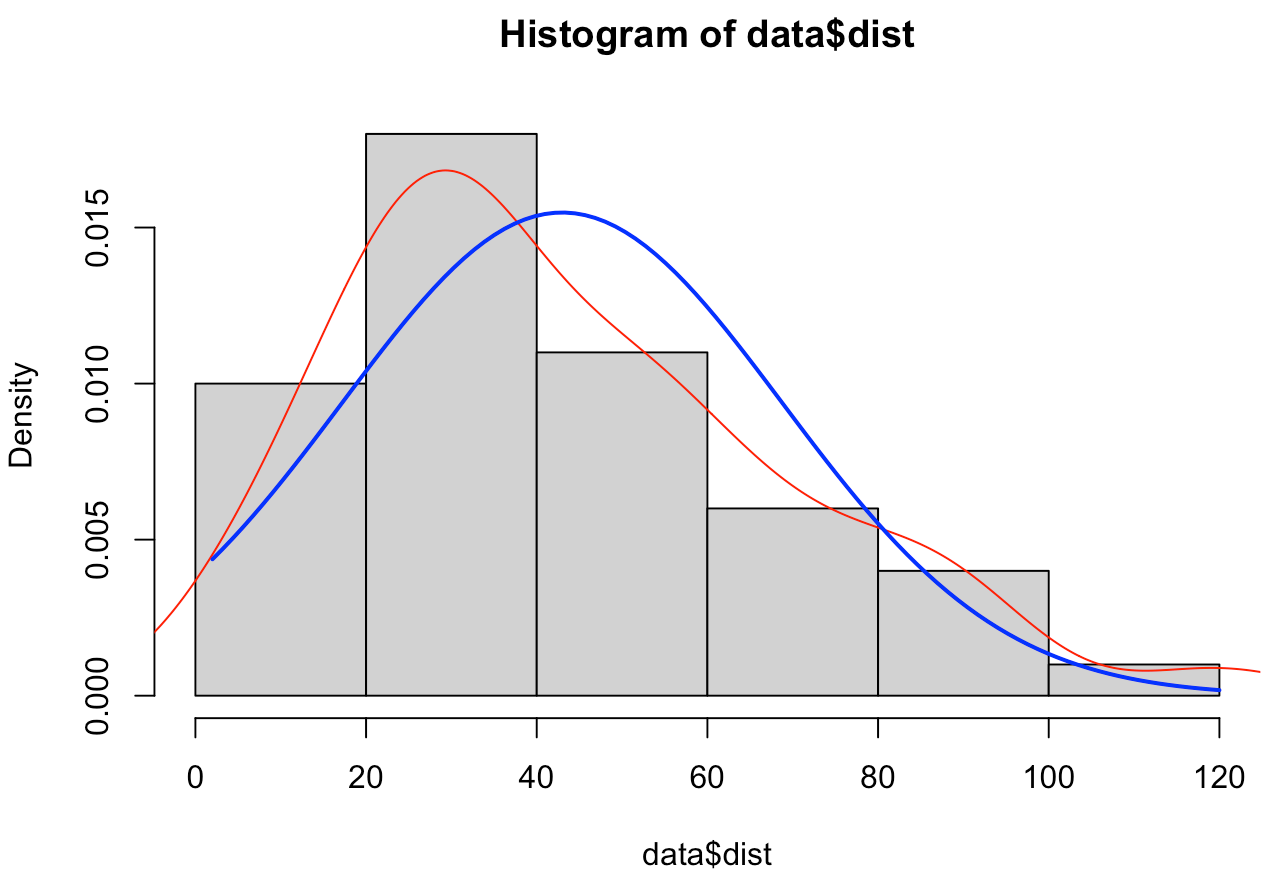

```
hist(data$speed,freq=FALSE, xlim = c(0,30), ylim=c(0,0.08))
lines(density(data$speed),col="red")
curve(dnorm(x,mean=mean(data$speed),sd=sd(data$speed)), from=min(data$speed), to=max(dat
a$speed), add=TRUE, col="blue",lwd=2)
```




```
qqnorm(data$dist)
qqline(data$dist)
```




```
hist(data$dist,freq=FALSE)
lines(density(data$dist),col="red")
curve(dnorm(x,mean=mean(data$dist),sd=sd(data$dist)), from=min(data$dist), to=max(data$d
ist), add=TRUE, col="blue",lwd=2)
```



Las gráficas de la variable speed se ven mucho más apegadas a lo que es una distribución normal que las gráficas de la variable dist, esta variable pierde un poco de su normalidad en las colas de la distribución. Sin embargo, las pruebas de hipótesis realizadas anteriormente confirman la normalidad de ambas variables.

Sesgo y kurtosis

```
cat('Sesgo speed=', skewness(data$speed))
```

```
## Sesgo speed= -0.1139548
```

```
cat('\nSesgo dist=', skewness(data$dist))
```

```
##
## Sesgo dist= 0.7824835
```

```
cat('\nCurtosis speed=', kurtosis(data$speed))
```

```
##
## Curtosis speed= 2.422853
```

```
cat('\nCurtosis dist=', kurtosis(data$dist))
```

```
##
## Curtosis dist= 3.248019
```

Los valors de sesgo están relativamente cerca de 0 y los de curtosis cerca de 3, lo que significa que estos valores se asemejan a los de una distribución normal.

Modelo de regresión lineal

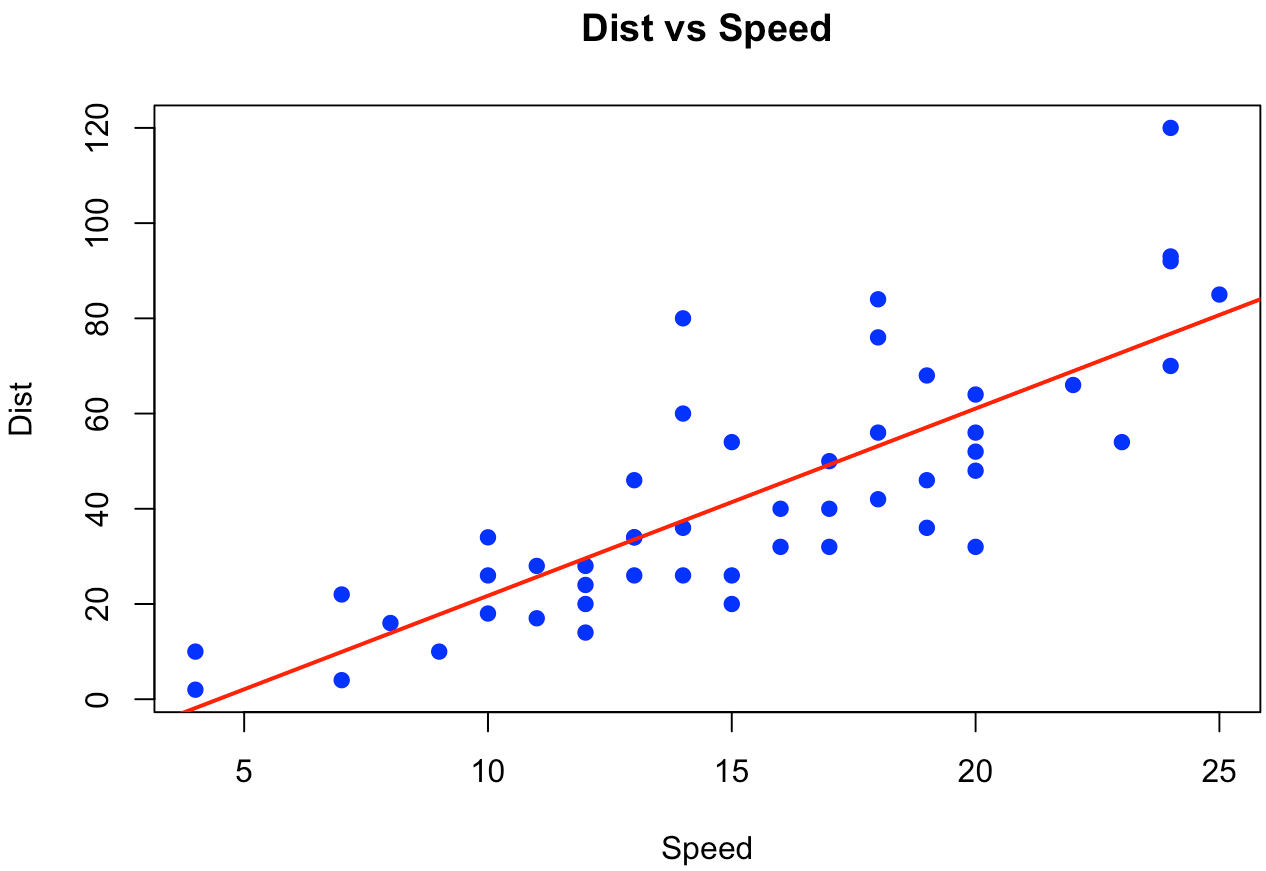
```
Modelo1 = lm(dist ~ speed, data = data)
summary(Modelo1)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = data)
##
## 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
```

