act_int_2_A01742161

Rogelio Lizárraga

2024-09-06

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

```
##   symboling                  CarName fueltype     carbody drivewheel
## 1         3       alfa-romero giulia      gas convertible        rwd
## 2         3      alfa-romero stelvio      gas convertible        rwd
## 3         1 alfa-romero Quadrifoglio      gas   hatchback        rwd
## 4         2              audi 100 ls      gas       sedan        fwd
## 5         2               audi 100ls      gas       sedan        4wd
## 6         2                 audi fox      gas       sedan        fwd
##   enginelocation wheelbase carlength carwidth carheight curbweight enginetype
## 1          front      88.6     168.8     64.1      48.8       2548       dohc
## 2          front      88.6     168.8     64.1      48.8       2548       dohc
## 3          front      94.5     171.2     65.5      52.4       2823       ohcv
## 4          front      99.8     176.6     66.2      54.3       2337        ohc
## 5          front      99.4     176.6     66.4      54.3       2824        ohc
## 6          front      99.8     177.3     66.3      53.1       2507        ohc
##   cylindernumber enginesize stroke compressionratio horsepower peakrpm citympg
## 1           four        130   2.68              9.0        111    5000      21
## 2           four        130   2.68              9.0        111    5000      21
## 3            six        152   3.47              9.0        154    5000      19
## 4           four        109   3.40             10.0        102    5500      24
## 5           five        136   3.40              8.0        115    5500      18
## 6           five        136   3.40              8.5        110    5500      19
##   highwaympg price
## 1         27 13495
## 2         27 16500
## 3         26 16500
## 4         30 13950
## 5         22 17450
## 6         25 15250
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
M2 <- M %>% select(wheelbase, fueltype, horsepower, price)
head(M2)
```

```
##   wheelbase fueltype horsepower price
## 1      88.6      gas        111 13495
## 2      88.6      gas        111 16500
## 3      94.5      gas        154 16500
## 4      99.8      gas        102 13950
## 5      99.4      gas        115 17450
## 6      99.8      gas        110 15250
```

Exploración de la base de datos

Exploración de la base de datos

Calcula medidas estadísticas apropiadas para las variables: cuantitativas (media, desviación estándar, cuantiles, etc), cualitativas: cuantiles, frecuencias (puedes usar el comando table o prop.table)

```
#Wheelbase
summary(M2$wheelbase)
```

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

```
print(IQR(M2$wheelbase))
```

```
## [1] 7.9
```

```
#Horsepower
summary(M2$horsepower)
```

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

```
print(IQR(M2$horsepower))
```

```
## [1] 46
```

```
#Fueltype
x = table(M2$fueltype)
print(x)
```

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

```
prop.table(x)*100
```

```
##
##    diesel       gas
##  9.756098 90.243902
```

```
#Price
summary(M2$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5118    7788   10295   13277   16503   45400
```

```
print(IQR(M2$price))
```

```
## [1] 8715
```

Como podemos observar, la mediana de wheelbase es de 97, mientras su media es de 98.76, por lo que no parecen estar tan alejados estos valores. Sin embargo, su rango intercuartílico es de 7.9, por lo que el máximo de 121 para estar relativamente alejado del resto de los valores,

La media de horsepower es de 104.1, mientras su media es de 95 con un IQR de 46, por lo que no parecen estar demasiado alejados estos valores. Por otro lado, su máximo es de 288, lo cual sí está bastante alejado del resto de valores.

Para fueltype, tenemos 20 datos de tipo diesel y 185 datos de tipo gas, por lo que en proporción sería 9.76% diésel y 90.24% gasolina, lo cual es una preferencia hacia la gasolina muy notable.

Para price, tenemos una media de 13277 y una mediana de 10295, con un IQR de 8715, por lo que estos valores sí parecen estar alejados. Además, el valor máximo es de 45000, lo cual está muy alejado del resto de los datos.

Analiza la correlación entre las variables (analiza posible colinealidad entre las variables)

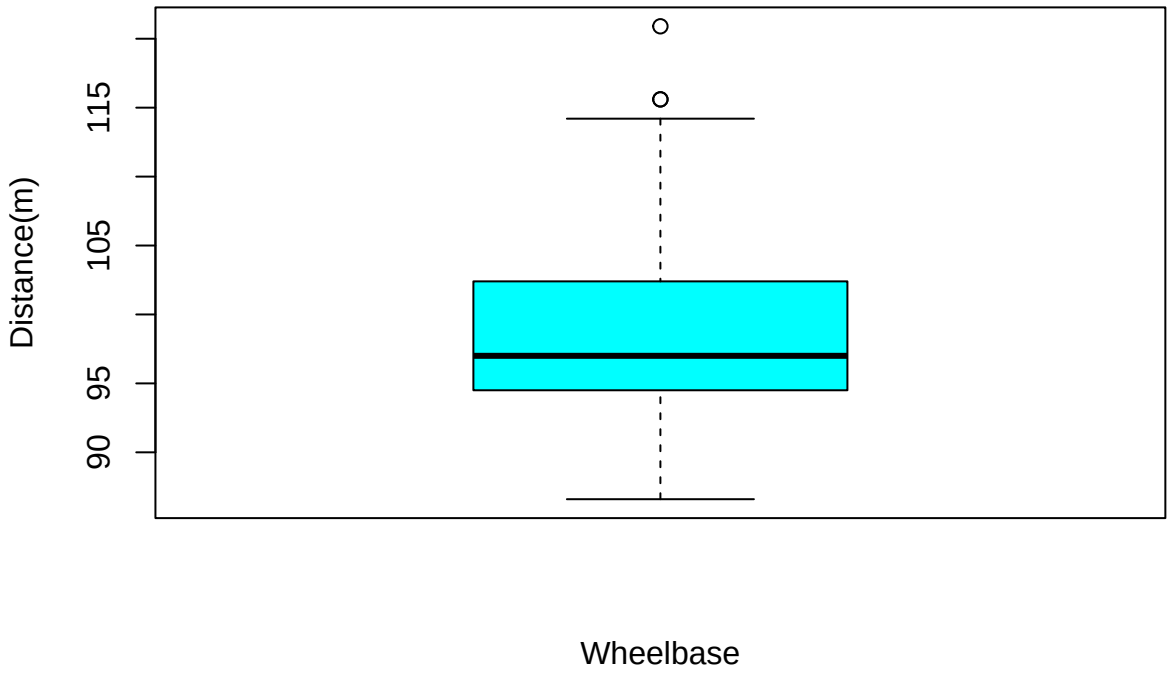
```
cor(M2[, sapply(M2, is.numeric)])
```

```
##            wheelbase horsepower     price
## wheelbase  1.0000000  0.3532945 0.5778156
## horsepower 0.3532945  1.0000000 0.8081388
## price      0.5778156  0.8081388 1.0000000
```

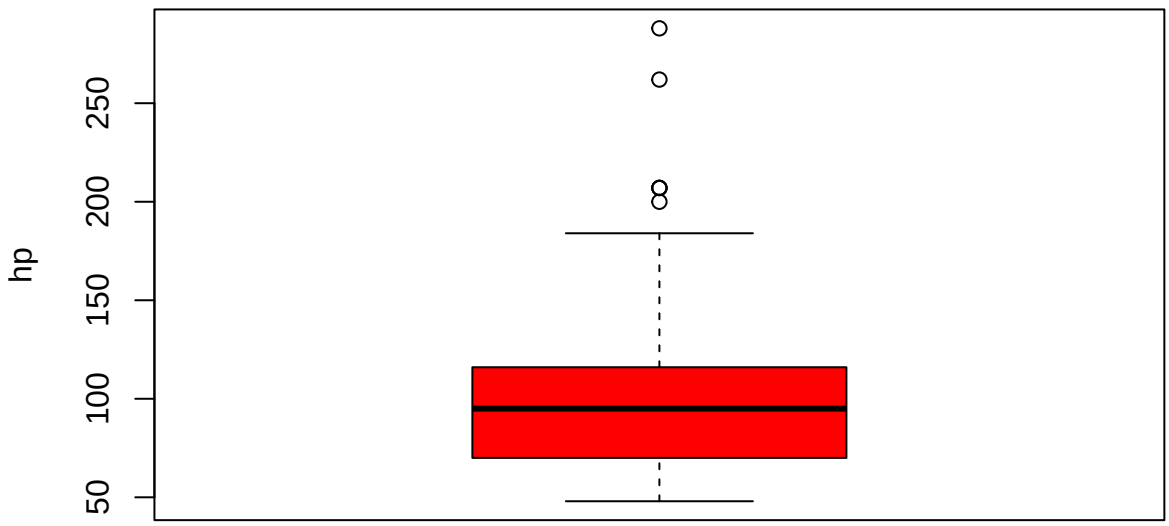
Observamos que tenemos una correlación de 0.57 entre wheelbase y price, lo cual parece ser una correlación no tan fuerte. Por otro lado, tenemos una correlación entre horsepower y price de 0.81, lo cual es una correlación fuerte y una correlación entre wheelbase y horsepower de 0.35, lo cual no es muy fuerte.

