Actividad Integradora 2

Oskar Arturo Gamboa Reyes

2024-09-06

Exploración de la base de datos

```
M=read.csv("precios_autos.csv")
```

Medidas estadísticas

```
print("Wheel base")
```

```
## [1] "Wheel base"
```

```
summary(M$wheelbase)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   86.60   94.50   97.00   98.76  102.40  120.90
```

```
print("Fuel Type")
```

```
## [1] "Fuel Type"
```

```
print(freq_fueltype <- table(M$fueltype))
```

```
## 
## diesel    gas 
##     20    185
```

```
print("Horsepower")
```

```
## [1] "Horsepower"
```

```
summary(M$horsepower)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    48.0    70.0    95.0   104.1   116.0   288.0
```

```
M$fueltype_dummy <- ifelse(M$fueltype == "gas", 1, 0)
M_subset <- M[c("wheelbase", "horsepower", "fueltype_dummy", "price")]
print(cor(M_subset, use = "complete.obs"))
```

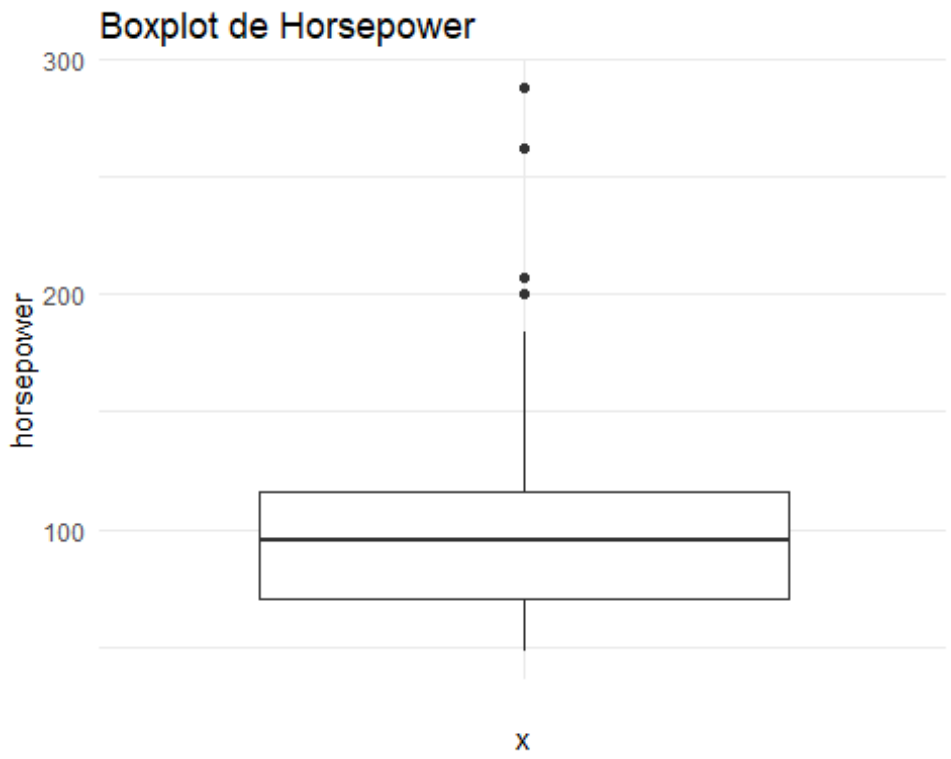
```
##                 wheelbase horsepower fueltype_dummy      price
## wheelbase       1.0000000  0.3532945     -0.3083459  0.5778156
## horsepower      0.3532945  1.0000000      0.1639262  0.8081388
## fueltype_dummy -0.3083459  0.1639262      1.0000000 -0.1056795
## price           0.5778156  0.8081388     -0.1056795  1.0000000
```

Diagramas

Boxplot (Cuantitativas)

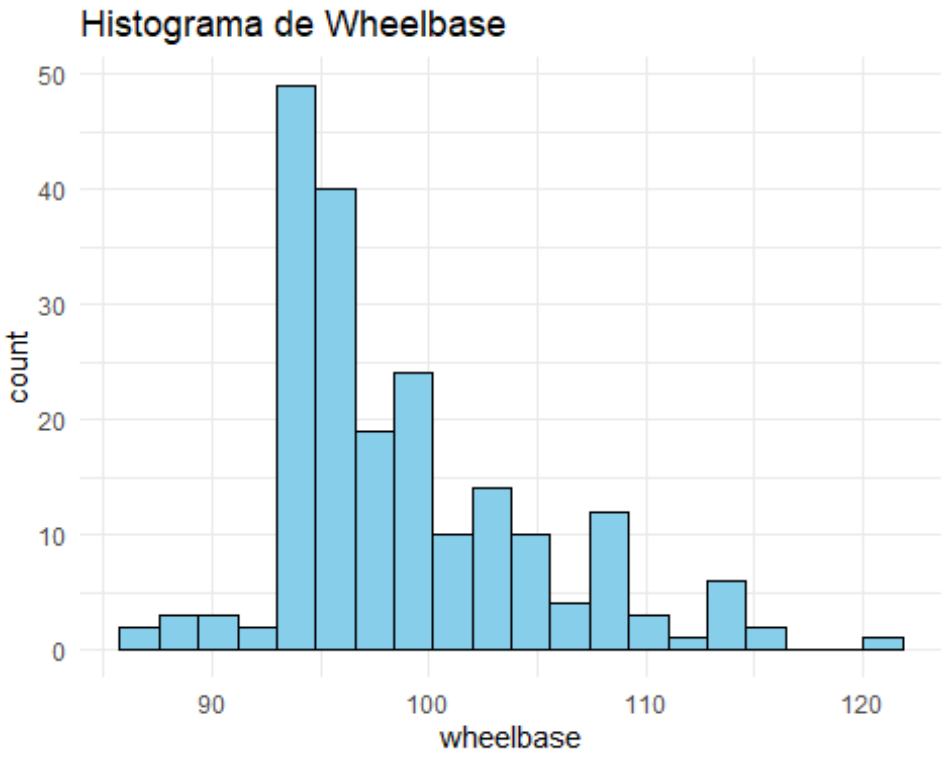
```
library(ggplot2)
library(GGally)

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

ggplot(M, aes(x = "", y = wheelbase)) +
  geom_boxplot() +
  labs(title = "Boxplot de Wheelbase") +
  theme_minimal()
```


```
ggplot(M, aes(x = "", y = horsepower)) +
  geom_boxplot() +
  labs(title = "Boxplot de Horsepower") +
  theme_minimal()
```