Tanto el intercept como la variable independiente (speed) son significantes con un alpha de 0.05. El modelo logra explicar un 64.38% de la variación en la variable de interés. El modelo en general es significativo ya que el valor p es menor a alpha.

El modelo obtenido dist = 3.9324*speed - 17.5791

```
plot(data$speed, data$dist, main="Dist vs Speed",
     xlab="Speed", ylab="Dist", pch=19, col="blue")
abline(Modelo1, col="red", lwd=2)
```



Análisis de residuos

Normalidad de residuos

H_0 : Los residuos se distribuyen normalmente H_1 : Los residuos no se distribuyen normalmente

```
ad.test(Modelo1$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  Modelo1$residuals
## A = 0.79406, p-value = 0.0369
```

```
jarque.test(Modelo1$residuals)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  Modelo1$residuals
## JB = 8.1888, p-value = 0.01667
## alternative hypothesis: greater
```

De acuerdo con estas dos pruebas, considerando un alpha de 0.05, los residuos no se distribuyen normalmente.

Comprobar media = 0

$H_0 : \mu = 0 \, H_1 : mu \neq 0$

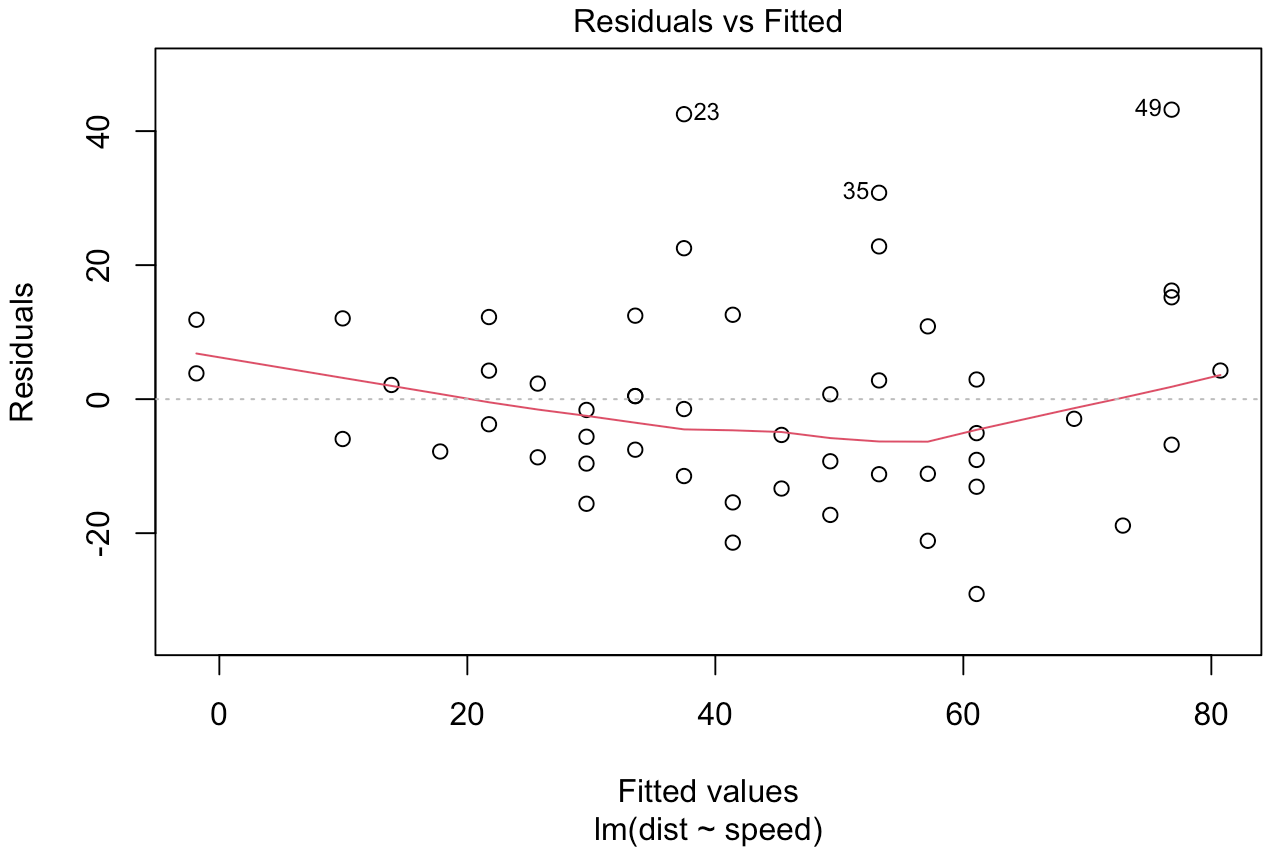
```
t.test(Modelo1$residuals)
```

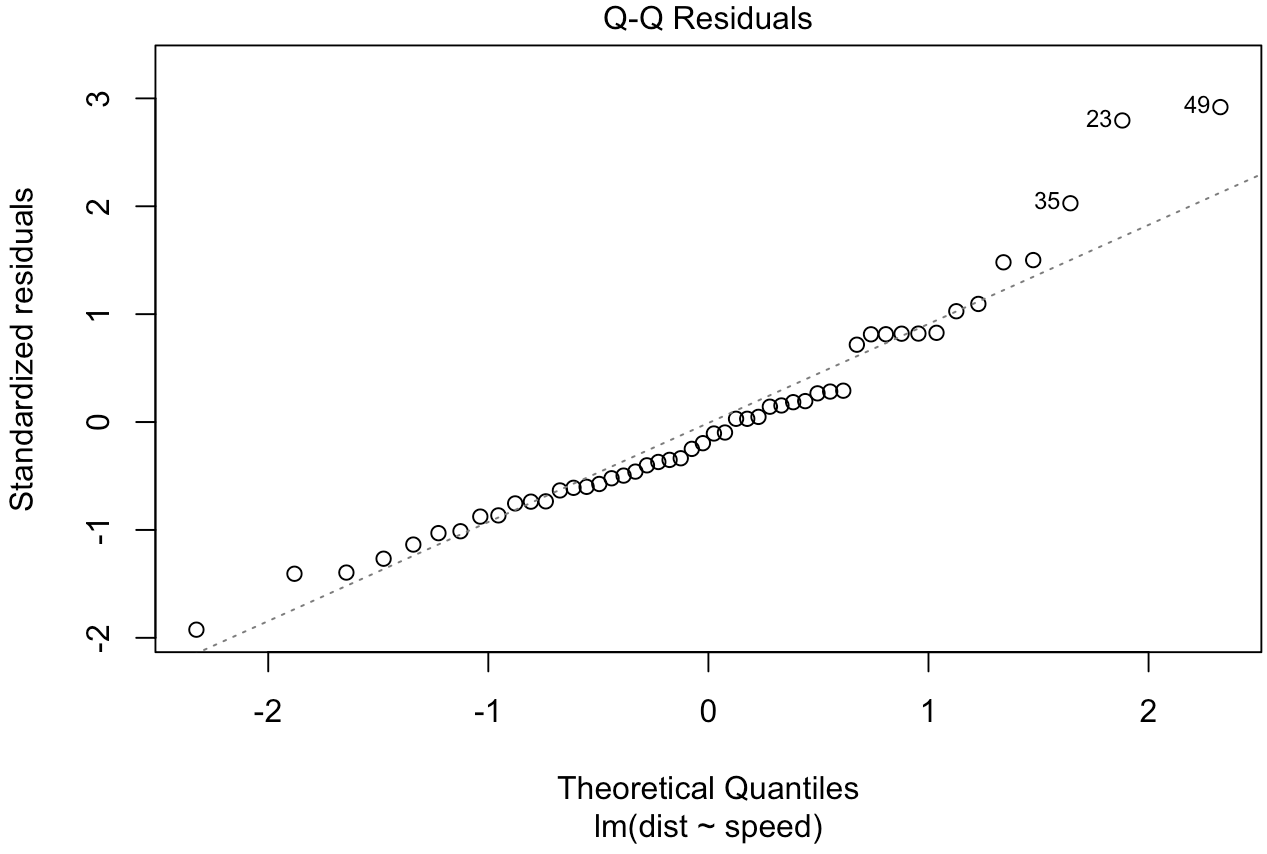
```
##
##  One Sample t-test
##
## data:  Modelo1$residuals
## t = -2.0629e-16, df = 49, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -4.326  4.326
## sample estimates:
##     mean of x
## -4.440892e-16
```

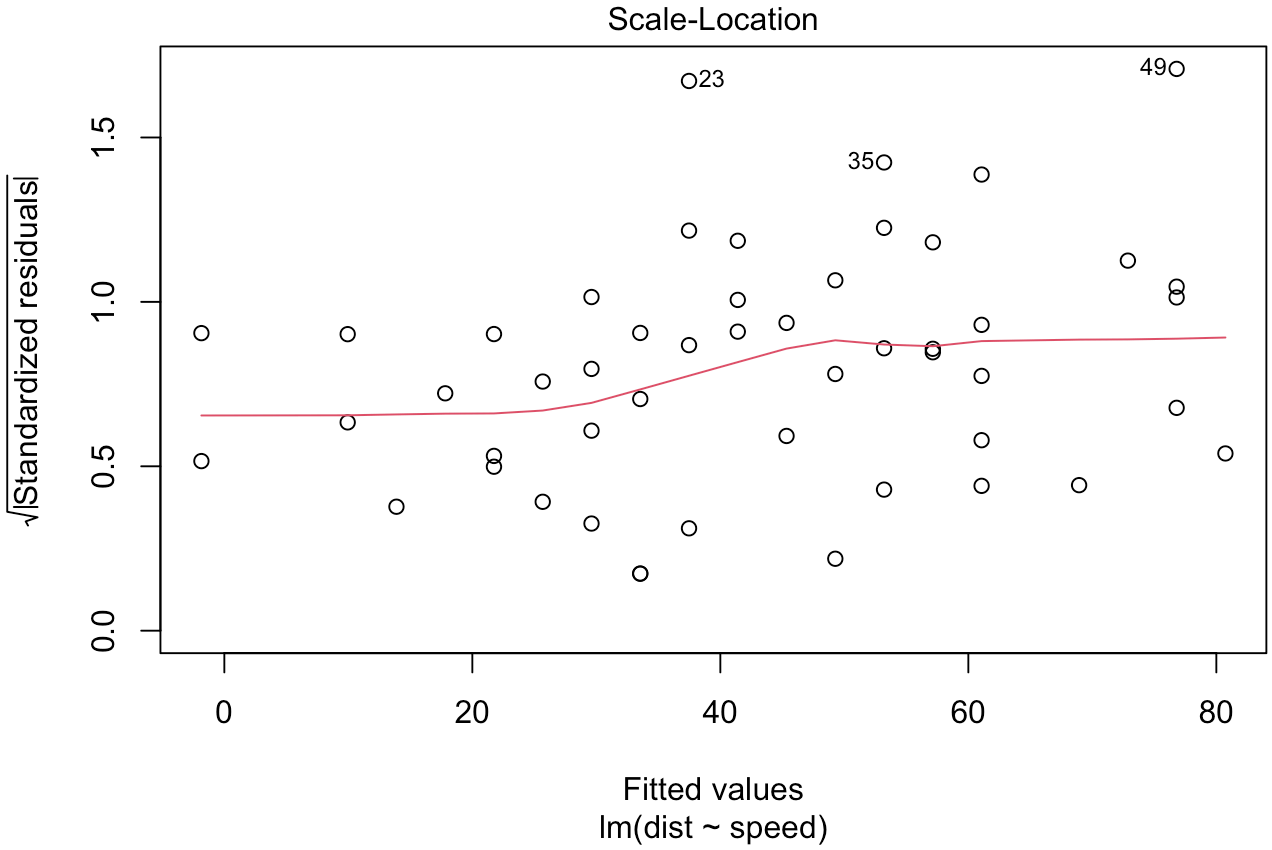
La media de los residuos sí es igual a 0, ya que el valor p es mayor a alpha.

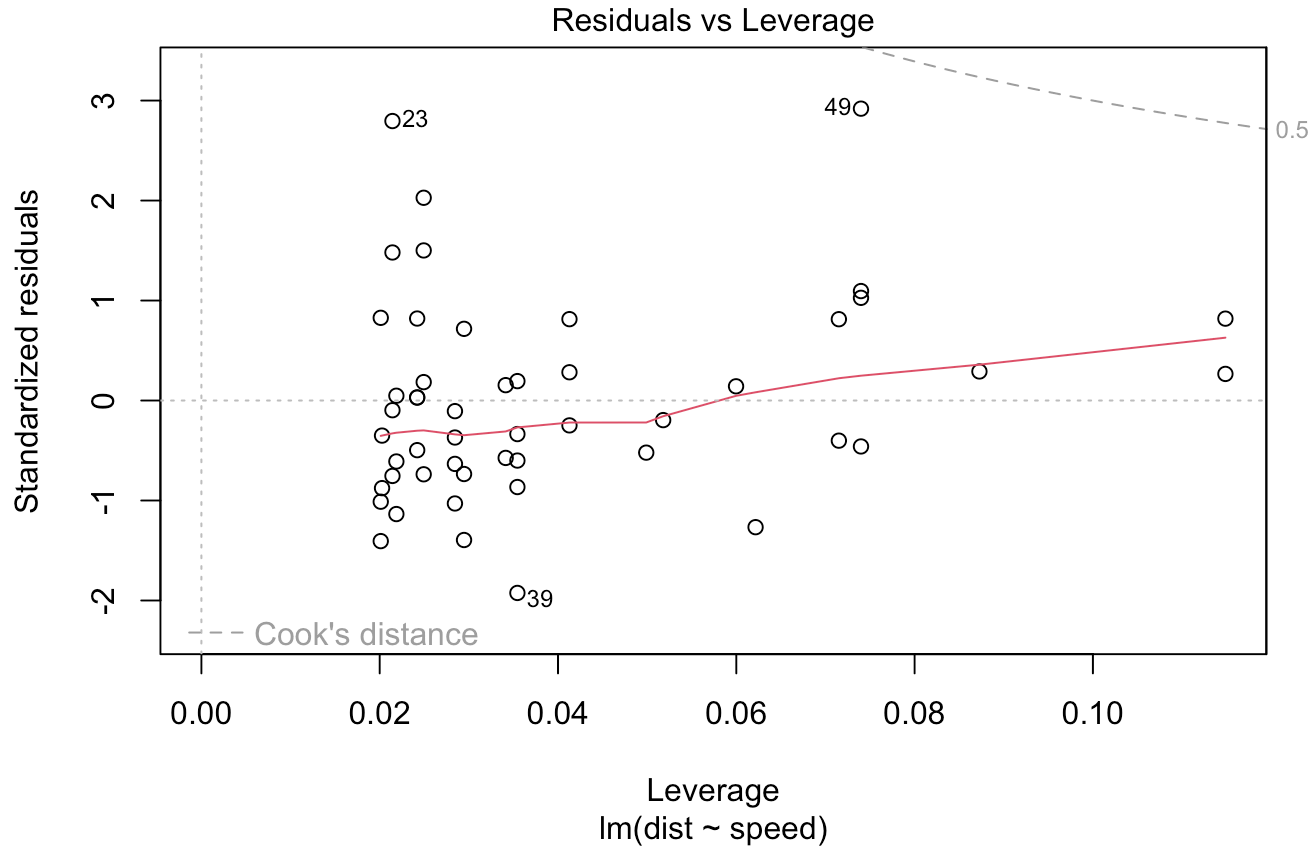
Homocedasticidad e independencia

```
plot(Modelo1)
```


