Actividad 2.1

Ricardo Kaleb Flores Alfonso

2024-09-20

0 Se importan las librerias

```
knitr::opts_chunk$set(echo = TRUE)
library("tidyverse")
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lmtest)
```

```
## Cargando paquete requerido: zoo
##
## Adjuntando el paquete: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

1) Analisis exploratorio

Se importan los datos

```
df <- mtcars
```

Se inspeccionan los datos

```
glimpse(df)
```

```
## Rows: 32
## Columns: 11
## $ mpg  <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2, 17.8,~
## $ cyl  <dbl> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8, 4, 4, 4, 4, 8,~
## $ disp <dbl> 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360.0, 146.7, 140.8, 16~
## $ hp   <dbl> 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180, 180, 180~
## $ drat <dbl> 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3.92, 3.92,~
## $ wt   <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3.150, 3.~
```

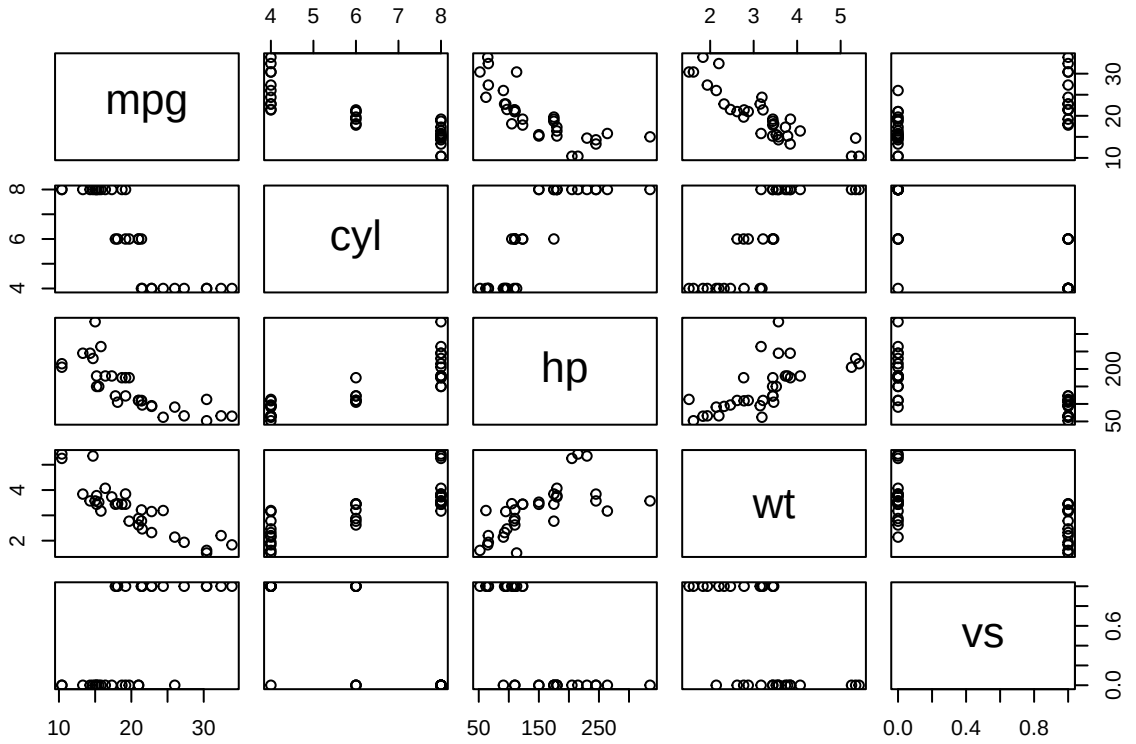
```
## $ qsec <dbl> 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15.84, 20.00, 22.90, 18~
## $ vs   <dbl> 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0,~
## $ am   <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0,~
## $ gear <dbl> 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3, 3,~
## $ carb <dbl> 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, 4, 4, 1, 2, 1, 1, 2,~
```

Se seleccionan las variables, mpg, cyl, hp, wt, vs. Para explorar su relación entre ellos