Explora los datos usando herramientas de visualización (si lo consideras necesario):

```
library(ggplot2)
boxplot(M2$wheelbase, col = 'cyan',
        horizontal = FALSE, xlab = "Wheelbase", ylab = 'Distance(m)')
```

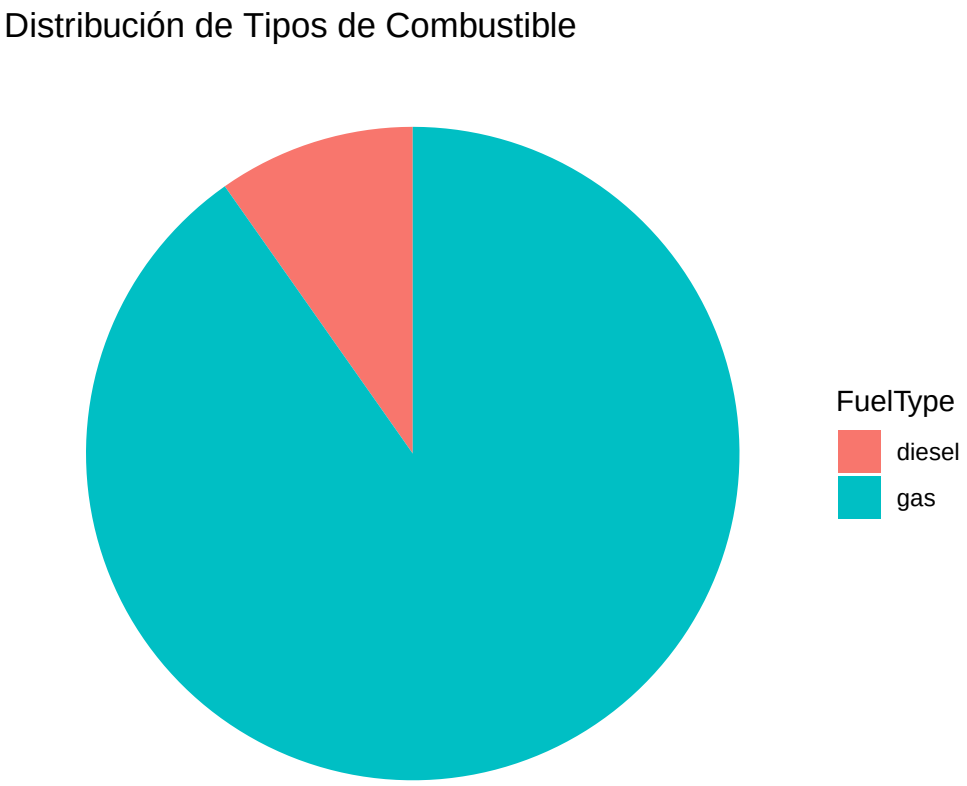



```
boxplot(M2$horsepower, col = 'red',
        horizontal = FALSE, xlab = "Horsepower", ylab = 'hp')
```



Horsepower

```
conteos <- as.data.frame(table(M2$fueltype))
colnames(conteos) <- c("FuelType", "Count")

ggplot(conteos, aes(x = "", y = Count, fill = FuelType)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar(theta = "y") +
  labs(title = "Distribución de Tipos de Combustible") +
  theme_void()
```



Como podemos observar, tenemos ciertos valores que se salen demasiado del rango aceptable para horsepower y que pueden afectar a nuestro modelo. Sin embargo, al ser datos representativos de la población y no ser erróneos (es decir, ser datos que sí pertenecen correctamente al conjunto de datos), no los eliminaremos y trabajaremos sobre ellos. Además, observamos un desbalanceo de clases muy alto para el tipo de combustible, pues la gran mayoría de personas utilizan gasolina, mientras muy pocas utilizan diésel.

Modelación y verificación del modelo

Encuentra la ecuación de regresión de mejor ajuste. Propón al menos 2 modelos de ajuste para encontrar la mejor forma de ajustar la variable precio.

Para cada uno de los modelos propuestos:

Realiza la regresión entre las variables involucradas

Modelo de horsepower, wheelbase y fueltype sin interacción para predecir el precio

```
Modelo2 = lm(M2$price~ M2$horsepower + M2$wheelbase + M2$fueltype, M2)
Modelo2
```

```
##
## Call:
```

```
## lm(formula = M2$price ~ M2$horsepower + M2$wheelbase + M2$fueltype,
##     data = M2)
##
## Coefficients:
##    (Intercept)   M2$horsepower    M2$wheelbase  M2$fueltypegas
##       -34754.3           148.3           364.7         -3794.5
```

```
summary(Modelo2)
```

```
##
## Call:
## lm(formula = M2$price ~ M2$horsepower + M2$wheelbase + M2$fueltype,
##     data = M2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8650  -2191   -197   1606  15816
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -34754.325   5314.194  -6.540 4.99e-10 ***
## M2$horsepower     148.323      7.723  19.205  < 2e-16 ***
## M2$wheelbase      364.657     52.594   6.933 5.48e-11 ***
## M2$fueltypegas  -3794.450   1009.750  -3.758 0.000225 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3884 on 201 degrees of freedom
## Multiple R-squared:  0.7671, Adjusted R-squared:  0.7636
## F-statistic: 220.7 on 3 and 201 DF,  p-value: < 2.2e-16
```

Nuestra ecuación nos queda price = -34754.3 + 148.3horsepower + 364.7wheelbase - 3794.5 * fueltypegas. En esta ecuación, fueltypegas representa si es gas o no. si es gas, el precio disminuye 3794.5, pero si no lo es, no disminuye nada.

```
alpha = 0.04
t_critical = abs(qt(alpha/2, length(M2$horsepower)-4))
cat('t frontera:', t_critical)
```

```
## t frontera: 2.067162
```

```
f_critical <- qf(1 - alpha, 3, length(M2$horsepower)-4)
cat('\nF frontera:', f_critical)
```

```
##
## F frontera: 2.82134
```

Analiza la significancia del modelo:

Valida la significancia del modelo con un alfa de 0.04 (incluye las hipótesis que pruebas y el valor frontera) H_0 : El modelo es estadísticamente significativo. H_1: El modelo es estadísticamente significativo.

Como nuestro estadístico F = 220.7 sobrepasa nuestro valor forntera de 2.82, y nuestro valor p < 0.04, la hipótesis inicial se rechaza, por lo que el modelo es estadísticamente significativo.

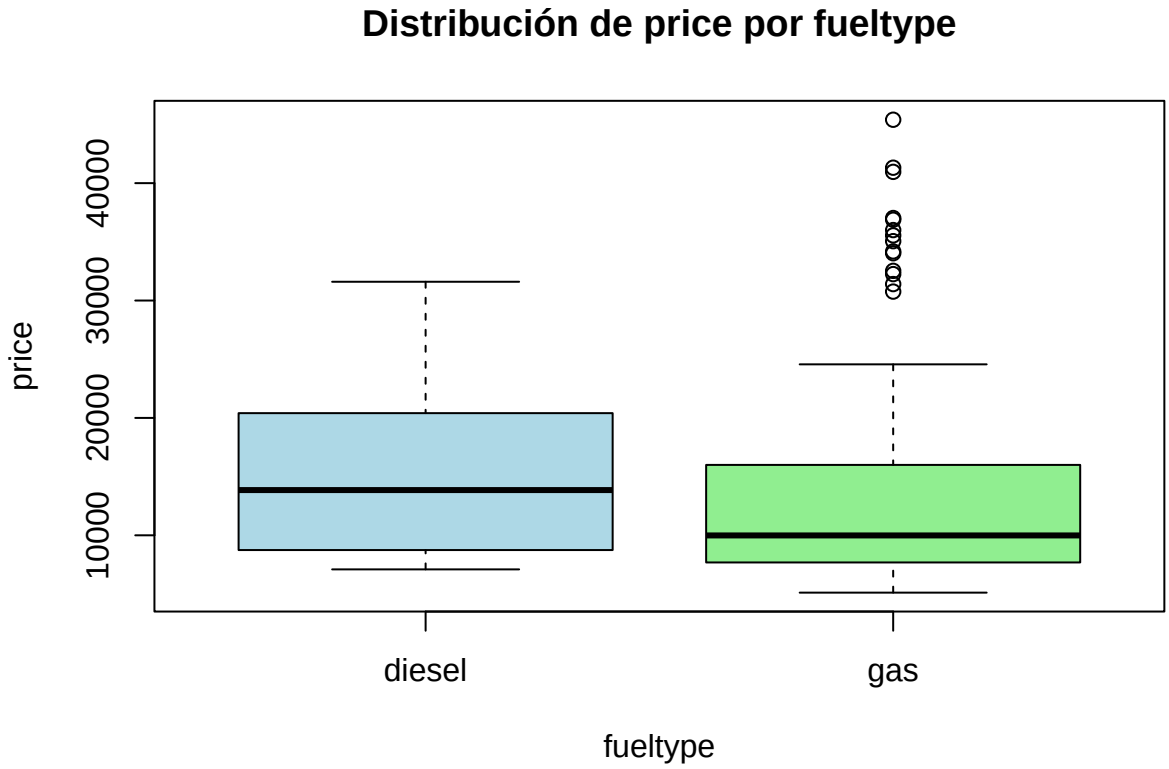
Valida la significancia de β_i con un alfa de 0.04 (incluye las hipótesis que pruebas y el valor frontera de cada una de ellas) $H_0 : \hat{\beta}_i = 0$. $H_1 : \hat{\beta}_i \neq 0$.

Como nuestros valores p de β 0, 1 2 y 3 son menores a 0.04 y los valores absolutos de t sobrepasan al valor t frontera, nuestra hipótesis inicial se rechaza, por lo que todos los coeficientes del modelo son estadísticamente significativos.