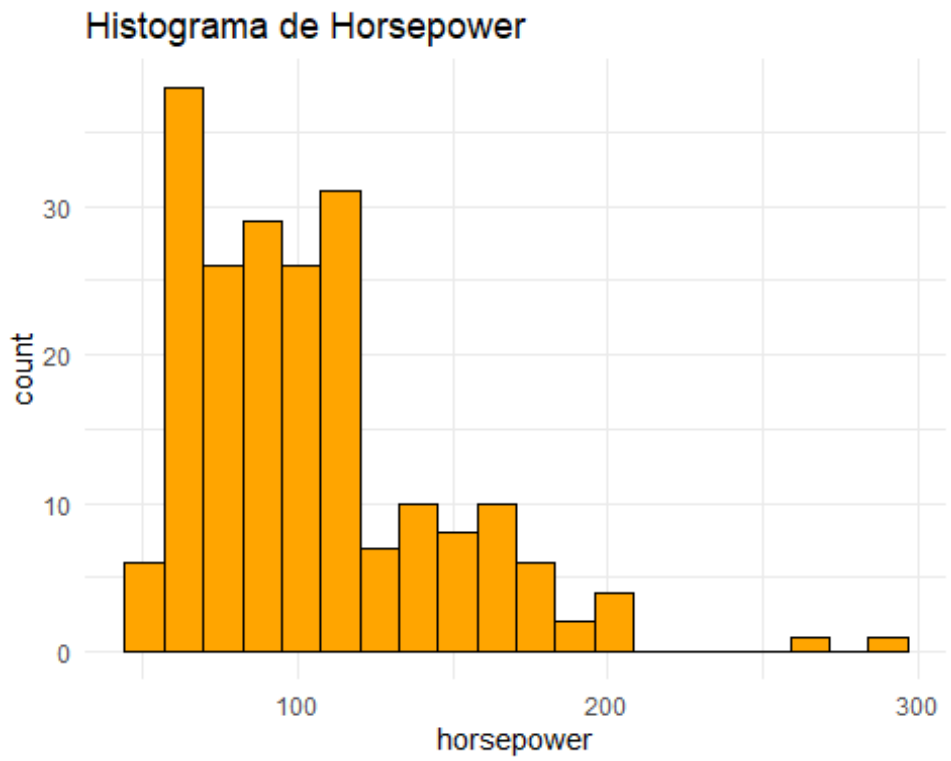


Histograma (Cuantitativas)

```
ggplot(M, aes(x = wheelbase)) +
  geom_histogram(bins = 20, fill = "skyblue", color = "black") +
  labs(title = "Histograma de Wheelbase") +
  theme_minimal()
```

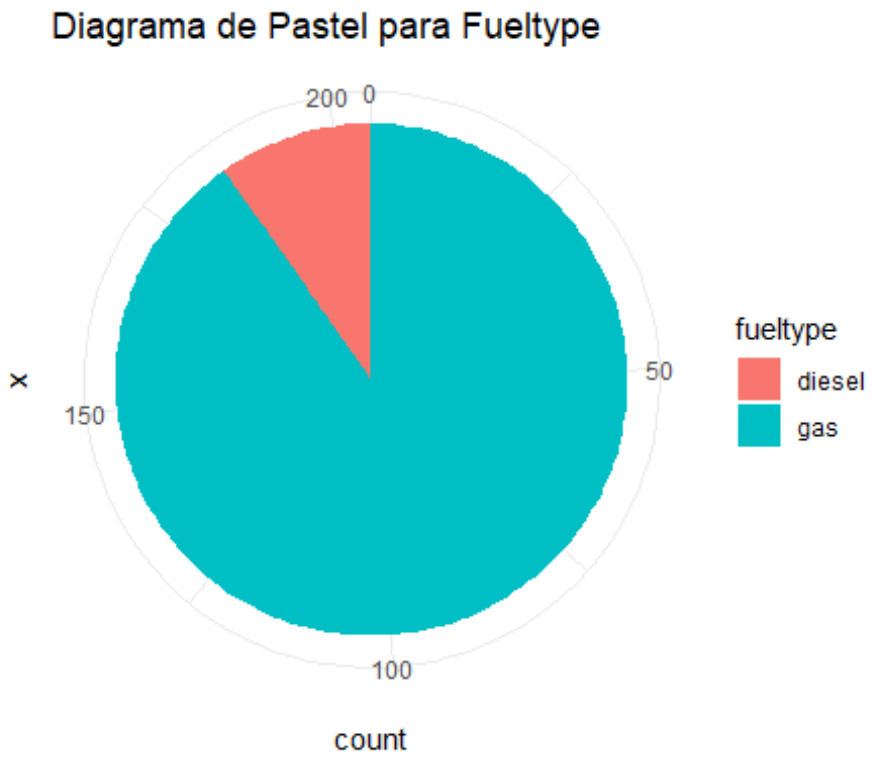



```r
ggplot(M, aes(x = horsepower)) +
  geom_histogram(bins = 20, fill = "orange", color = "black") +
  labs(title = "Histograma de Horsepower") +
  theme_minimal()
```