```
datos <- df %>% select(mpg,cyl,hp,wt,vs)
```

A) Gráficos de dispersión

```
plot(datos)
```



B) Matriz de varianzas y covarianzas

```
cov(datos)
```

```
##             mpg         cyl         hp         wt          vs
## mpg   36.324103  -9.1723790 -320.73206 -5.1166847   2.0171371
## cyl   -9.172379   3.1895161  101.93145  1.3673710  -0.7298387
## hp  -320.732056 101.9314516 4700.86694 44.1926613 -24.9879032
## wt    -5.116685   1.3673710   44.19266  0.9573790  -0.2736613
## vs     2.017137  -0.7298387  -24.98790 -0.2736613   0.2540323
```

C) Matriz de correlación

```
cor(datos)
```

```
##            mpg        cyl         hp         wt         vs
## mpg  1.0000000 -0.8521620 -0.7761684 -0.8676594  0.6640389
## cyl -0.8521620  1.0000000  0.8324475  0.7824958 -0.8108118
## hp  -0.7761684  0.8324475  1.0000000  0.6587479 -0.7230967
## wt  -0.8676594  0.7824958  0.6587479  1.0000000 -0.5549157
## vs   0.6640389 -0.8108118 -0.7230967 -0.5549157  1.0000000
```

D) Seleccione una variable predictora y argumente la selección de su variable independiente.

Se selecciona la variable de peso como variable predictora pues esta cuenta con un valor cercano a -1 de correlación, lo que significa que estará relacionada de manera inversamente proporcional. Y una covarianza baja lo que hace que los valores no cambien tanto dependiendo de que valor tenga cada uno.

```
y <- datos %>% select(mpg) %>% .$mpg
x <- datos %>% select(wt)
x <- cbind(1, x$wt)
```

```
y
```

```
##  [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4
## [16] 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26.0 30.4 15.8 19.7
## [31] 15.0 21.4
```

2) Metodo de minimos cuadrados

Modelo de regresión lineal

$$\beta = (X^T X)^{-1} X^T Y$$

Se resuelve para el modelo de regresión lineal

```
beta <- solve(t(x) %*% x) %*% t(x) %*% y
```

```
beta
```

```
##           [,1]
## [1,] 37.285126
## [2,] -5.344472
```

Modelo de regresión obtenido

$$y = 37.28 - 5.34x$$

3) Regresión lineal en R

```
model <- lm(y ~ x[,2])
intercept <- model$coefficients[1]
values <- model$coefficients[2]
model
```

```
##
## Call:
## lm(formula = y ~ x[, 2])
##
```

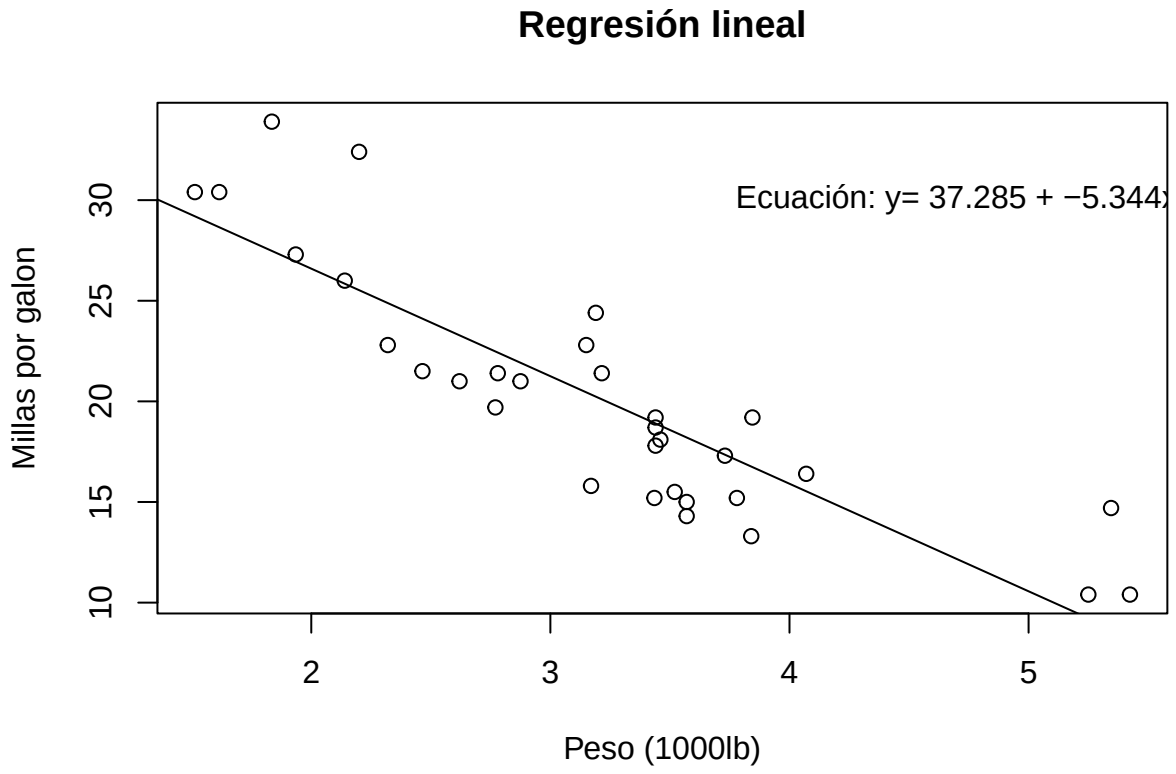
```
## Coefficients:
## (Intercept)       x[, 2]
##      37.285       -5.344
```

Modelo obtenido

$$y = 37.28 - 5.34x$$

4) Representación gráfica