Indica cuál es el porcentaje de variación explicada por el modelo. El modelo tiene un R^2 de 0.7636 ajustado. Es decir, el modelo explica el 76.36% de la variación del precio.

```r
# Cargar ggplot2
library(ggplot2)

# Diagrama de dispersión para x1
p1 <- ggplot(M2, aes(x = wheelbase, y = price)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Diagrama de dispersión: wheelbase",
       x = "wheelbase",
       y = "price")

# Diagrama de dispersión para x2
p2 <- ggplot(M2, aes(x = horsepower, y = price)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Diagrama de dispersión: horsepower",
       x = "horsepower",
       y = "price")

boxplot(price ~ fueltype, data = M2,
        main = "Distribución de price por fueltype",
        xlab = "fueltype",
        ylab = "price",
        col = c("lightblue", "lightgreen", "lightpink"))
```

Dibuja el diagrama de dispersión de los datos por pares y la recta de mejor ajuste.




```
# Mostrar los gráficos
library(gridExtra)
```

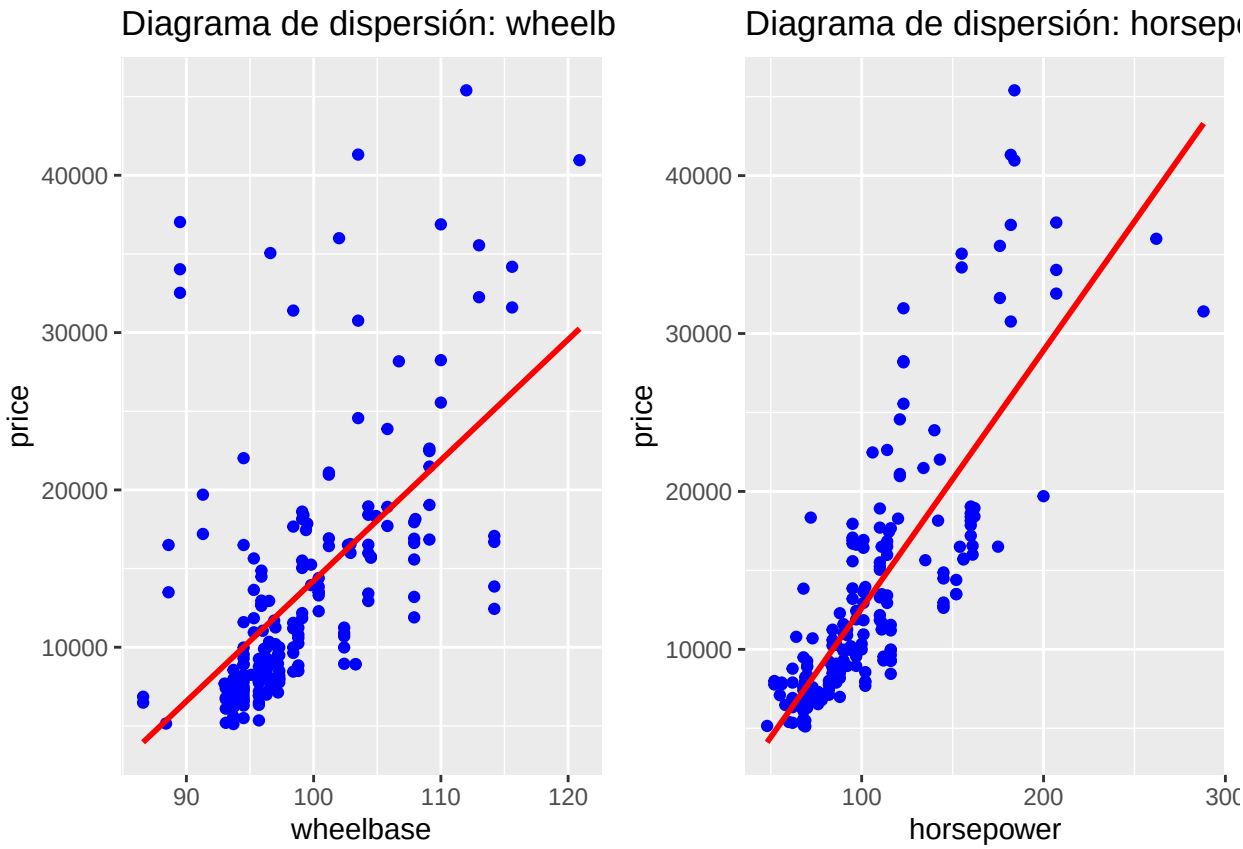
```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
grid.arrange(p1, p2, ncol = 2)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Interpreta en el contexto del problema cada uno de los análisis que hiciste. Como podemos observar, este modelo nos indica que podemos explicar el 76.36% de la variación del precio, por medio de la cantidad de caballos de fuerza el tipo de combustible que se utilizó y la distancia que hay entre los ejes.

Analiza la validez de los modelos propuestos:

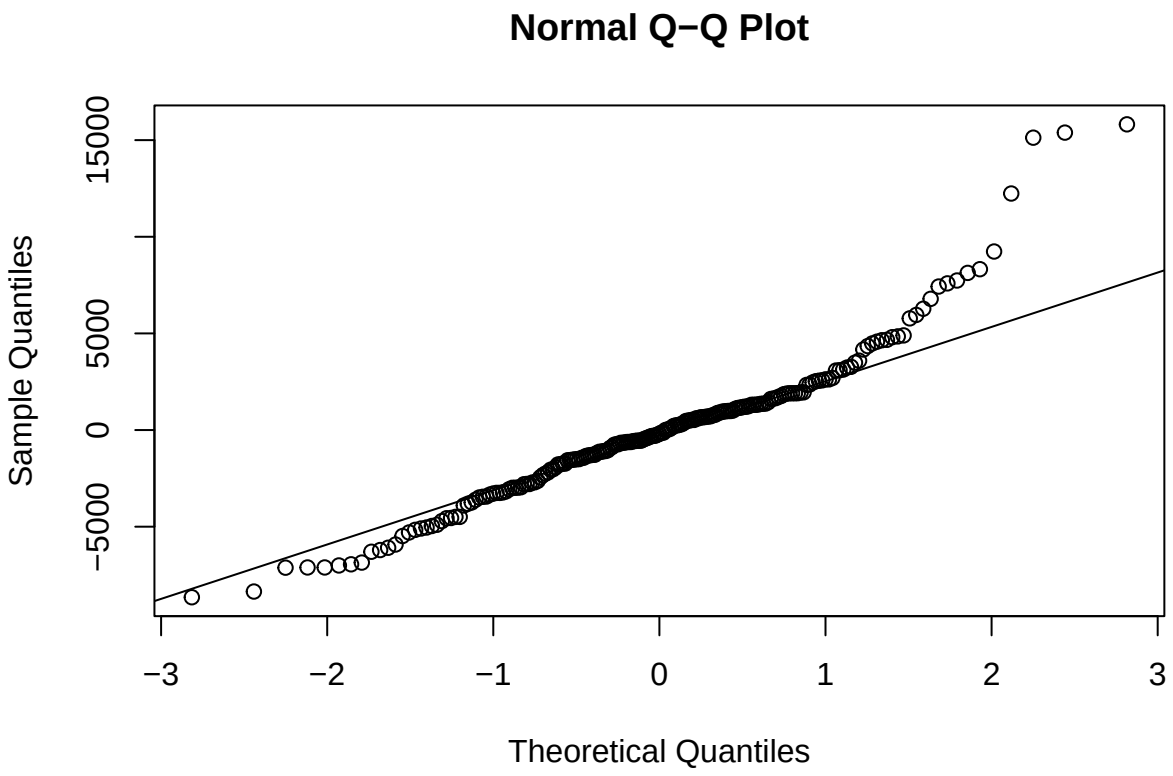
Normalidad

H_0 : Los residuos siguen una distribución normal. H_1 : Los residuos NO siguen una distribución normal.

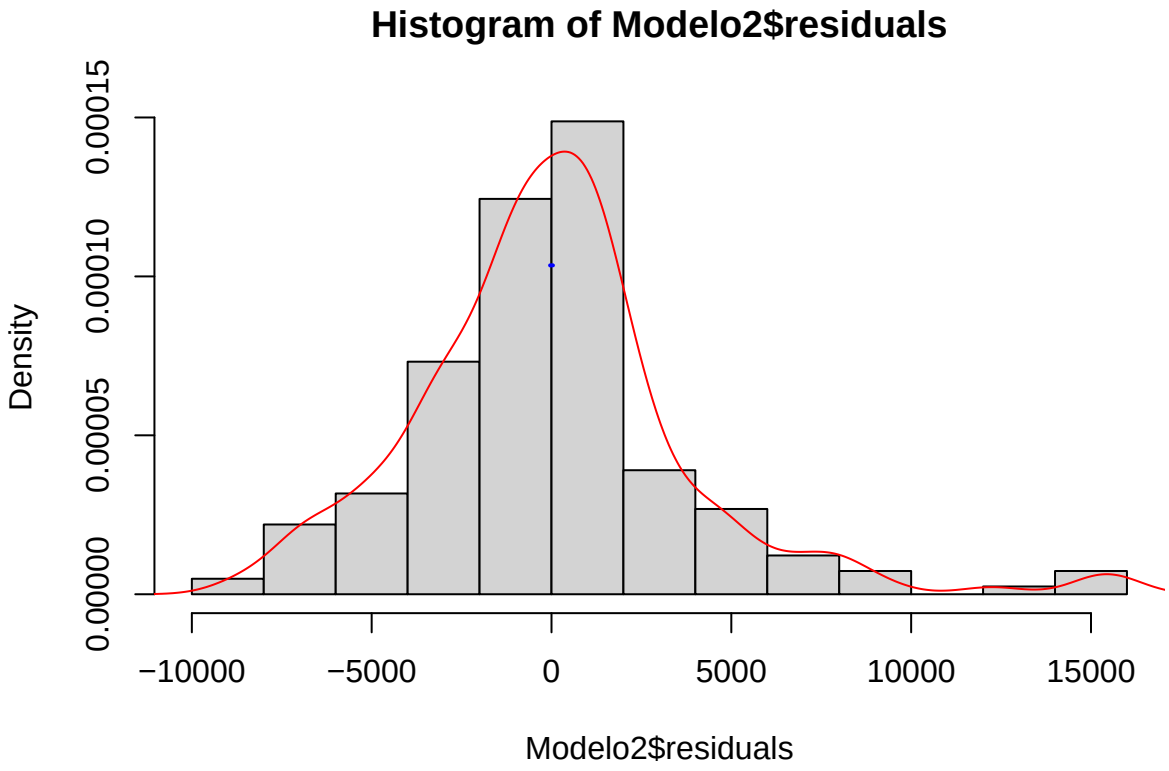
```
library(nortest)
ad.test(residuals(Modelo2))
```

```
##
##  Anderson-Darling normality test
##
## data:  residuals(Modelo2)
## A = 2.7561, p-value = 5.82e-07
```

```
qqnorm(Modelo2$residuals)
qqline(Modelo2$residuals)
```




```
hist(Modelo2$residuals,freq=FALSE)
lines(density(Modelo2$residuals),col="red")
curve(dnorm(x,mean=mean(Modelo2$residuals),sd=sd(Modelo2$residuals)), from=-
40, to=40, add=TRUE, col="blue",lwd=2)
```



Como podemos observar, el valor p < 0.04, por lo que se rechaza H_0 y los residuos no siguen una distribución normal. Esto también se observa en el gráfico qq, pues tenemos colas muy pesadas.

Verificación de media cero

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

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

```
##
##  One Sample t-test
##
## data:  Modelo2$residuals
## t = -2.6083e-16, df = 204, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -530.9376  530.9376
## sample estimates:
##     mean of x
## -7.023758e-14
```

Como tenemos un valor p ≈ 1, H_0 no se rechaza, por lo que los residuos tienen media cero.

Homocedasticidad

H_0 : La varianza de los errores es constante (Hay homocedasticidad). H_1: La varianza de los errores NO es constannte (Hay heterocedasticidad).