Distribución de datos (Cualitativas)

```
M_fueltype <- data.frame(table(M$fueltype))
colnames(M_fueltype) <- c("fueltype", "count")
ggplot(M_fueltype, aes(x = "", y = count, fill = fueltype)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y") +
  labs(title = "Diagrama de Pastel para Fueltype") +
  theme_minimal()
```



Modelos de predicción

Sin interacción

```
Modelo1 = lm(price~horsepower, M)
print(Modelo1)
```

```
##
## Call:
## lm(formula = price ~ horsepower, data = M)
##
## Coefficients:
## (Intercept)   horsepower
##     -3721.8        163.3
```

```
summary(Modelo1)
```

```
##
## Call:
## lm(formula = price ~ horsepower, data = M)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11897.5  -2350.4   -711.1   1644.6  19081.4
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3721.761    929.849  -4.003 8.78e-05 ***
## horsepower    163.263      8.351  19.549  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4717 on 203 degrees of freedom
## Multiple R-squared:  0.6531, Adjusted R-squared:  0.6514
## F-statistic: 382.2 on 1 and 203 DF,  p-value: < 2.2e-16
```

Hipotesis de prueba del modelo

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

```
summary_model1 <- summary(Modelo1)
p_value_model1 <- summary_model1$fstatistic[1]
cat("P-valor del modelo completo:", pf(p_value_model1,
summary_model1$fstatistic[2], summary_model1$fstatistic[3], lower.tail =
FALSE), "\n")

## P-valor del modelo completo: 1.483437e-48

#Valor frontera
df_residual <- df.residual(Modelo1)
alpha <- 0.04
valor_frontera <- qt(1 - alpha / 2, df_residual)
cat("Valor frontera para alfa =", alpha, ":", valor_frontera, "\n")

## Valor frontera para alfa = 0.04 : 2.067029
```

El valor es menor a alpha 0.04, por lo que nuestro modelo es significativo.

Hipotesis para β_i

Para cada i

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

```
p_values_coef1 <- summary_model1$coefficients[, 4]
significant_coefs1 <- p_values_coef1 < 0.04
cat("Coeficientes significativos a nivel de alfa 0.04:", significant_coefs1,
"\n")

## Coeficientes significativos a nivel de alfa 0.04: TRUE TRUE
```

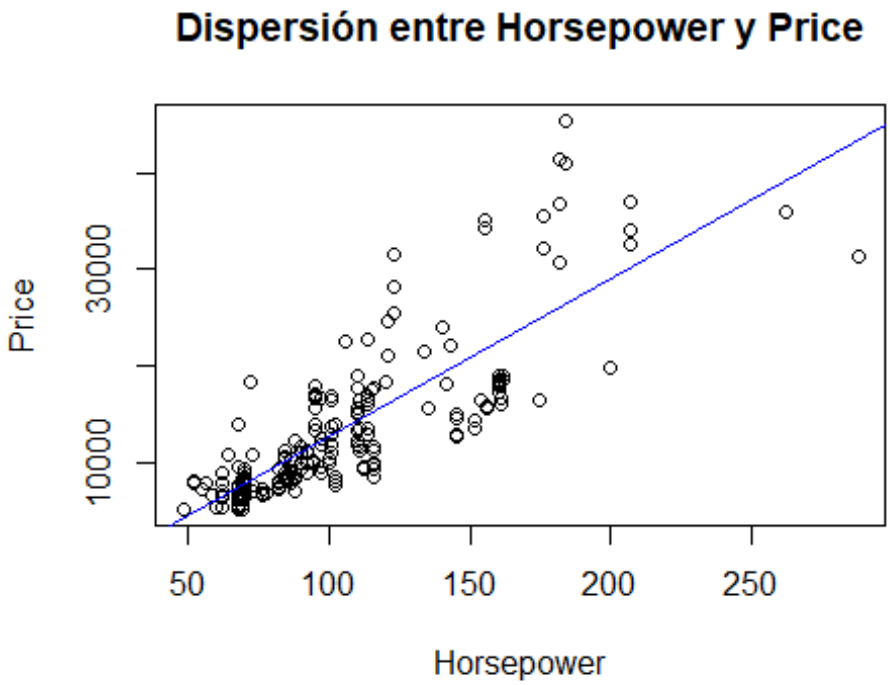
Comprobamos la hipótesis, por lo que cada coeficiente es significativo en nuestro modelo.