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


Scale-Location
23
49
35
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
20
40
60
80
Fitted values
lm(dist ~ speed)


Residuals vs Leverage
23
49
39
0.5
Cook's distance
Standardized residuals
-2
-1
0
1
2
3
0.00
0.02
0.04
0.06
0.08
0.10
Leverage
lm(dist ~ speed)

De acuerdo con la gráfica de residuals vs fitted values parece haber una tendencia en los errores. El QQ plot muestra que no hay normalidad a lo largo de toda la distribución, especialmente en las colas.

Prueba de independencia

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

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

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

```
dwtest(Modelo1)
```

```
##
##  Durbin-Watson test
##
## data:  Modelo1
## DW = 1.6762, p-value = 0.09522
## alternative hypothesis: true autocorrelation is greater than 0
```

```
bgtest(Modelo1)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  Modelo1
## LM test = 1.2908, df = 1, p-value = 0.2559
```

Las pruebas de hipótesis muestran que no hay autocorrelación entre errores ya que el valor p es mayor a alpha.

Prueba de homocedasticidad

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

```
gqtest(Modelo1)
```

```
##
##  Goldfeld-Quandt test
##
## data:  Modelo1
## GQ = 1.5512, df1 = 23, df2 = 23, p-value = 0.1498
## alternative hypothesis: variance increases from segment 1 to 2
```

Sí hay varianza constante en los errores, hay homocedasticidad.

Significancia de modelo 1

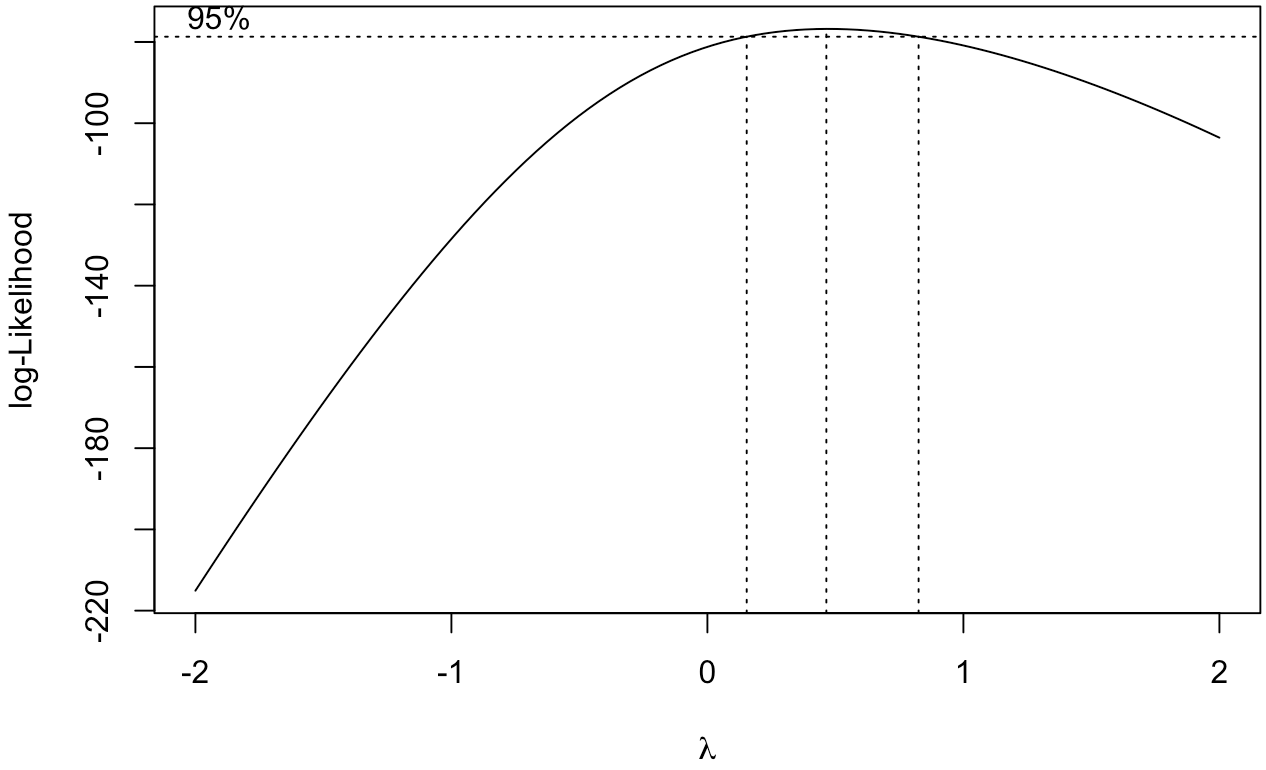
El modelo logra explicar un 64% de la variación, este modelo puede ser útil para predecir los valores de dist en función de speed, las variable dependiente, el intercept y el modelo son significativos de acuerdo con las pruebas de hipótesis, sin embargo, el modelo no cumple todos los supuestos de la regresión lineal, por lo que es posible encontrar un mejor modelo, especialmente un modelo no lineal ya que las gráficas de errores muestran una posible relación que este modelo no está considerando.

Regresión no lineal

Transformacion Box-Cox

Se aplica una transformación de box cox para la variable dist ya que es la más alejada de distribuirse como una normal.

```
library(MASS)
bc<-boxcox((data$dist+1)~1)
```




```
l=bc$x[which.max(bc$y)]
cat('Lambda=',l)
```

```
## Lambda= 0.4646465
```

Aplicamos la transformación. La ecuacion para la transformacion exacta es $\frac{x^\lambda + 1}{\lambda}$. Para la transformada simple es $\sqrt{x+1}$

```
dist_te = ((data$dist +1 )^l - 1)/l

dist_ta = sqrt(data$dist+1)
```

Una vez realizadas las transformaciones, se comprueba normalidad.

```
cat('Sesgo dist transf exacta=', skewness(dist_te))
```

```
## Sesgo dist transf exacta= -0.03965266
```

```
cat('\nCurtosis dist tranf exacta=', kurtosis(dist_te))
```

```
##
## Curtosis dist tranf exacta= 2.779542
```

```
cat('\nSesgo dist trasf simple', skewness(dist_ta))
```

```
##
## Sesgo dist trasf simple 0.0250565
```

```
cat('\nCurtosis dist transf simple', kurtosis(dist_ta))
```