```
plot(x[,2], y ,ylab = "Millas por galon" ,xlab = "Peso (1000lb)")
abline(model)
title("Regresión lineal")
text(x=4.7, y=30, "Ecuación: y= 37.285 + -5.344x")
```



5) Coeficiente de determinación

```
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x[, 2])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
```

```
## x[, 2]       -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

El modelo se ajusta al 74.46% de los datos

6) Validación del modelo

6.1 Significancia de los coeficientes de regresión

T Test

- H_0 := El coeficiente es igual a 0.
- H_A := El coeficiente no es igual a 0.

Significancia de los coeficientes de regresión:

```
summary(model)
```

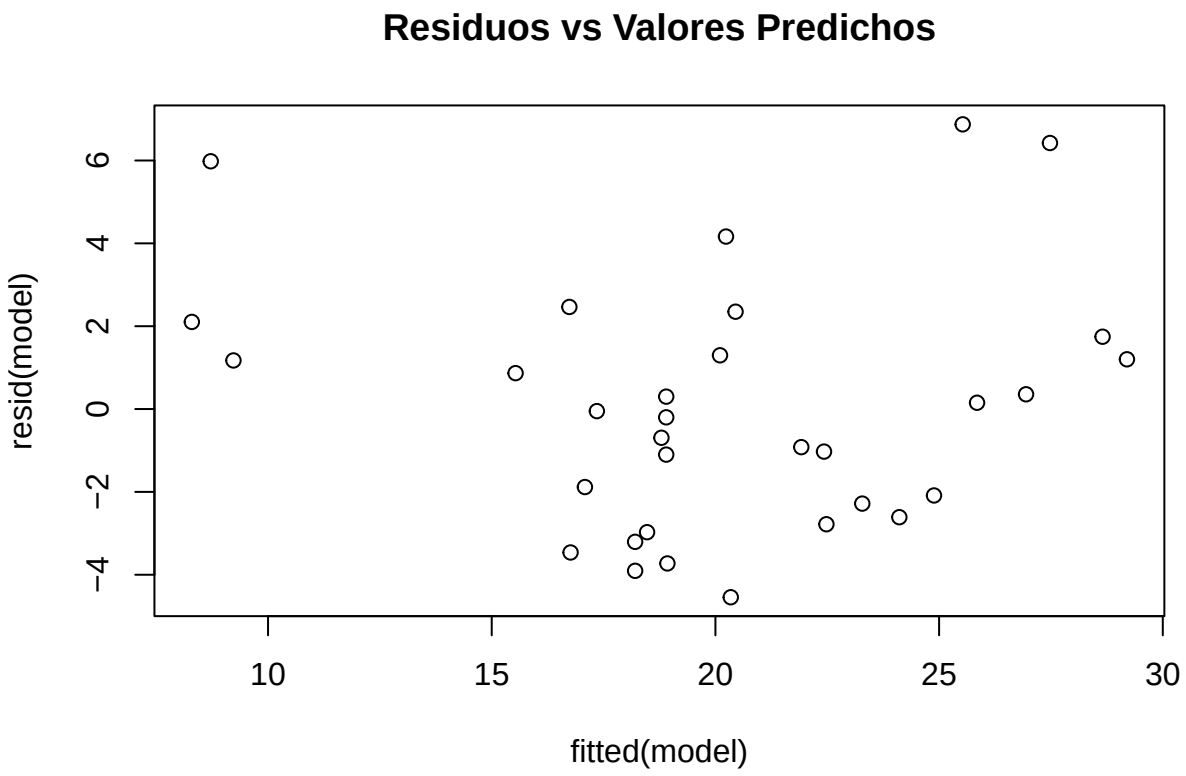
```
##
## Call:
## lm(formula = y ~ x[, 2])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## x[, 2]       -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

El resumen muestra los resultados de la prueba de T regresó valores de 19.858 y -9.559, por lo que los coeficientes del modelo sugerido son significativos. Dado que $p < 0.05$, se rechaza la hipotesis nula, lo que demuestra que los coeficientes son diferentes a 0, asi se demuestra que los coeficientes obtenidos son significativos.

6.2 Linealidad

- H_0 := La relación entre la variable independiente y dependiente es lineal
- H_A := La relación es no lineal