```
library(lmtest)
```

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

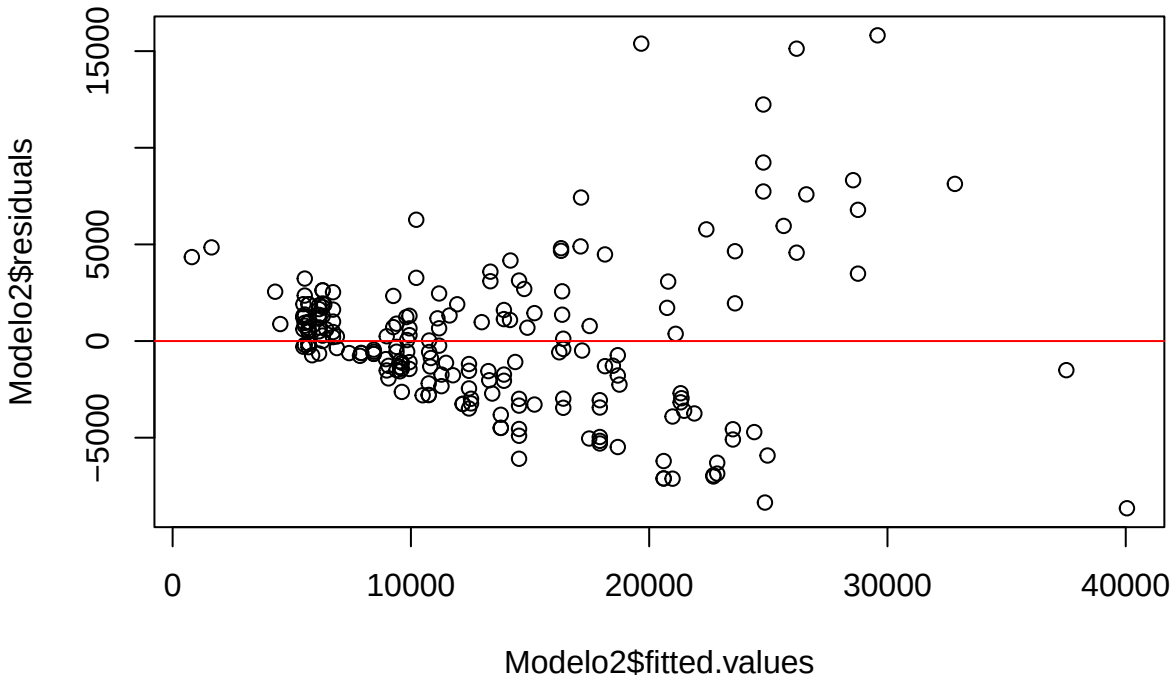
```
##
## Attaching package: 'zoo'
```

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

```
bptest(Modelo2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Modelo2
## BP = 62.451, df = 3, p-value = 1.759e-13
```

```
plot(Modelo2$fitted.values,Modelo2$residuals)
abline(h=0, col= 'red')
```



Como el valor p < 0.03, Se rechaza H_0, por lo que la varianza de los errores NO es constante (hay heterocedasticidad). Además, esto se puede observar en el gráfico, pues la varianza no fluctúa dentro de un rango, sino de manera caótica.

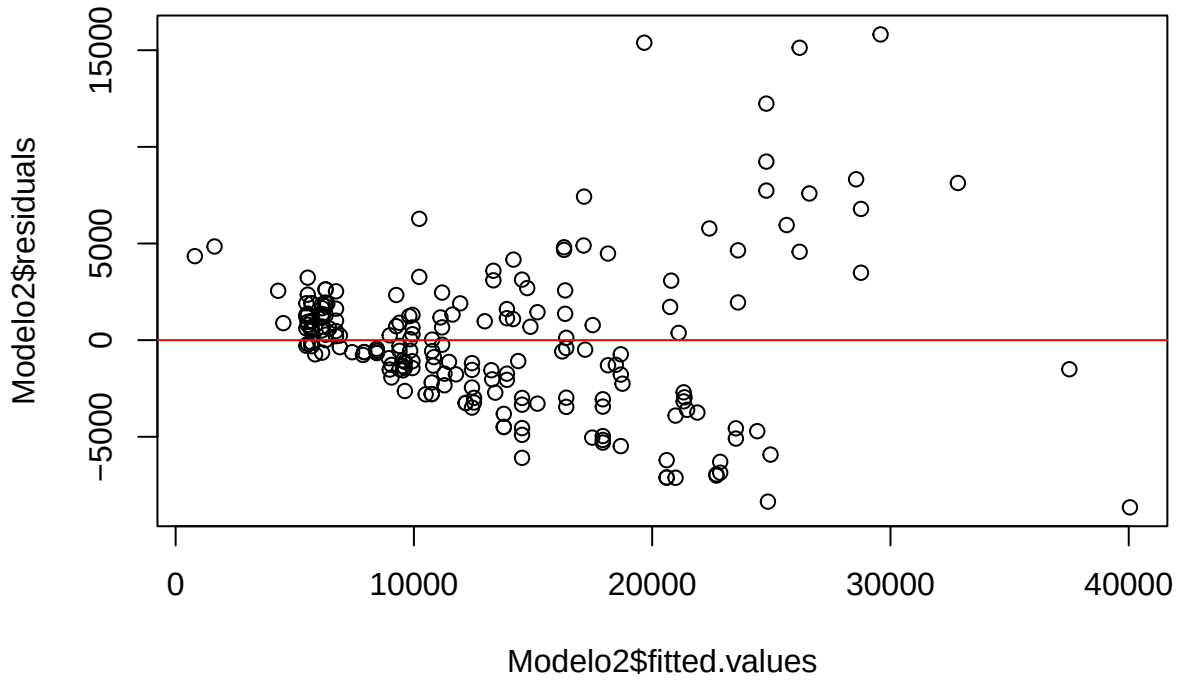
Independencia

H_0: La autocorrelación de los residuos es 0 (hay independencia). H_1: La autocorrelación de los residuos $\neq 0$ (no hay independencia).

```
dwtest(Modelo2)
```

```
##
##  Durbin-Watson test
##
## data:  Modelo2
## DW = 0.97856, p-value = 4.496e-14
## alternative hypothesis: true autocorrelation is greater than 0
```

```
plot(Modelo2$fitted.values,Modelo2$residuals)
abline(h=0, col= 'red')
```



Como el valor p < 0.03, se rechaza H_0, por lo que no hay independencia en los residuos. Además, se observa un patrón en la gráfica, pues entre más alejado, mayor el valor del residuo.

Conclusión

Debido a que el modelo no cumple con homocedasticidad, ni normalidad, ni independencia, este modelo no es adecuado, por lo que haremos un modelo con interacción entre horsepower y fueltype para predecir el precio.

Modelo de horsepower, fueltype con interacción para predecir el precio

```
Modelo1 = lm(M2$price~ M2$horsepower * M2$fueltype, M2)
Modelo1
```

```
##
## Call:
## lm(formula = M2$price ~ M2$horsepower * M2$fueltype, data = M2)
##
## Coefficients:
##                   (Intercept)                 M2$horsepower
##                       -7731.4                         279.1
##               M2$fueltypegas  M2$horsepower:M2$fueltypegas
##                        3016.8                        -112.4
```

```
summary(Modelo1)
```

```
##
## Call:
## lm(formula = M2$price ~ M2$horsepower * M2$fueltype, data = M2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11904.3  -1776.2   -381.8   1458.9  19435.5
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -7731.37    3298.65  -2.344  0.02006 *
## M2$horsepower                   279.09      37.42   7.459 2.56e-12 ***
## M2$fueltypegas                 3016.83    3414.35   0.884  0.37798
## M2$horsepower:M2$fueltypegas  -112.36      38.21  -2.940  0.00366 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4234 on 201 degrees of freedom
## Multiple R-squared:  0.7233, Adjusted R-squared:  0.7191
## F-statistic: 175.1 on 3 and 201 DF,  p-value: < 2.2e-16
```

Nuestra ecuación nos queda price = -7731.4 + 279.1horsepower + 3016.8fueltypegas - 112.4 (horsepower:fueltypegas). En esta ecuación, fueltypegas representa si es gas o no. si es gas, el precio aumenta 3016.8, pero si no lo es, no aumenta nada. Además, si es fueltypegas, el coeficiente de horsepower disminuye en 112.4.