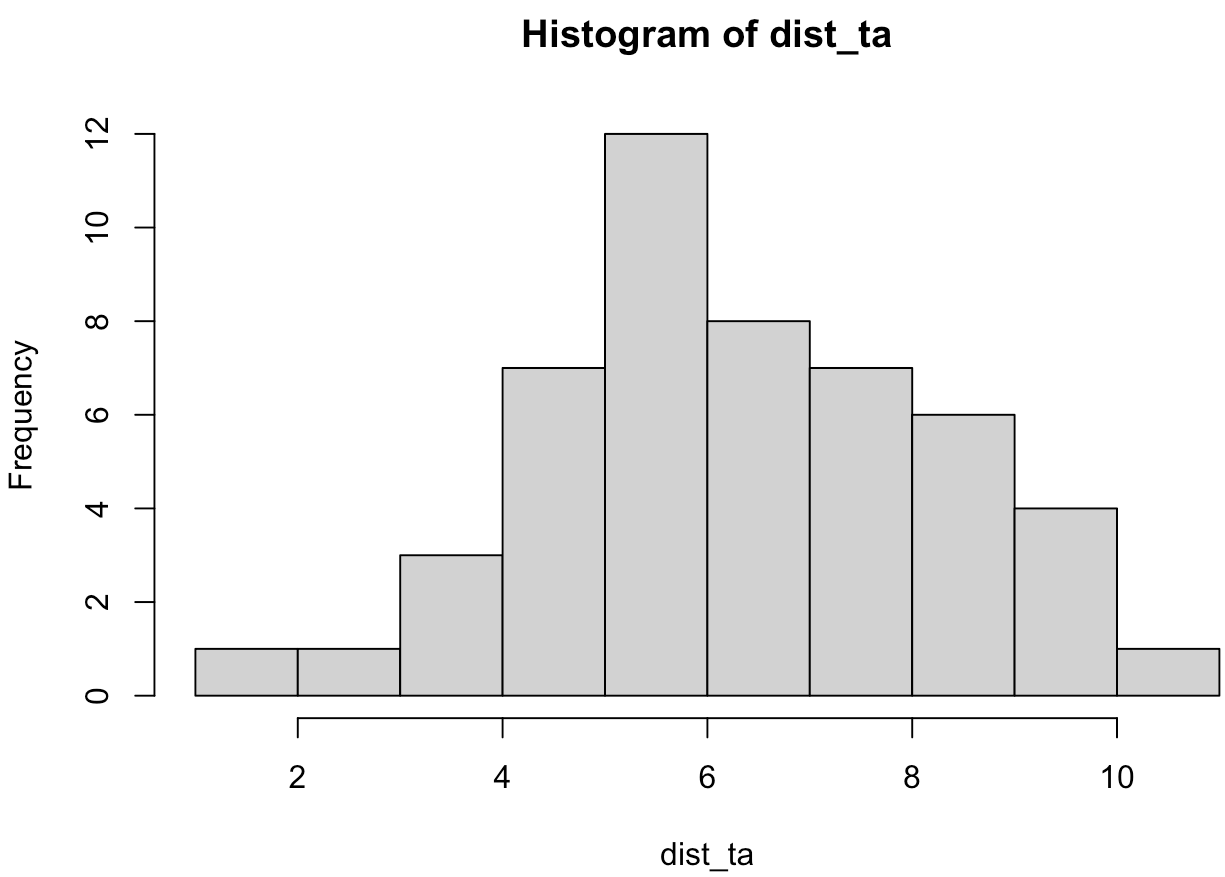
```
##
## Curtosis dist transf simple 2.743943
```

Los valores para sesgo y curtosis de ambas tansformaciones son muy similares, se obtienen valores similares a los de una distribucion normal.

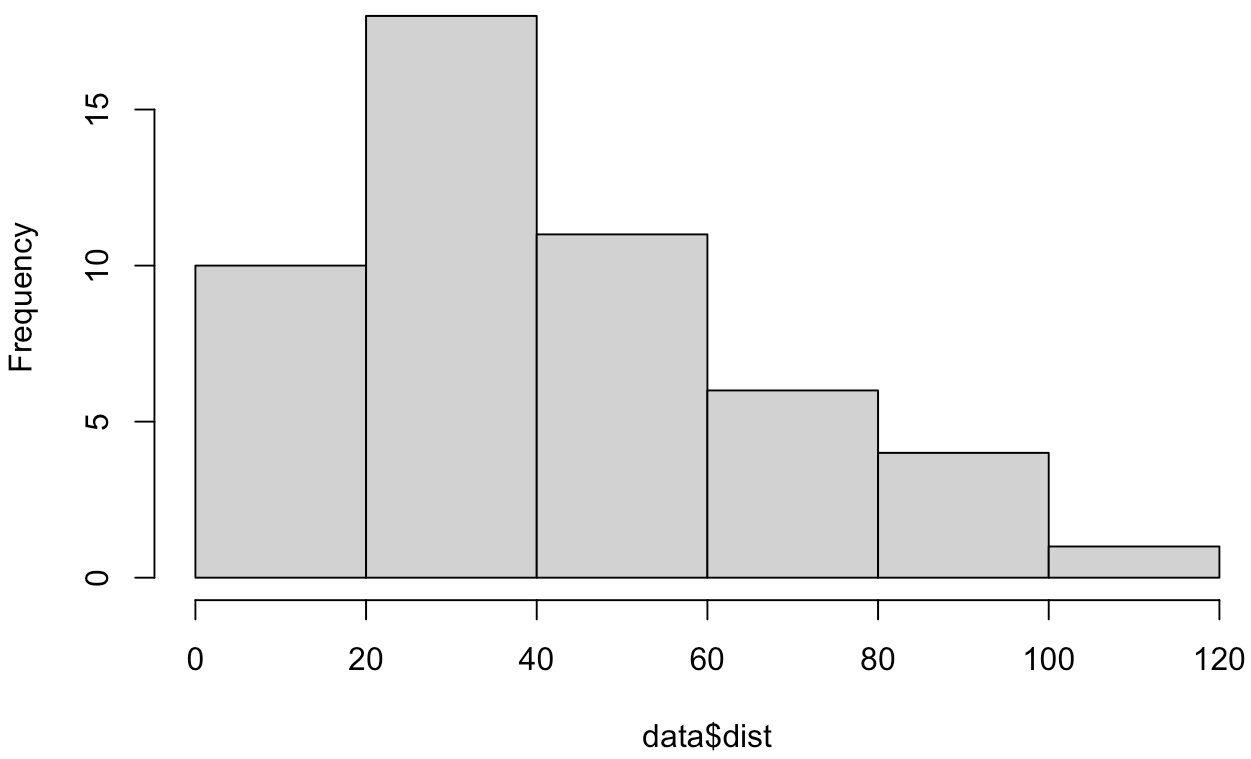
Ahora visualicemos los histogramas.

```
hist(dist_te)
```


```
hist(dist_ta)
```




```
hist(data$dist)
```



Ambas tranformaciones se ven mucho más normal que la variable sin transformar, sin embargo, ambas fallan un poco en las colas de la distribucion.

Ahora confirmemos con pruebas de hipótesis.

H_0 : Se distribuye normalmente H_1 : No se distribuye normalmente

```
ad.test(dist_te)
```

```
##
##  Anderson-Darling normality test
##
## data:  dist_te
## A = 0.13452, p-value = 0.9775
```

```
jarque.test(dist_te)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  dist_te
## JB = 0.11436, p-value = 0.9444
## alternative hypothesis: greater
```

```
ad.test(dist_ta)
```

```
##
##  Anderson-Darling normality test
##
## data:  dist_ta
## A = 0.14191, p-value = 0.97
```

```
jarque.test(dist_ta)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  dist_ta
## JB = 0.14183, p-value = 0.9315
## alternative hypothesis: greater
```

La transformada exacta obtiene valores un poco mejores que la aproximada, sin embargo la mejor transformación es la aproximada ya que es una transformación mucho más simple y los resultados fueron muy similares.