```
plot(fitted(model), resid(model), main = "Residuos vs Valores Predichos")
```




```
resettest(model)
```

```
##
##  RESET test
##
## data:  model
## RESET = 5.1315, df1 = 2, df2 = 28, p-value = 0.01263
```

Se obtiene un p-value para la prueba de linealidad de 0.01263, por lo que se rechaza la hipotesis nula, de esta manera se demuestra que existe otro tipo de relación no lineal en el modelo.

6.3) Media de cero de los residuos

T Test

- H_0 := La media de los errores es igual a 0.
- H_A := La media de los errores no es igual a 0.

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

```
##
##  One Sample t-test
##
## data:  model$residuals
## t = 1.3586e-16, df = 31, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1.0803  1.0803
```

```
## sample estimates:
##    mean of x
## 7.196392e-17
```

Se obtuvo que el resultado de la prueba da como resultado que p = 1. Por lo que no es posible rechazar la hipotesis nula. Por lo tanto se tiene un error promedio de 0 en el modelo.

6.4) Normalidad de residuos

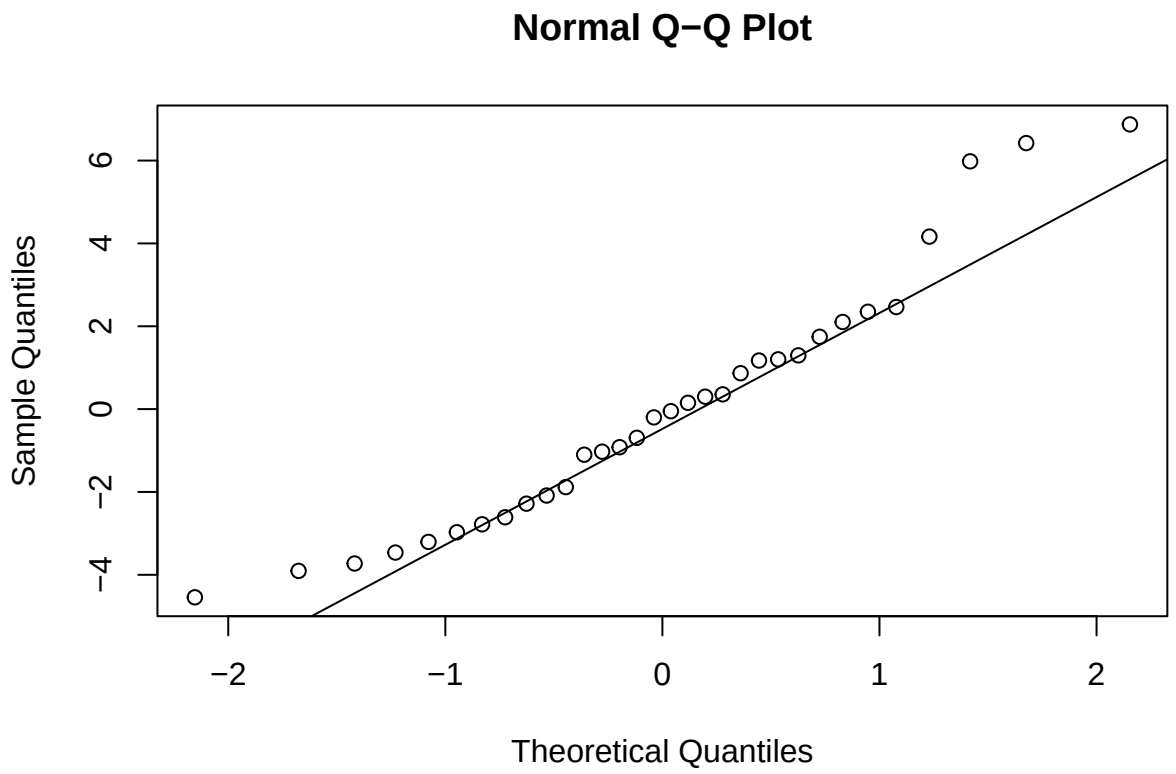
T Test

- H_0 := Los residuos tienen una distribución normal
- H_A := Los residuos no tienen una distribución normal

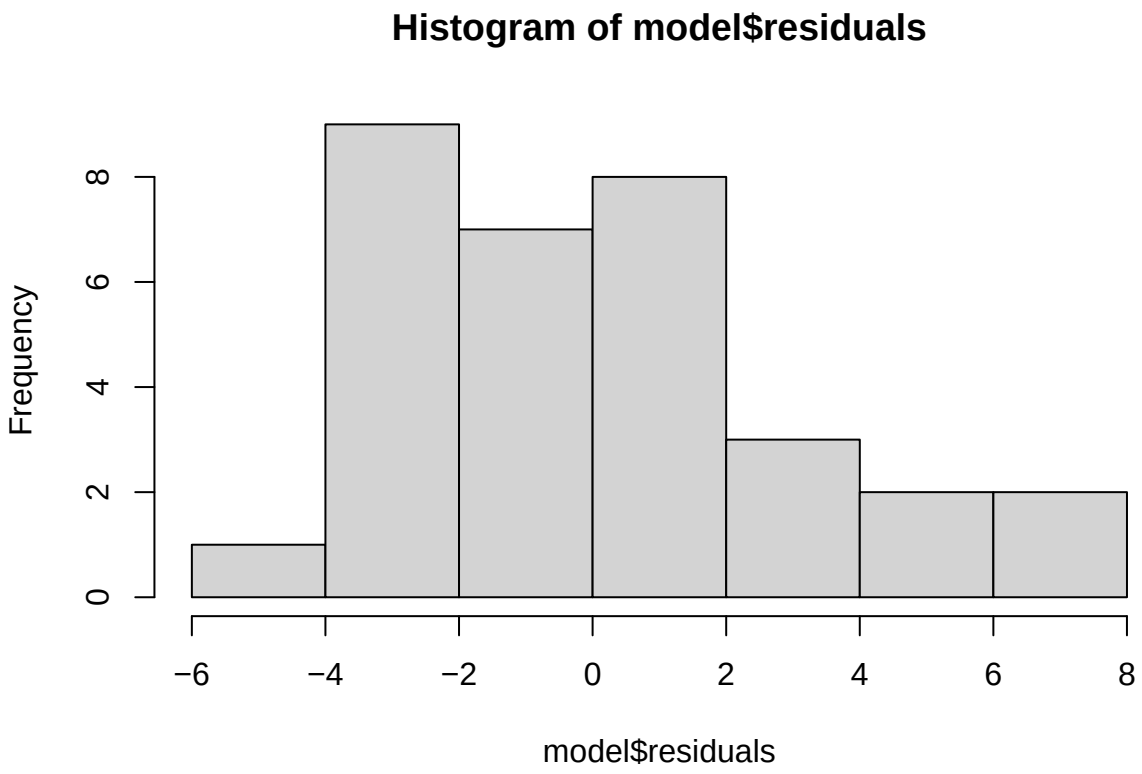
```
shapiro.test(model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.94508, p-value = 0.1044
```

```
qqnorm(model$residuals)
qqline(model$residuals)
```




```
hist(model$residuals)
```



Se obtiene un valor de p=0.1044, por lo tanto no existe evidencia para rechazar la hipotesis nula. De esta manera se sabe que los residuos no se desvian de una distribución normal.

6.5) Breusch-Pagan

- H_0 := Los datos tienen homocedasticidad.
- H_A := Los datos no tienen homocedasticidad.