```r
alpha = 0.04
t_critical = abs(qt(alpha/2, length(M2$horsepower)-4))
cat('t frontera:', t_critical)
```

```
## t frontera: 2.067162
```

```r
f_critical <- qf(1 - alpha, 3, length(M2$horsepower)-4)
cat('\nF frontera:', f_critical)
```

```
##
## F frontera: 2.82134
```

Analiza la significancia del modelo:

Valida la significancia del modelo con un alfa de 0.04 (incluye las hipótesis que pruebas y el valor frontera) H_0 : El modelo es estadísticamente significativo. H_1: El modelo es estadísticamente significativo.

Como nuestro estadístico F = 220.7 sobrepasa nuestro valor forntera de 2.82, y nuestro valor p < 0.04, la hipótesis inicial se rechaza, por lo que el modelo es estadísticamente significativo.

Valida la significancia de β_i con un alfa de 0.04 (incluye las hipótesis que pruebas y el valor frontera de cada una de ellas) $H_0 : \hat{\beta}_i = 0$. $H_1 : \hat{\beta}_i \neq 0$.

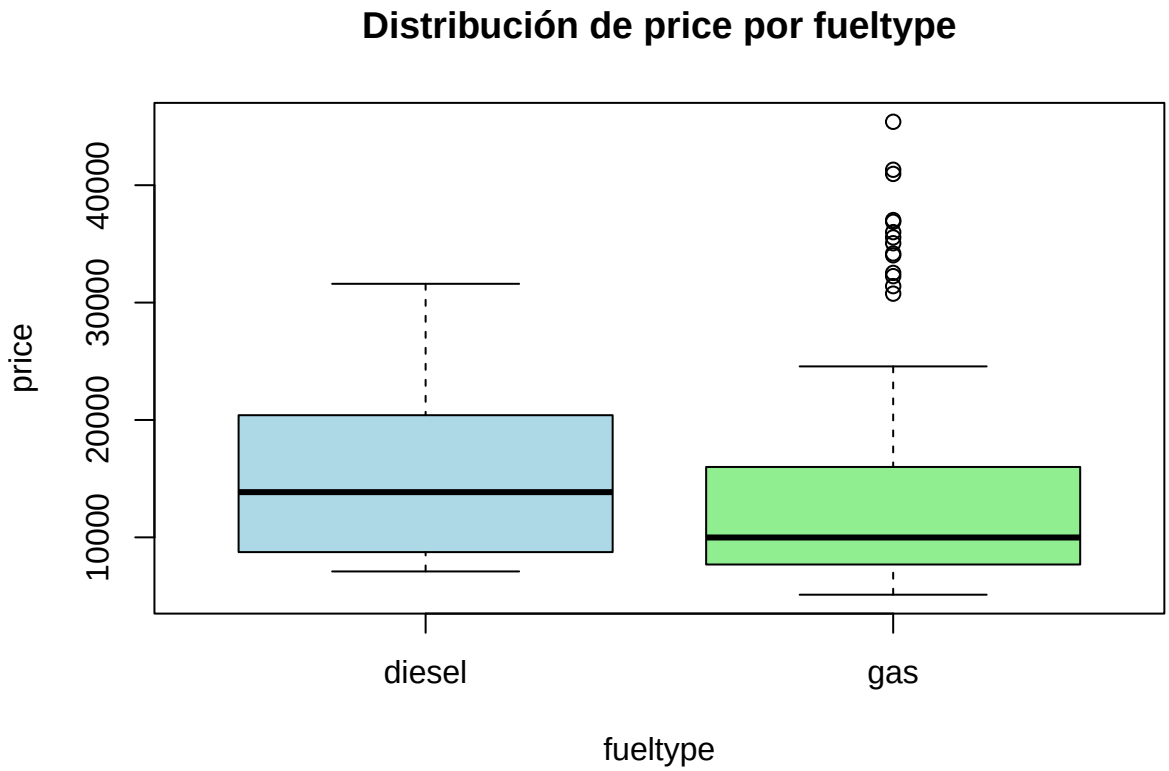
Como nuestros valores p de β 0, 1 y 3 son menores a 0.04 y los valores absolutos de t sobrepasan al valor t frontera, estos coeficientes del modelo son estadísticamente significativos. Por otro lado β_2 tiene un valor t menor al valor de la frontera y y su valor p > 0.04, por lo que este coeficiente (fueltypegas) no es estadísticamente significativo.

Indica cuál es el porcentaje de variación explicada por el modelo. El modelo tiene un R^2 de 0.7191 ajustado. Es decir, el modelo explica el 71.91% de la variación del precio.

```r
# Cargar ggplot2
library(ggplot2)

# Diagrama de dispersión para x1
p1 <- ggplot(M2, aes(x = wheelbase, y = price)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Diagrama de dispersión: wheelbase",
       x = "wheelbase",
       y = "price")

# Diagrama de dispersión para x2
p2 <- ggplot(M2, aes(x = horsepower, y = price)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Diagrama de dispersión: horsepower",
```

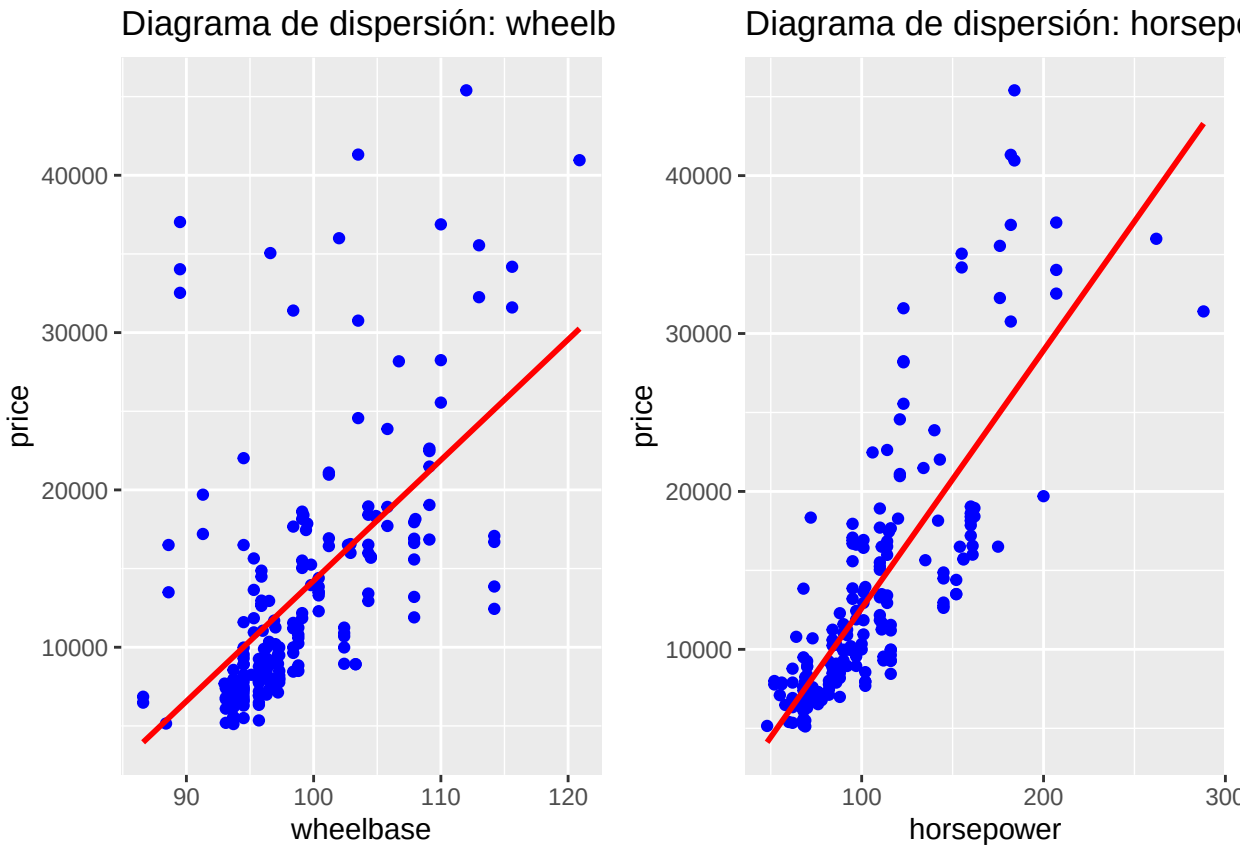
```
       x = "horsepower",
       y = "price")

boxplot(price ~ fueltype, data = M2,
        main = "Distribución de price por fueltype",
        xlab = "fueltype",
        ylab = "price",
        col = c("lightblue", "lightgreen", "lightpink"))
```

Dibuja el diagrama de dispersión de los datos por pares y la recta de mejor ajuste.




```
# Mostrar los gráficos
library(gridExtra)
grid.arrange(p1, p2, ncol = 2)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



Interpreta en el contexto del problema cada uno de los análisis que hiciste. Como podemos observar, este modelo nos indica que podemos explicar el 71.91% de la variación del precio, por medio de la cantidad de caballos de fuerza el tipo de combustible que se utilizó y la distancia que hay entre los ejes con interacción entre sí.

Analiza la validez de los modelos propuestos:

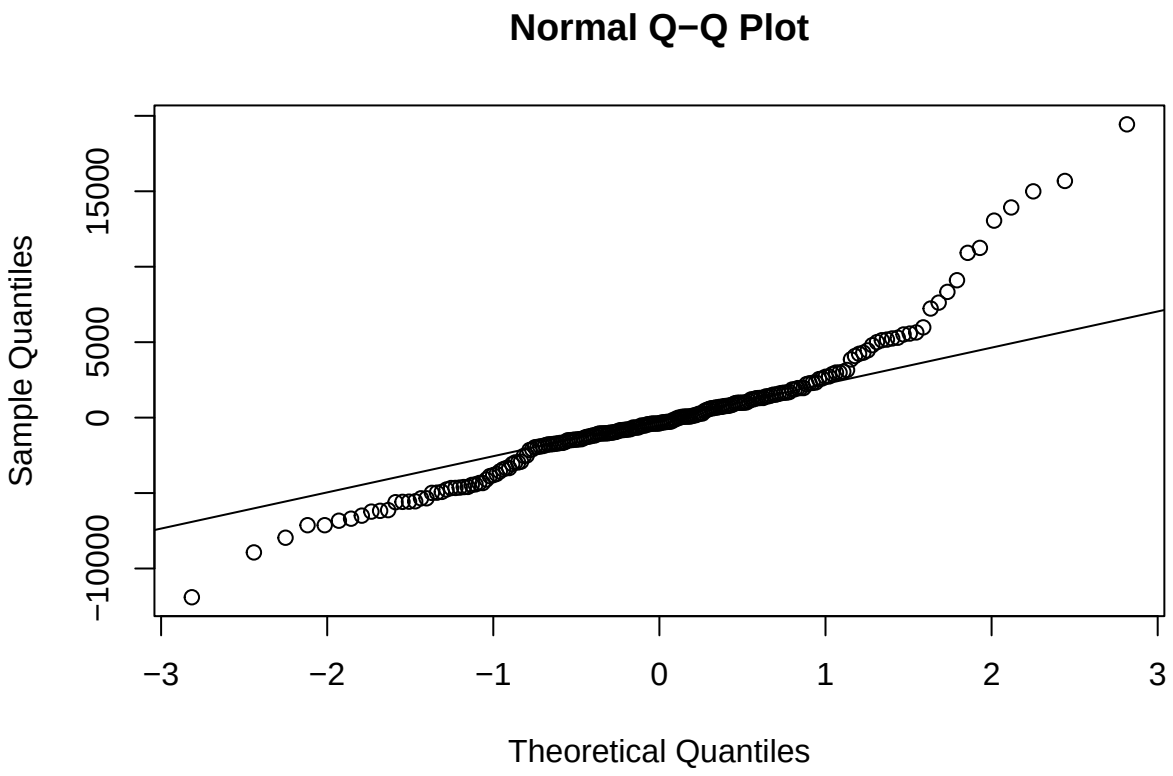
Normalidad

H_0 : Los residuos siguen una distribución normal. H_1 : Los residuos NO siguen una distribución normal.