Modelo 2

Ahora realicemos un modelo utilizando esta nueva variable dist con una transformación aproximada.

```
Modelo2 = lm(dist_ta~data$speed)
summary(Modelo2)
```

```
##
## Call:
## lm(formula = dist_ta ~ data$speed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0430 -0.6992 -0.1773  0.5815  3.1087
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.46680    0.47605   3.081  0.00341 **
## data$speed   0.31604    0.02927  10.798 1.93e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.083 on 48 degrees of freedom
## Multiple R-squared:  0.7084, Adjusted R-squared:  0.7023
## F-statistic: 116.6 on 1 and 48 DF,  p-value: 1.931e-14
```

Tanto el modelo como las variables involucradas son significativas con un alpha de 0.05, este segundo modelo logra explicar un poco más de la variación en la variable de interés, obteniendo un coeficiente de determinación de 0.7023.

Análisis de residuos

Normalidad de residuos

H_0 : Los residuos se distribuyen normalmente H_1 : Los residuos no se distribuyen normalmente

```
ad.test(Modelo2$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  Modelo2$residuals
## A = 0.41653, p-value = 0.3198
```

```
jarque.test(Modelo2$residuals)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  Modelo2$residuals
## JB = 3.0788, p-value = 0.2145
## alternative hypothesis: greater
```

Los residuos se distribuyen normalmente ya que el valor p es mayor a un alpha de 0.05.

Comprobar media = 0

$H_0 : \mu = 0$ $H_1 : mu \neq 0$

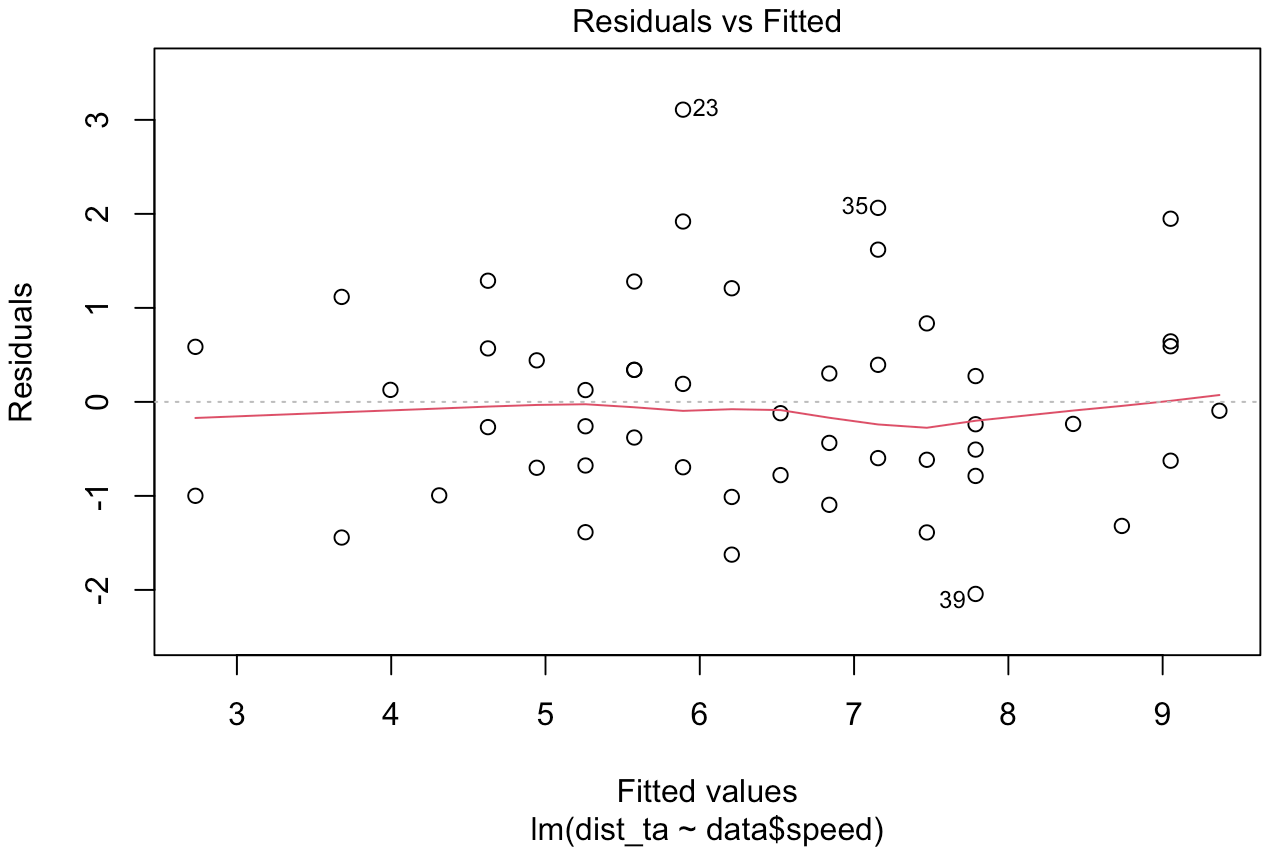
```
t.test(Modelo2$residuals)
```

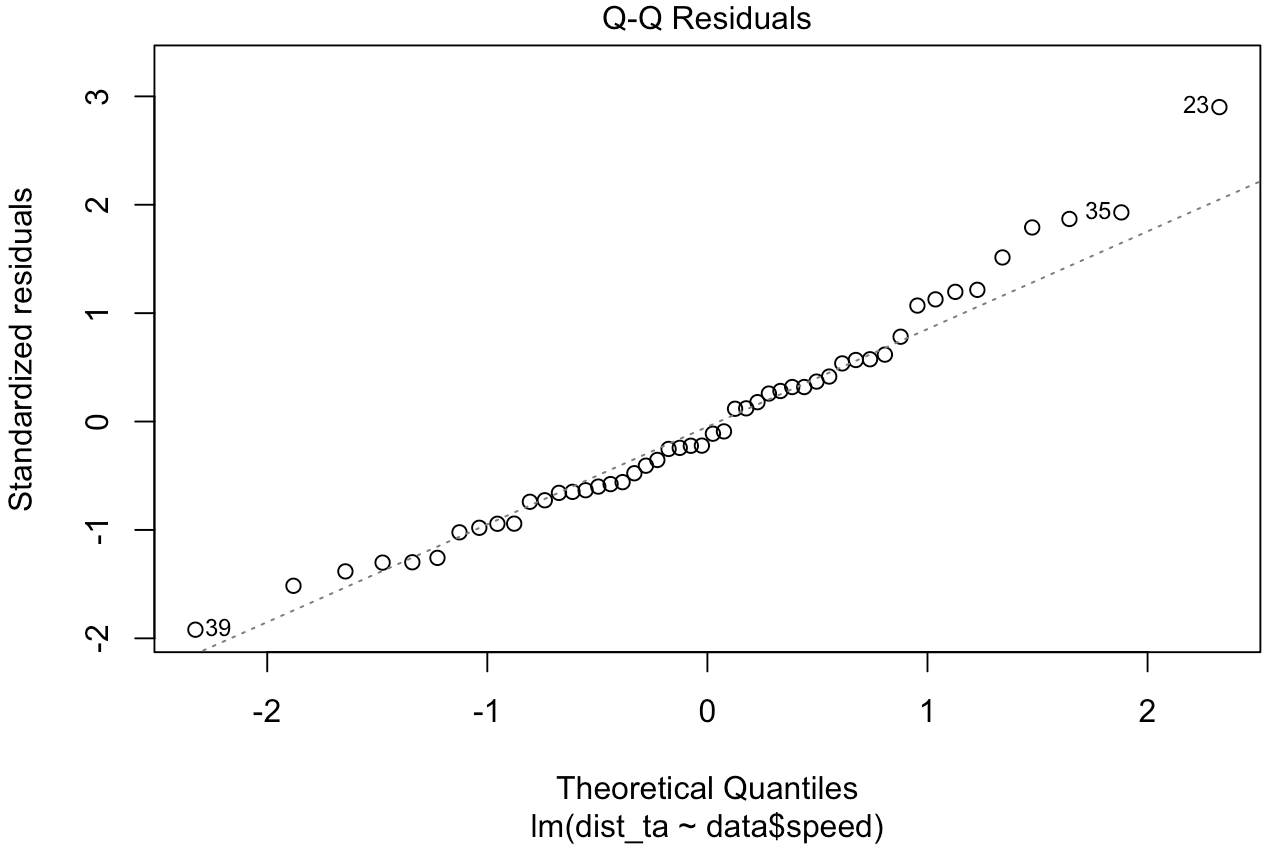
```
##
##  One Sample t-test
##
## data:  Modelo2$residuals
## t = 3.2216e-16, df = 49, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.3047124  0.3047124
## sample estimates:
##    mean of x
## 4.884981e-17
```