```
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 0.040438, df = 1, p-value = 0.8406
```

El test de Breusch-Pagan produce un valor de p de 0.8406, dado que es mayor que 0.05. Se falla en rechazar la hipotesis nula. Esto demuestra que el modelo tiene varianza constante.

Test de white

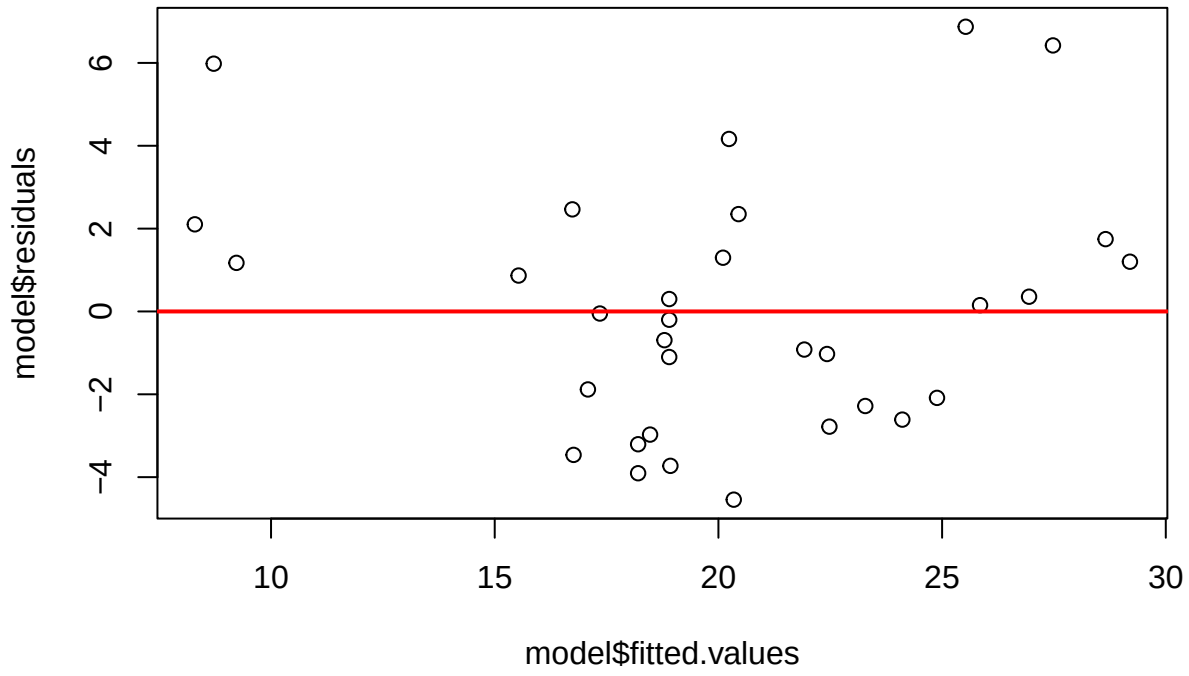
```
# Subset the relevant variables: mpg (y) and wt (x)
datos_subset <- datos %>% select(mpg, wt)

model <- lm(mpg ~ wt, data = datos_subset)

bptest(model, varformula = ~ wt + I(wt^2), data = datos_subset)
```

```
##
```

```
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 1.3663, df = 2, p-value = 0.505
```

El test de White produce un valor de p de 0.505, dado que es mayor que 0.05. Se falla en rechazar la hipotesis nula. Esto demuestra que el modelo tiene varianza constante.