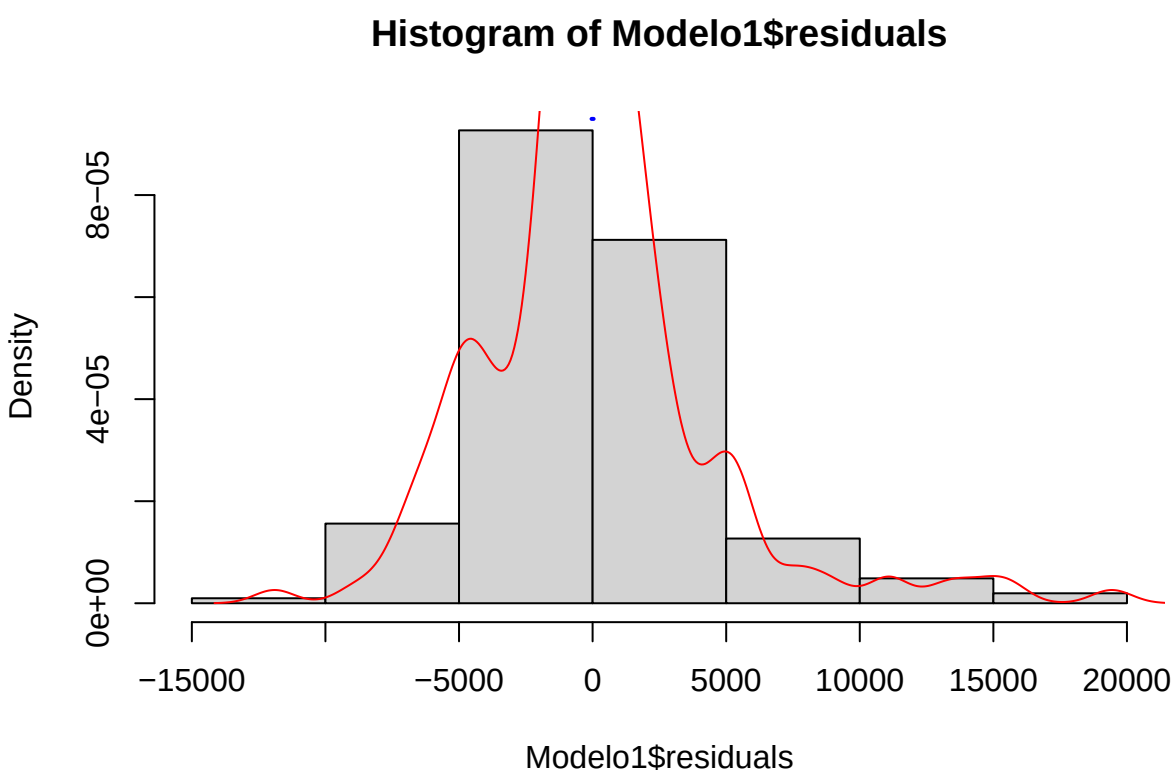
```
library(nortest)
ad.test(residuals(Modelo1))
```

```
##
##  Anderson-Darling normality test
##
## data:  residuals(Modelo1)
## A = 4.7325, p-value = 9.266e-12
```

```
qqnorm(Modelo1$residuals)
qqline(Modelo1$residuals)
```




```
hist(Modelo1$residuals,freq=FALSE)
lines(density(Modelo1$residuals),col="red")
curve(dnorm(x,mean=mean(Modelo1$residuals),sd=sd(Modelo1$residuals)), from=-
40, to=40, add=TRUE, col="blue",lwd=2)
```



Como podemos observar, el valor p < 0.04, por lo que se rechaza H_0 y los residuos no siguen una distribución normal. Esto también se observa en el gráfico qq, pues tenemos colas muy pesadas.

Verificación de media cero

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

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

```
##
##  One Sample t-test
##
## data:  Modelo1$residuals
## t = -1.5791e-15, df = 204, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -578.714  578.714
## sample estimates:
##     mean of x
## -4.634905e-13
```

Como tenemos un valor p ≈ 1, H_0 no se rechaza, por lo que los residuos tienen media cero.

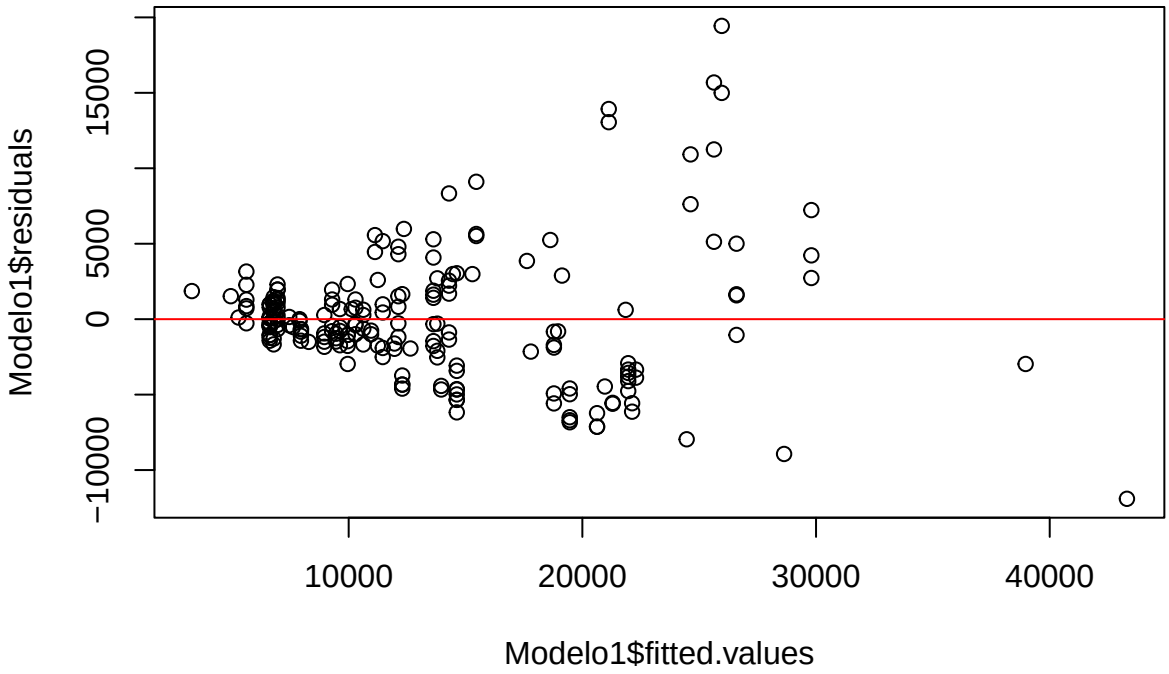
Homocedasticidad

H_0 : La varianza de los errores es constante (Hay homocedasticidad). H_1: La varianza de los errores NO es constannte (Hay heterocedasticidad).

```
library(lmtest)
bptest(Modelo1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Modelo1
## BP = 62.878, df = 3, p-value = 1.426e-13
```

```
plot(Modelo1$fitted.values,Modelo1$residuals)
abline(h=0, col= 'red')
```



Como el valor p < 0.03, Se rechaza H_0, por lo que la varianza de los errores NO es constante (hay heterocedasticidad). Además, esto se puede observar en el gráfico, pues la varianza no fluctúa dentro de un rango, sino de manera caótica.