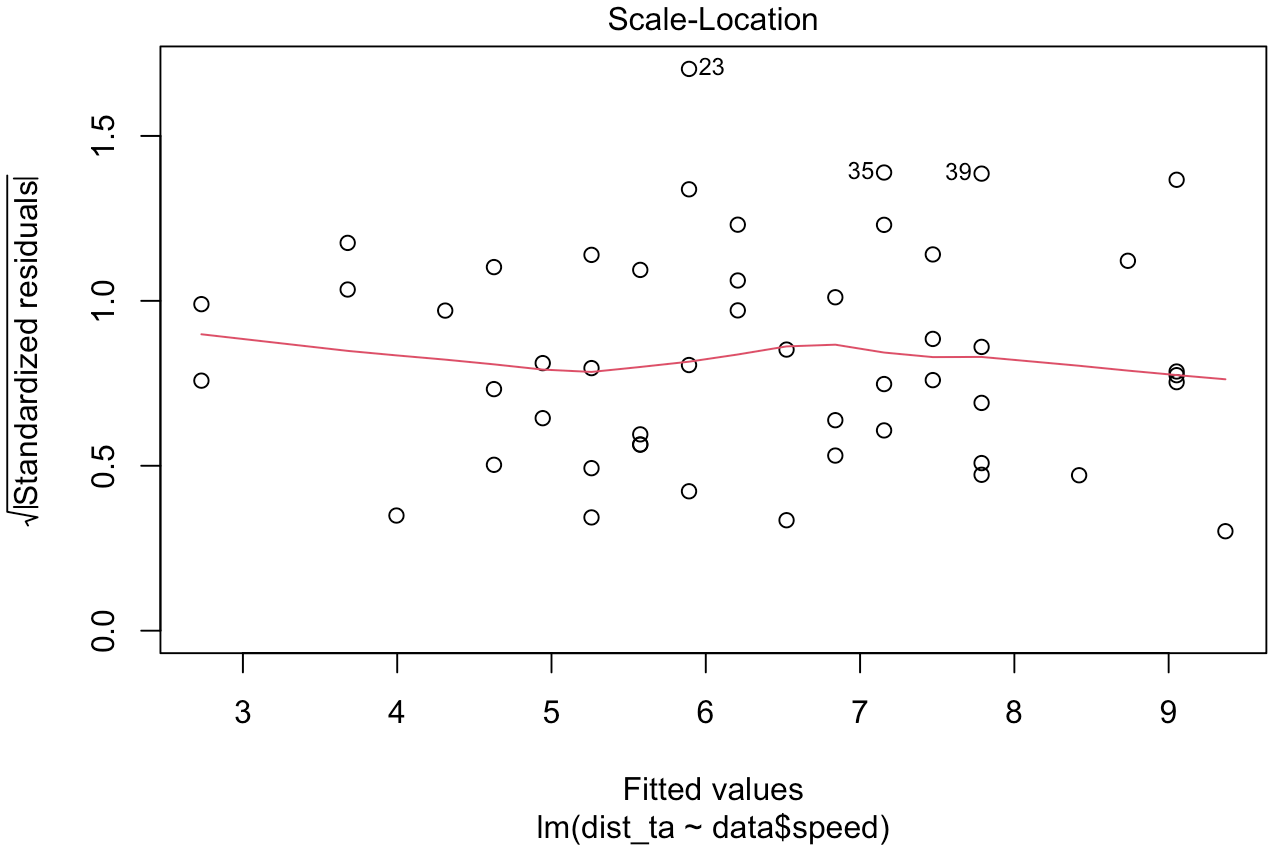
La media de los residuos es igual a 0.

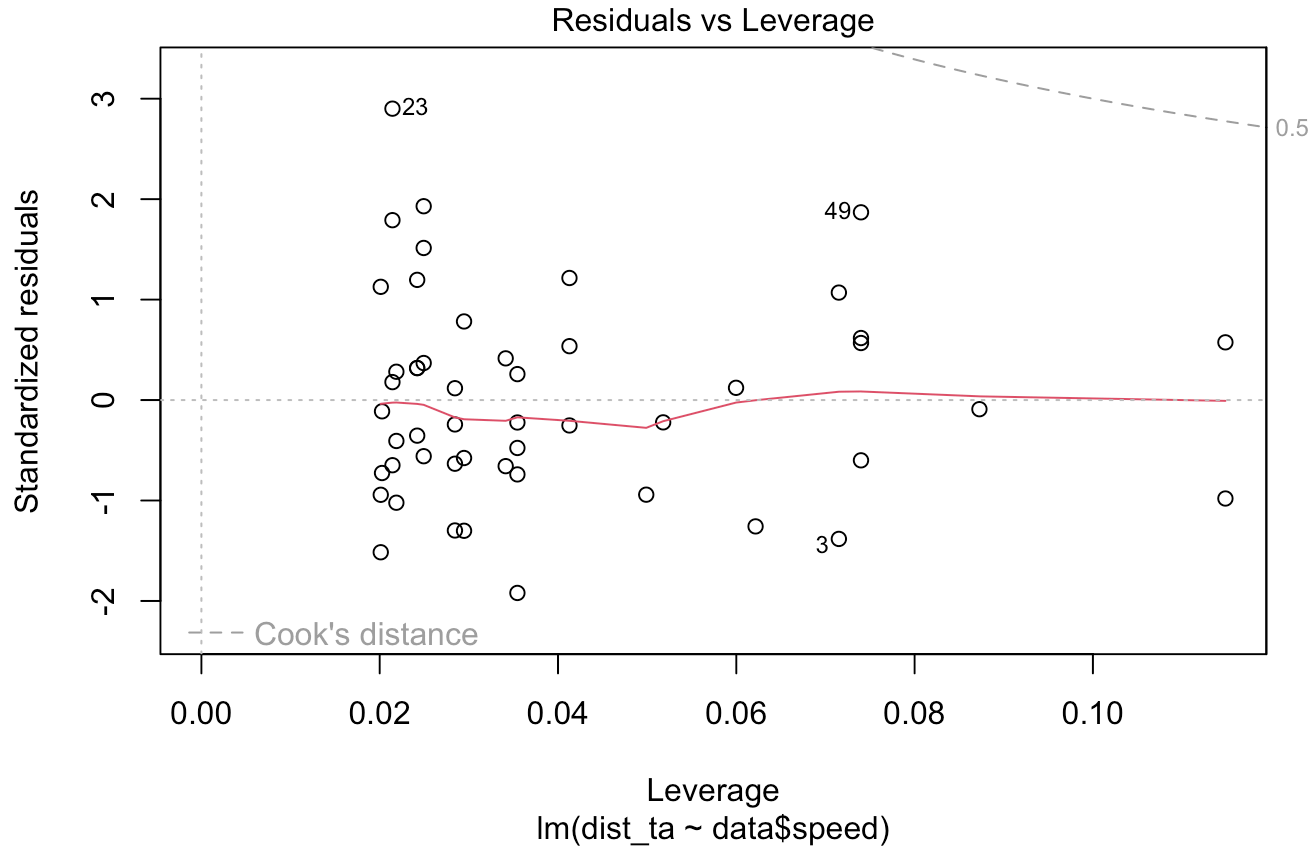
Homocedasticidad e independencia

```
plot(Modelo2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(dist_ta ~ data$speed)
23
35
39


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


Scale-Location
23
35
39
√|Standardized residuals|
0.0
0.5
1.0
1.5
3
4
5
6
7
8
9
Fitted values
lm(dist_ta ~ data$speed)


Residuals vs Leverage
23
49
3
0.5
Cook's distance
Standardized residuals
-2
-1
0
1
2
3
0.00
0.02
0.04
0.06
0.08
0.10
Leverage
lm(dist_ta ~ data$speed)

Los residuos se distribuyen de manera que no se logra distinguir algun patrón, en el QQ plot se nota que los residuos siguen una distribución casi normal, se pierde un poco de normalidad en las colas.

Prueba de independencia

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