```
plot(model$fitted.values, model$residuals)
abline(h=0, col = "red", lwd = 2)
```



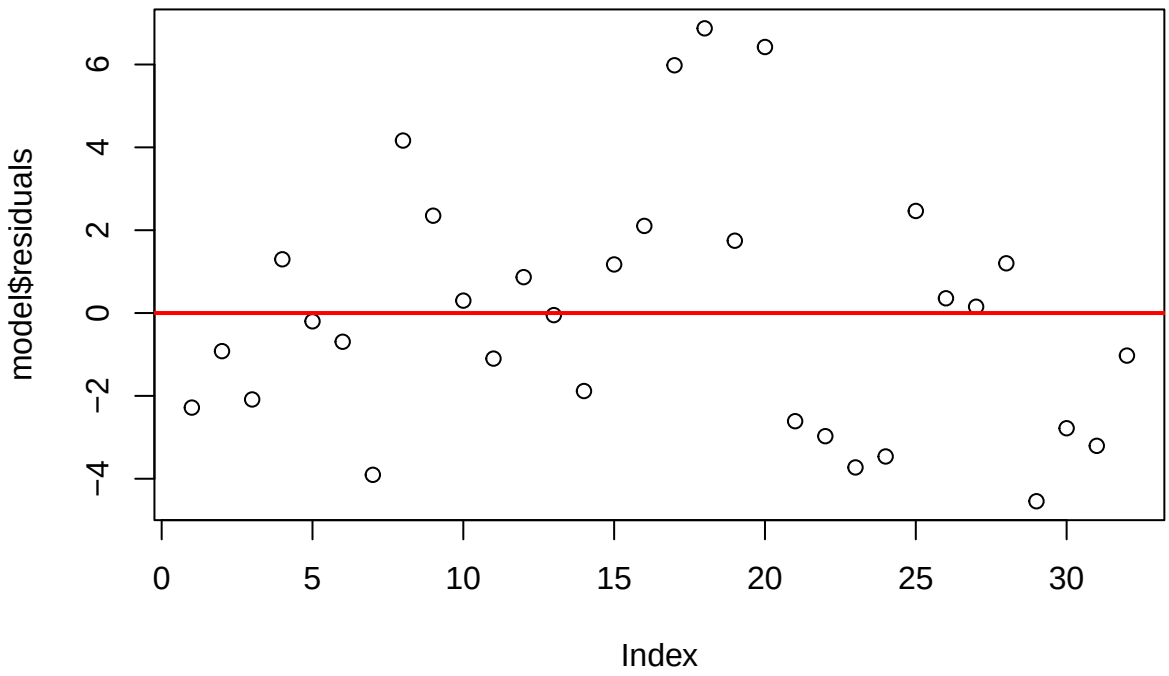
6.6) Independencia Test de Durbin Watson

- H_0 := No existe autocorrelación en los datos
- H_A := Existe autocorrelacion en los datos.

```
dwtest(model)
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 1.2517, p-value = 0.0102
## alternative hypothesis: true autocorrelation is greater than 0
```

```
plot(model$residuals)
abline(h=0, col = "red", lwd = 2)
```



Dado que el valor obtenido por el test de durbin watson da un valor de 0.0102 y esto es menor que 0.05, se rechaza la hipotesis nula. Esto significa que existe una correlación positiva en los residuos, por lo que estos no son independientes. La estadistica de durbin watson cercana a 1.25 sugiere esta misma correlación positiva.

Conclusiones

El análisis realizado demuestra una relación lineal negativa entre el consumo de combustible de los automóviles y su peso. Esto indica que, a medida que aumenta el peso, el consumo de combustible aumenta.

A pesar de la solidez de los coeficientes encontrados, el modelo puede ser mejorado. Dado que se rechaza la hipotesis nula de linealidad, significa que puede ser que la relación presente no sea completamente lineal y el modelo podria mejorar. Ademas existe autocorrelación positiva en los residuos, lo cual muestra que estos no son independientes, afectando asi la presición del modelo.

Por otro lado la media de los residuos es cero, siguen una distribución normal y se mantiene la homocedasticidad.

Esto muestra que el modelo de regresión obtenido es estadísticamente significativo y captura de manera correcta la relación entre el peso y las millas por galon. Es importante abordar la no linealidad observada para futuros modelo. De manera que se pueda mejorar su precisión.