Porcentaje de variación

65%

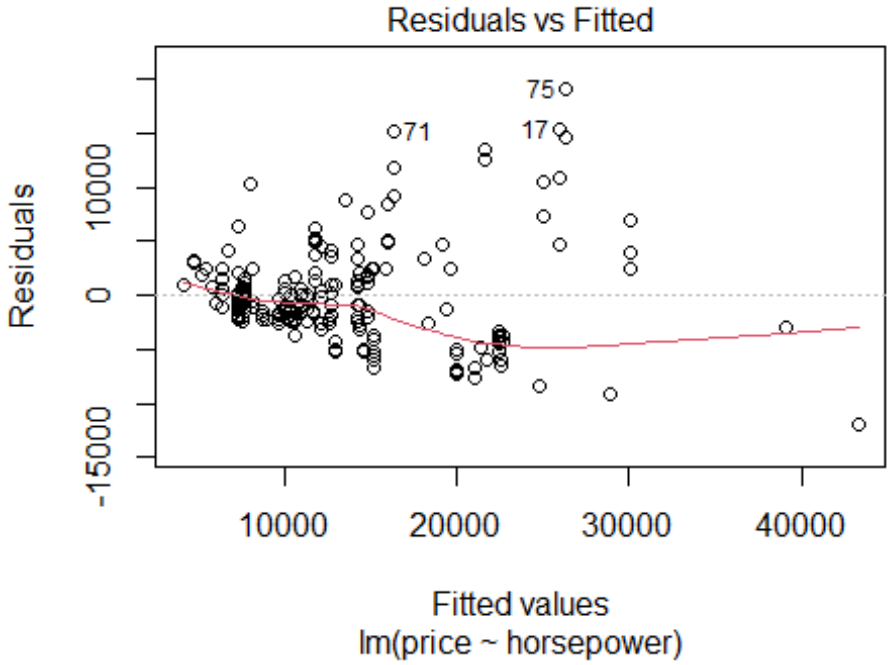
Diagrama de dispersión

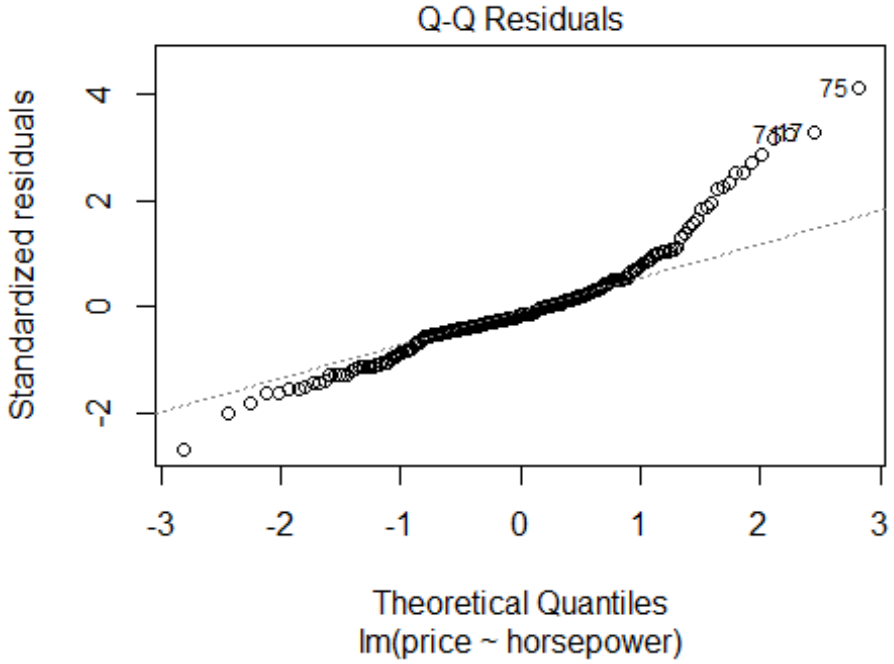
```r
plot(M$horsepower, M$price, main = "Dispersión entre Horsepower y Price",
xlab = "Horsepower", ylab = "Price")
abline(lm(price ~ horsepower, data = M),col="blue")
```