```
dwtest(Modelo2)
```

```
##
##  Durbin-Watson test
##
## data:  Modelo2
## DW = 1.9356, p-value = 0.3527
## alternative hypothesis: true autocorrelation is greater than 0
```

```
bgtest(Modelo2)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  Modelo2
## LM test = 0.027137, df = 1, p-value = 0.8692
```

Los errores no están autocorrelacionados de acuerdo con estas pruebas de hipótesis ya que el valor p es menor a alpha.

Prueba de homocedasticidad

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

```
gqtest(Modelo2)
```

```
##
##  Goldfeld-Quandt test
##
## data:  Modelo2
## GQ = 0.85916, df1 = 23, df2 = 23, p-value = 0.6405
## alternative hypothesis: variance increases from segment 1 to 2
```

Se confirma que la varianza es constante, hay homocedasticidad.

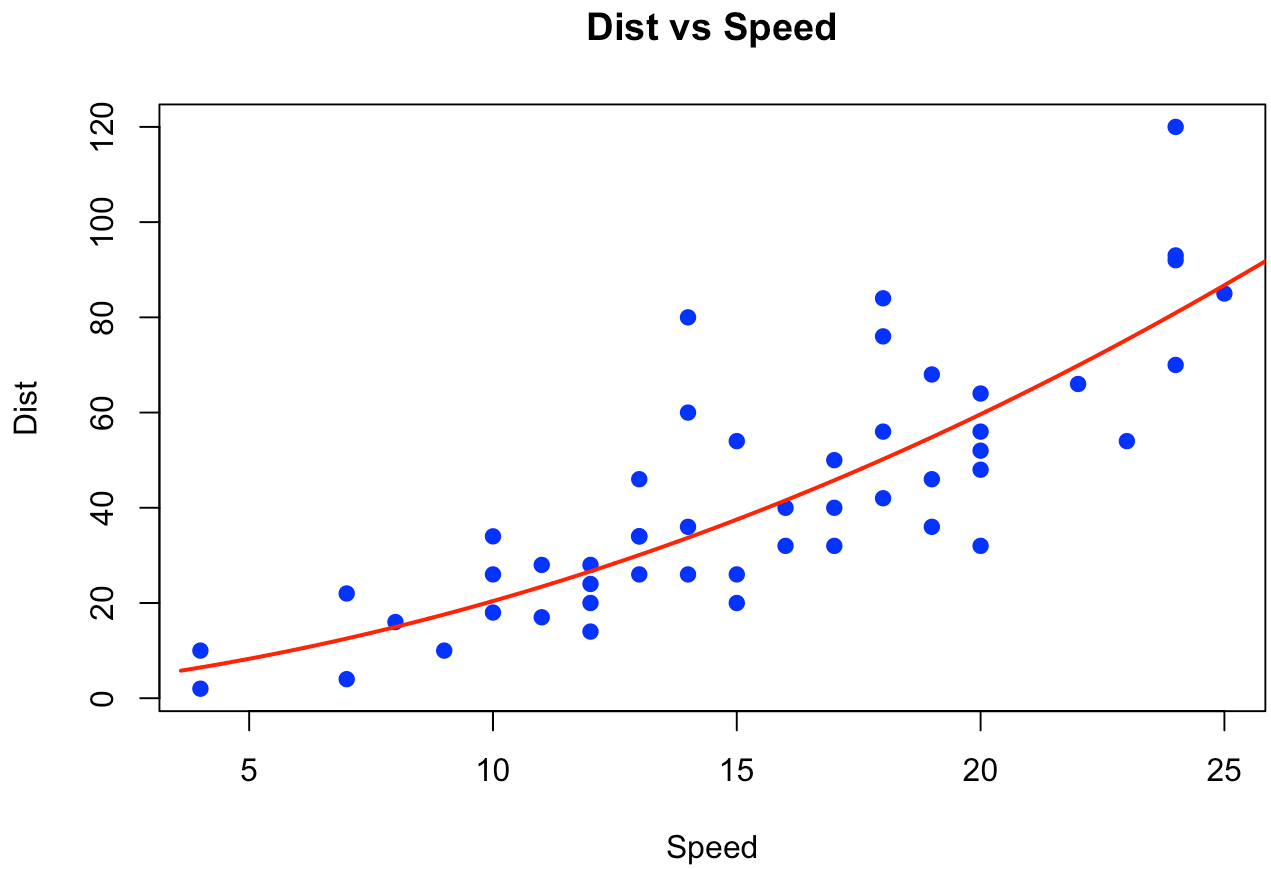
Ecuación y gráfica del modelo no lineal

La ecuación de este modelo es $\text{dist} = (0.31604 * \text{speed} + 1.46680)^2 - 1$, de esta manera se relaciona la variable original con un modelo no lineal.

```
b0 = Modelo2$coefficients[1]
b1 = Modelo2$coefficients[2]

Y = function(x){(b1*x + b0)^2 - 1}
x = seq(min(data$speed)*0.9, max(data$speed)*1.1, 0.01)

plot(data$speed, data$dist, main="Dist vs Speed",
     xlab="Speed", ylab="Dist", pch=19, col="blue")
lines(x, Y(x), col = "red", lwd = 2)
```



Se puede observar en la gráfica que este modelo no lineal se ajusta mejor a los datos que el modelo lineal. Este modelo cumple con todos los supuestos de la regresión lineal, además logra explicar la mayoría de la variación en la variable de interés, aproximadamente un 70%.

Conclusión

De los dos modelos analizados se debe escoger uno, el mejor en este caso sería el no lineal, a pesar de que el modelo lineal logra explicar una gran parte de la variación en la variable de interés, este primer modelo no cumple con todos los supuestos de la regresión lineal ya que los residuos no se distribuyen normalmente. Por otro lado, el modelo no lineal sí cumple con los supuestos y logra explicar un porcentaje más alto de la variación en la variable dependiente. Es por esto que desde el punto de vista estadístico, debe escogerse el segundo modelo.

Problemas a considerar con este modelo son similares a cualquier otro, si se tienen datos atípicos el modelo puede verse influenciado puesto que cambiaría la recta de regresión obtenida, o tal vez si el fenómeno fuera más complejo tendríamos que considerar otras posibles variables que pudieran estar involucradas de lo contrario esto

provocaría un alejamiento de los supuestos de la regresión lineal, obteniendo un modelo que podría no ser válido. Otra desventaja de los modelos no lineales es que dependiendo de la complejidad, puede volverse mas difícil el realizarse los cálculos, aumentando así el costo computacional y podría también dificultarse la interpretación de los resultados.