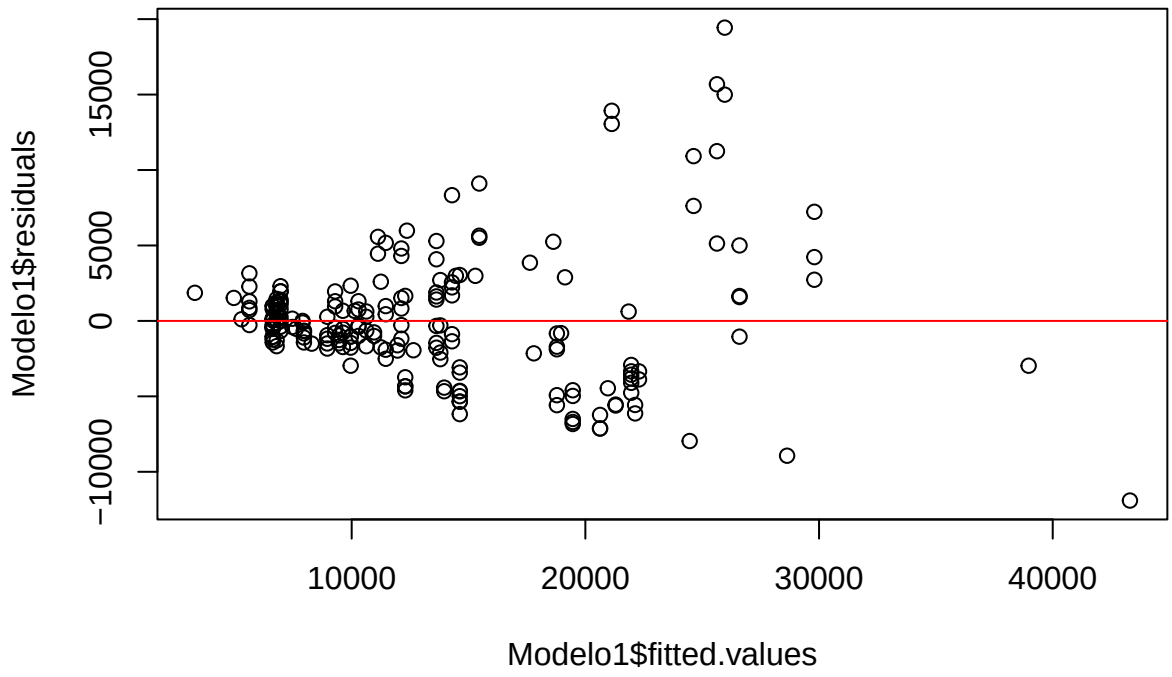
Independencia

H_0: La autocorrelación de los residuos es 0 (hay independencia). H_1: La autocorrelación de los residuos $\neq 0$ (no hay independencia).

```r
dwtest(Modelo1)
```

```
##
##  Durbin-Watson test
##
## data:  Modelo1
## DW = 1.0589, p-value = 3.272e-12
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
plot(Modelo1$fitted.values,Modelo1$residuals)
abline(h=0, col= 'red')
```



Como el valor p < 0.03, se rechaza H_0, por lo que no hay independencia en los residuos. Además, se observa un patrón en la gráfica, pues entre más alejado, mayor el valor del residuo.

Conclusión

Debido a que el modelo no cumple con homocedasticidad, ni normalidad, ni independencia, este modelo no es adecuado. Sin embargo, tampoco lo fue el otro.

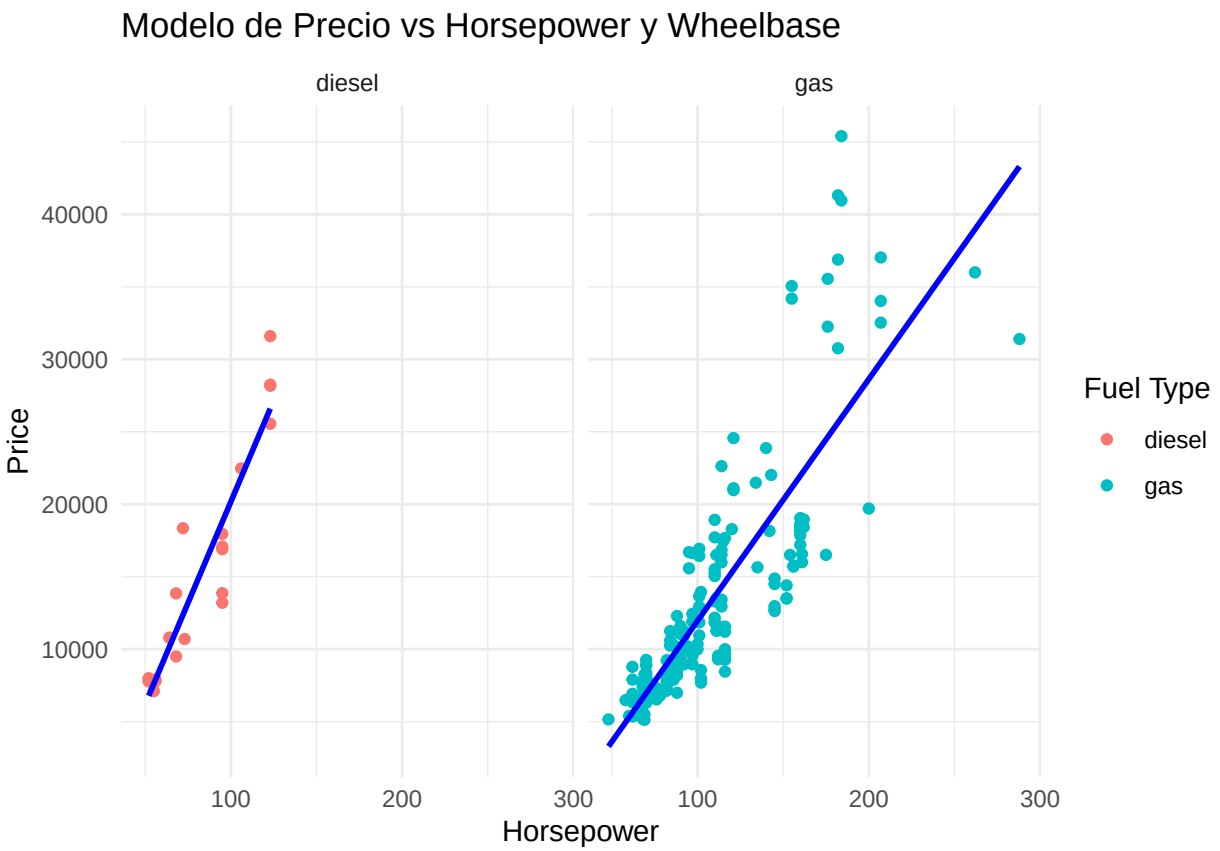
Conclusión final

Debido a que ningún modelo fue adecuado, escogeremos el que explica una mayor variación. Es decir, el modelo de wheelbase, horsepower y fueltype sin interacción para predecir el precio da mejores resultados, pues logra explicar el 76.36% de la variación en el precio del modelo.