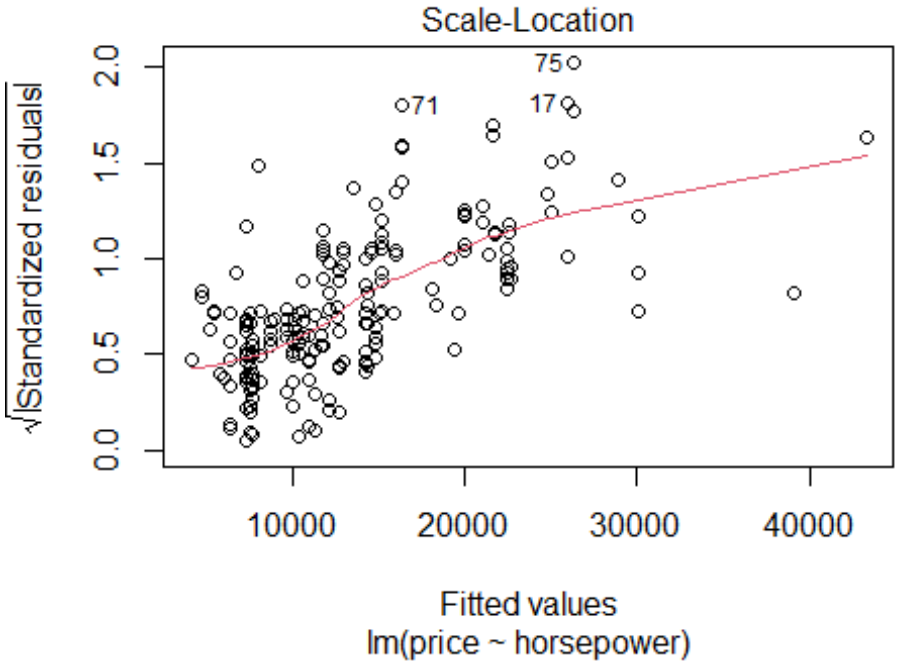


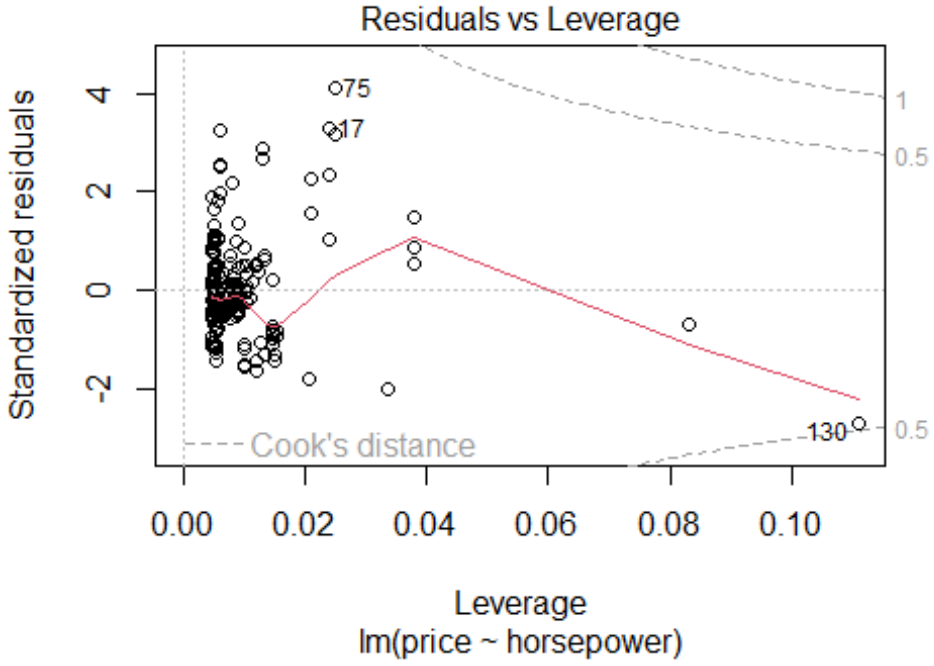
Residuos

```r
plot(Modelo1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(price ~ horsepower)
75
71
17
10000
0
-15000
10000
20000
30000
40000


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(price ~ horsepower)
75
71
17
4
2
0
-2
-3
-2
-1
0
1
2
3





Homocedasticidad

```
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

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

bptest(Modelo1)

##
##  studentized Breusch-Pagan test
##
## data:  Modelo1
## BP = 54.573, df = 1, p-value = 1.497e-13

gqtest(Modelo1)

##
##  Goldfeld-Quandt test
##
## data:  Modelo1
## GQ = 0.42709, df1 = 101, df2 = 100, p-value = 1
## alternative hypothesis: variance increases from segment 1 to 2
```

Independencia

```
dwtest(Modelo1)

##
##  Durbin-Watson test
##
## data:  Modelo1
## DW = 0.79229, p-value < 2.2e-16
## 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 = 81.074, df = 1, p-value < 2.2e-16
```

El modelo usando horsepower tiene una buena significancia en predecir el precio del automovil, sin embargo los autos que tienen mayores caballos de fuerza suelen incrementarse exponencialmente y no seguir la linea de predicción. Podemos ver por el porcentaje de significancia que todavia hay datos en los errores que se pueden explicar con variables. Ademas no tiene ni homocedasticidad ni independencia ya que no es constante y los valores no estan correlacionados. Voy a incluir para el siguiente modelo la siguiente variable que este correlacionada al precio.

Con Interacción

```
Modelo2 = lm(price~horsepower*wheelbase, M)
print(Modelo2)
```

```
##
## Call:
## lm(formula = price ~ horsepower * wheelbase, data = M)
##
## Coefficients:
##          (Intercept)            horsepower             wheelbase
##           -17059.574               -89.721               155.900
## horsepower:wheelbase
##                2.342
```

```
summary(Modelo2)
```

```
##
## Call:
## lm(formula = price ~ horsepower * wheelbase, data = M)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8847  -2050   -177   1350  15889
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -17059.574  14377.287  -1.187   0.2368
## horsepower              -89.721    111.777  -0.803   0.4231
## wheelbase               155.900    148.256   1.052   0.2943
## horsepower:wheelbase      2.342      1.140   2.055   0.0412 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3977 on 201 degrees of freedom
## Multiple R-squared:  0.7558, Adjusted R-squared:  0.7522
## F-statistic: 207.4 on 3 and 201 DF,  p-value: < 2.2e-16
```

Hipotesis de prueba del modelo

- $H_0: \beta_1 = \beta_2 = \beta_1 * \beta_2 = 0$
- $H_1: \beta_i \neq 0$

```
summary_model2 <- summary(Modelo2)
p_value_model2 <- summary_model2$fstatistic[1]
cat("P-valor del modelo completo:", pf(p_value_model2,
summary_model2$fstatistic[2], summary_model2$fstatistic[3], lower.tail =
FALSE), "\n")
```

```
## P-valor del modelo completo: 2.855049e-61
```

```r
#Valor frontera
df_residual <- df.residual(Modelo2)
alpha <- 0.04
valor_frontera <- qt(1 - alpha / 2, df_residual)
cat("Valor frontera para alfa =", alpha, ":", valor_frontera, "\n")

## Valor frontera para alfa = 0.04 : 2.067162
```

El valor es menor a alpha 0.04, por lo que nuestro modelo es significativo.

Hipotesis para β_i

Para cada i

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

```r
p_values_coef2 <- summary_model2$coefficients[, 4]
significant_coefs2 <- p_values_coef2 < 0.04
cat("Coeficientes significativos a nivel de alfa 0.04:", significant_coefs1,
"\n")

## Coeficientes significativos a nivel de alfa 0.04: TRUE TRUE
```

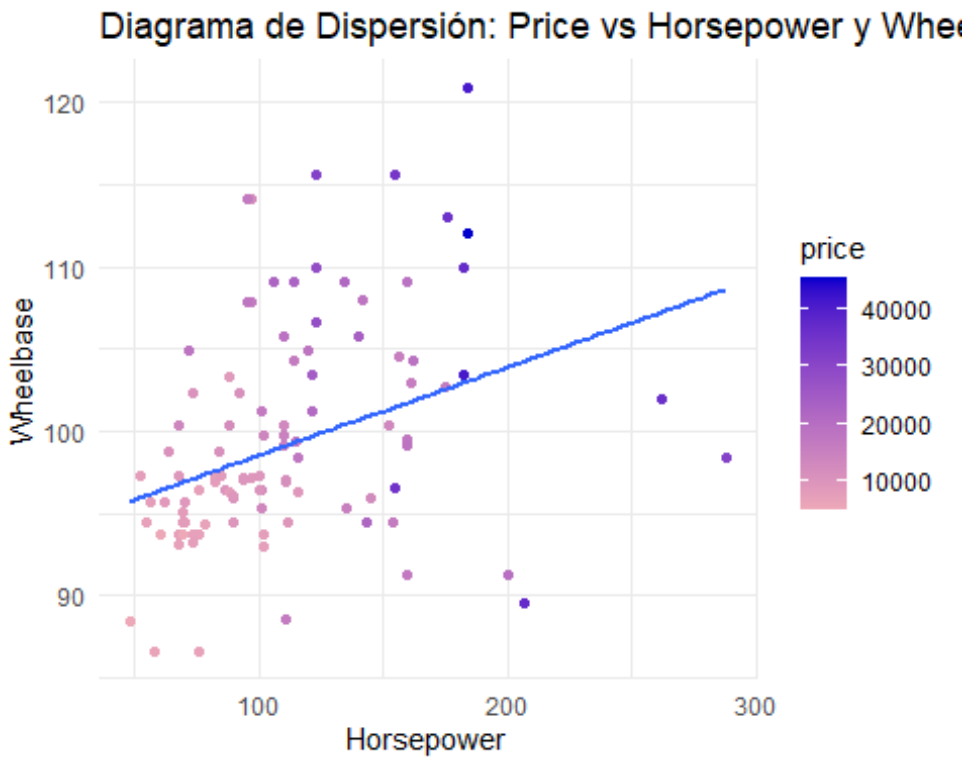
Comprobamos la hipótesis, por lo que cada coeficiente es significativo en nuestro modelo.

Porcentaje de variación

75%

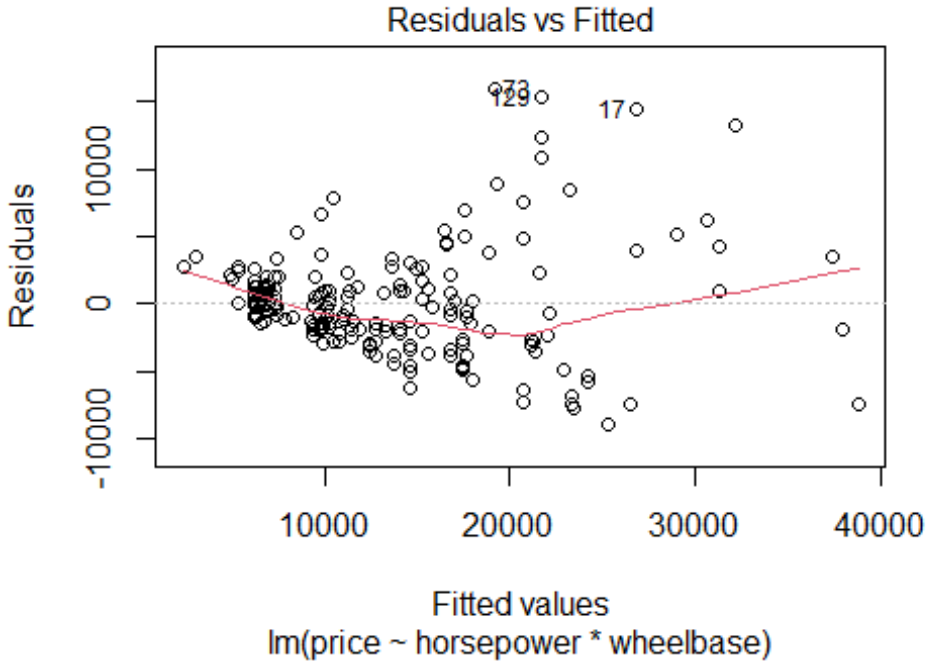
Diagrama de dispersión

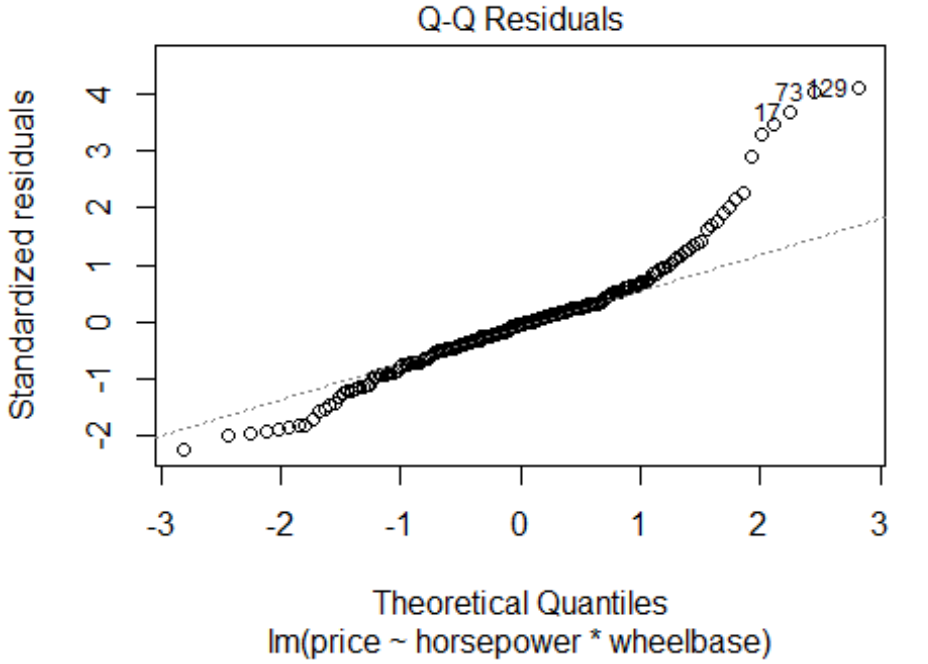
```r
ggplot(M, aes(x = horsepower, y = wheelbase, color = price)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, aes(color = NULL), se = FALSE)
+
  scale_color_gradient(low = "pink2", high = "blue3") +
  labs(title = "Diagrama de Dispersión: Price vs Horsepower y Wheelbase",
       x = "Horsepower",
       y = "Wheelbase") +
  theme_minimal()
```

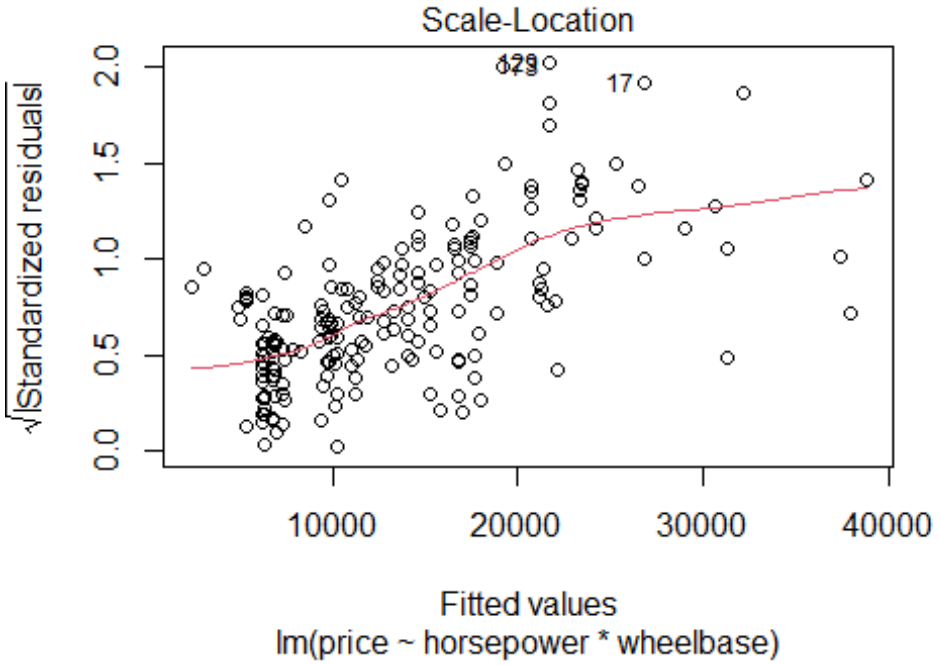


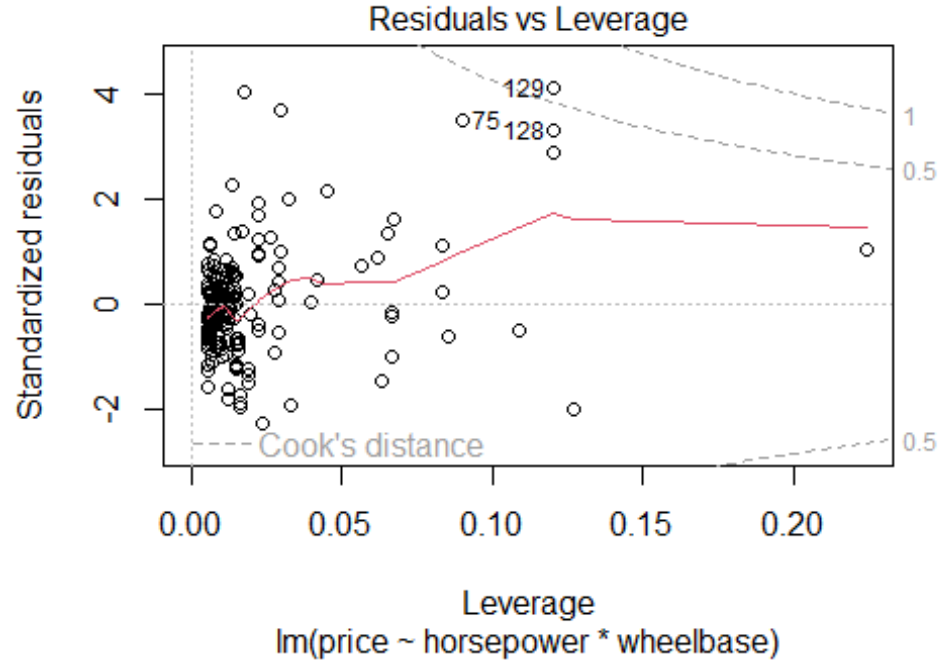
Residuos

```
plot(Modelo2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(price ~ horsepower * wheelbase)
73
129
17
10000
0
-10000
10000
20000
30000
40000


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(price ~ horsepower * wheelbase)
73
129
17
-2
-1
0
1
2
3
4
-3
-2
-1
0
1
2
3





Homocedasticidad

```
library(lmtest)
bptest(Modelo2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Modelo2
## BP = 60.863, df = 3, p-value = 3.845e-13
```

```
gqtest(Modelo2)
```

```
##
##  Goldfeld-Quandt test
##
## data:  Modelo2
## GQ = 0.5584, df1 = 99, df2 = 98, p-value = 0.9979
## alternative hypothesis: variance increases from segment 1 to 2
```

Independencia

```
dwtest(Modelo2)
```

```
##
##  Durbin-Watson test
##
## data:  Modelo2
## DW = 1.0509, p-value = 1.575e-12
## 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 = 52.006, df = 1, p-value = 5.534e-13
```

Podemos ver que los modelos no pasan las pruebas de normalidad en los errores, sin embargo tienen un buen nivel de predicción, el último modelo explica el 75% del comportamiento del precio, ademas a una significancia de 0.04 todas nuestras variables logran pasar esta prueba. Seguimos sin obtener ni homocedasticidad ni indepencia, pero creo que esto se debe a que hay demasiadas variables que determinan el precio de un automovil y no solamente se puede resolver con 2 de estas variables, tendriamos que hacer un analisis más profundo para obtener un mejor modelo.

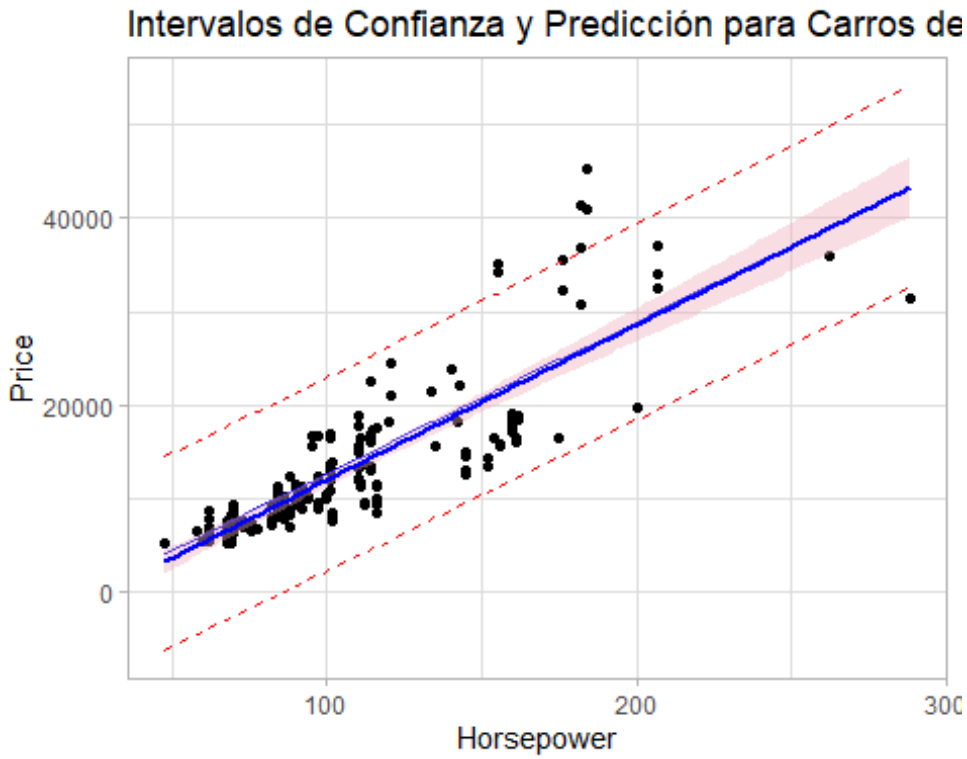
Intervalos de predicción y confianza

```
Ip <- predict(object = Modelo1, interval = "prediction", level = 0.97)
```

```
## Warning in predict.lm(object = Modelo1, interval = "prediction", level =
0.97): predictions on current data refer to _future_ responses
```

```
datos1 <- cbind(M, Ip)
```

```r
gas <- subset(datos1, datos1$fueltype == "gas")


library(ggplot2)
ggplot(gas, aes(x = horsepower, y = price)) +
geom_point() +
geom_line(aes(y = fit), color = "blue") +
geom_line(aes(y = lwr), color = "red", linetype = "dashed") +
geom_line(aes(y = upr), color = "red", linetype = "dashed") +
geom_smooth(method = lm, formula = y ~ x, se = TRUE, level = 0.97, col =
"blue", fill = "pink2") +
theme_light() +
labs(title = "Intervalos de Confianza y Predicción para Carros de Gasolina",
x = "Horsepower", y = "Price")
```



Todavia existen datos que no se acoplan a nuestro modelo, sin embargo esto se puede explicar ya que los autos con mayor caballos de poder se suelen vender como un carro de lujo y el precio aumenta exponencialmente, lo que hace que estos datos no se encuentren en nuestro rango de predicción, además podemos ver que los que tienen menor horsepower tienen un intervalo de confianza menor ya que todos se apegan al modelo establecido.

Conclusión

Creo que es un buen grupo de variables, sin embargo usaria otro tipo de modelos para poder determinar el precoi de una manera más acertada, ya que creo que la linealidad no ayuda mucho a este problema en específico.