Recta de mejor ajuste

```
modelo = Modelo2


beta_0 <- coef(modelo)[1]
beta_1 <- coef(modelo)[2]  # Para 'horsepower'
beta_2 <- coef(modelo)[3]  # Para 'wheelbase'
beta_3 <- coef(modelo)[4]  # Para 'fueltype' (este será un poco diferente porque es categórica)

# Escoge un nivel de fueltype (por ejemplo, el primer nivel)
nivel_fueltype <- levels(M2$fueltype)[1]

# Actualiza el modelo para un nivel específico de fueltype
beta_fueltype <- coef(modelo)[which(names(coef(modelo)) == paste0("fueltype", nivel_fueltype))]

Ym <- function(horsepower, wheelbase) {
  beta_0 + beta_1 * horsepower + beta_2 * wheelbase + beta_fueltype
}

# Gráfico de dispersión
library(ggplot2)

ggplot(M2, aes(x = horsepower, y = price, color = fueltype)) +
  geom_point() +
  labs(title = "Modelo de Precio vs Horsepower y Wheelbase",
       x = "Horsepower",
       y = "Price",
       color = "Fuel Type") +
  theme_minimal() +
  facet_wrap(~ fueltype) +
  geom_smooth(method = "lm", se = FALSE, aes(group = fueltype), color = "blue")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



¿Cuáles de las variables asignadas influyen en el precio del auto? ¿de qué manera lo hacen? En este modelo, las variables que influyen sobre el precio son horsepower, wheelbase y fueltype. La que más influye es wheelbase en nuestra ecuación deprice = -34754.3 + 148.3horsepower + 364.7wheelbase - 3794.5 * fueltypegas. Además, importa demasiado el tipo de combustible, pues la gente que usa gasolina paga 3794.5 menos a la gente que usa diésel. Los coeficientes representan los cambios de unidad. Es decir, por cada unidad de wheelbase y horsepower, el precio subre 364.7 y 148.3 unidades, respectivamente.

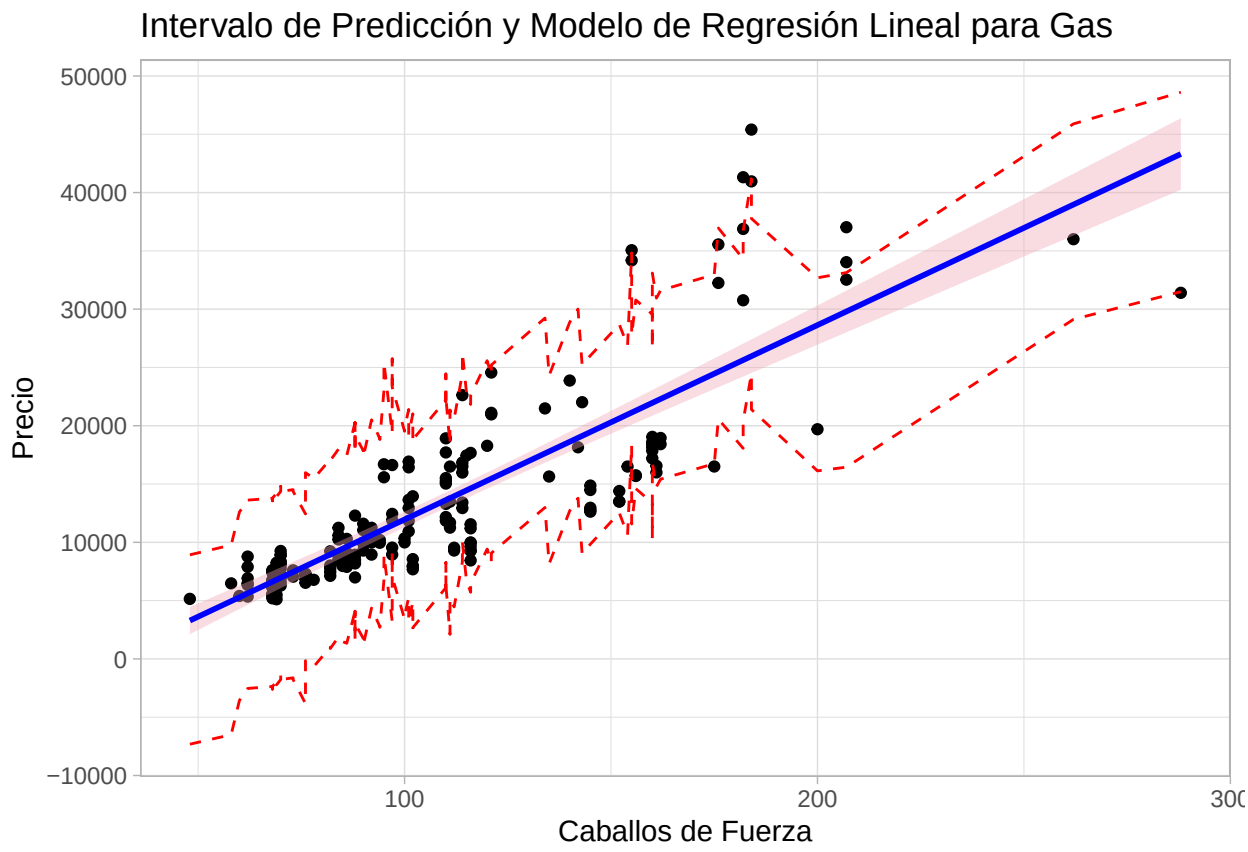
Intervalos de predicción y confianza

Con los datos de las variables asignadas construye la gráfica de los intervalos de confianza y predicción para la estimación y predicción del precio para el mejor modelo seleccionado: ## Calcula los intervalos para la variable Y ## Selecciona la categoría de la variable cualitativa que, de acuerdo a tu análisis resulte la más importante, y separa la base de datos por esa variable categórica. ## Grafica por pares de variables numéricas

En este caso, la categoría que vemos que tiene mayor relevancia es gas, pues el 95% de los datos pertenecen a esta categoría. De igual manera, haremos el análisis para diesel también.

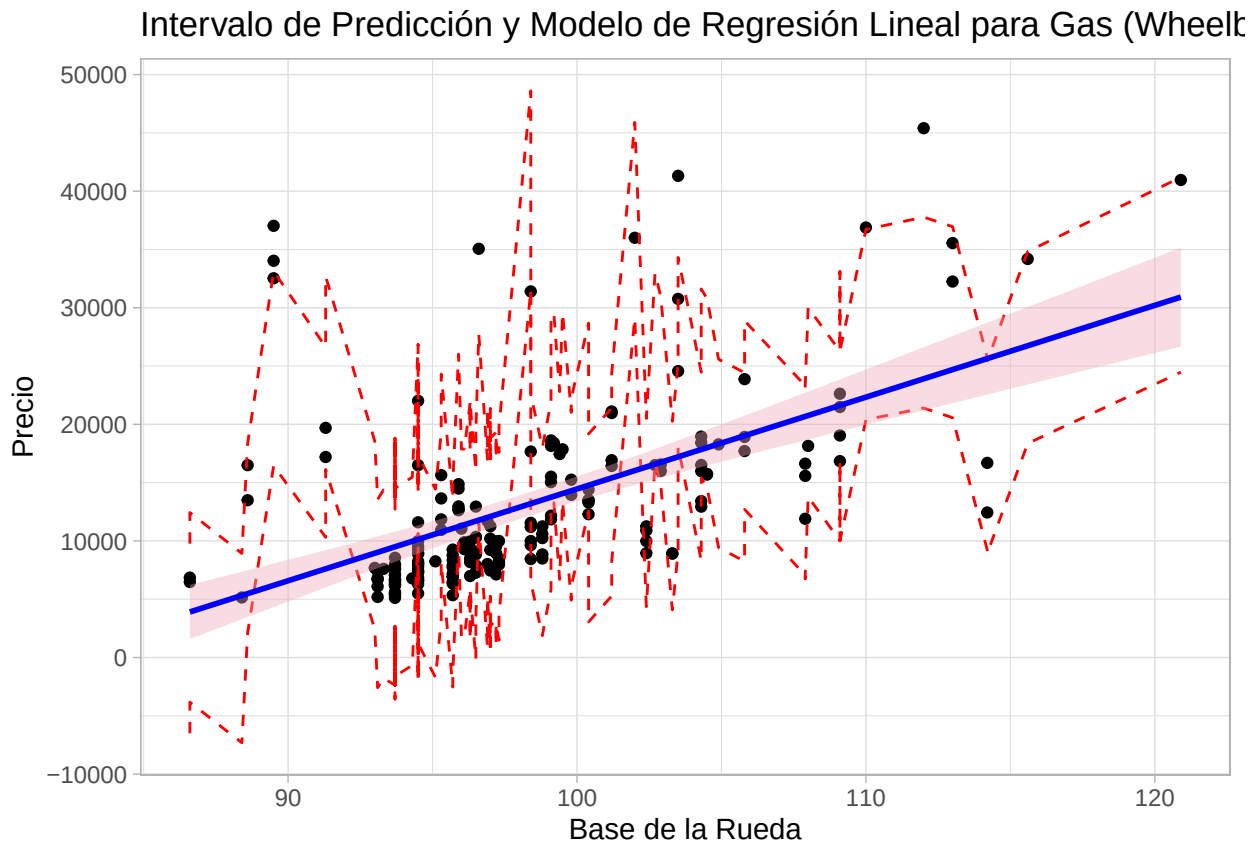
```
suppressWarnings({
ModeloSeleccionado = Modelo2

Ip <- predict(object = ModeloSeleccionado, interval = "prediction", level = 0.96)
datos <- cbind(M2, Ip)
```

```
datos_Gas <- subset(datos, fueltype == "gas")
datos_Diesel <- subset(datos, fueltype == "diesel")
})
library(ggplot2)
ggplot(datos_Gas, aes(x = horsepower, y = price)) +
  geom_point() +
  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 = 0.96, col = "blue", fill = "pink2") +
  labs(
    title = "Intervalo de Predicción y Modelo de Regresión Lineal para Gas",
    x = "Caballos de Fuerza",
    y = "Precio"
  ) +
  theme_light()
```

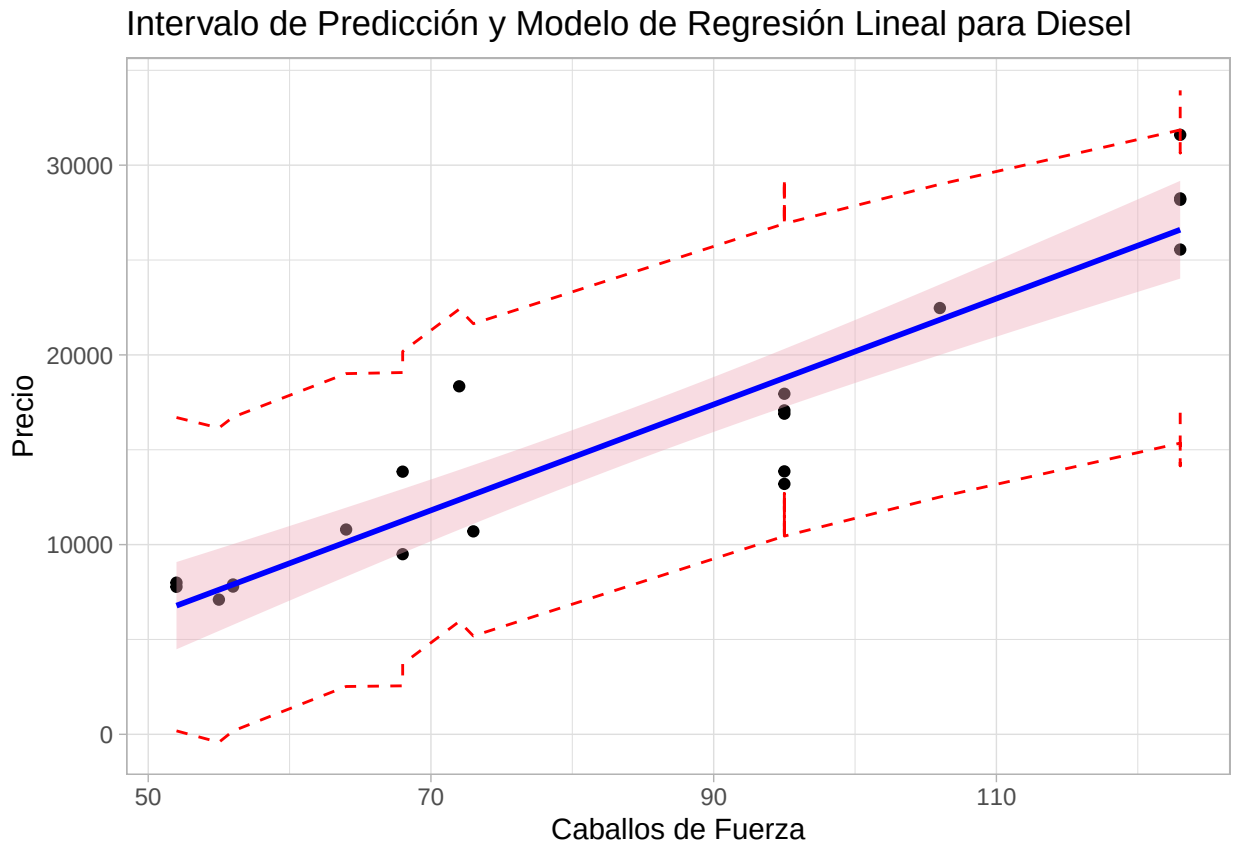



```
ggplot(datos_Gas, aes(x = wheelbase, y = price)) +
  geom_point() +
  geom_line(aes(y = lwr), color = "red", linetype = "dashed") +
  geom_line(aes(y = upr), color = "red", linetype = "dashed") +
  geom_smooth(method = lm, formula = y ~ x, se = TRUE, level = 0.96, col = "blue", fill = "pink2") +
  labs(
    title = "Intervalo de Predicción y Modelo de Regresión Lineal para Gas (Wheelbase)",
    x = "Base de la Rueda",
    y = "Precio"
```

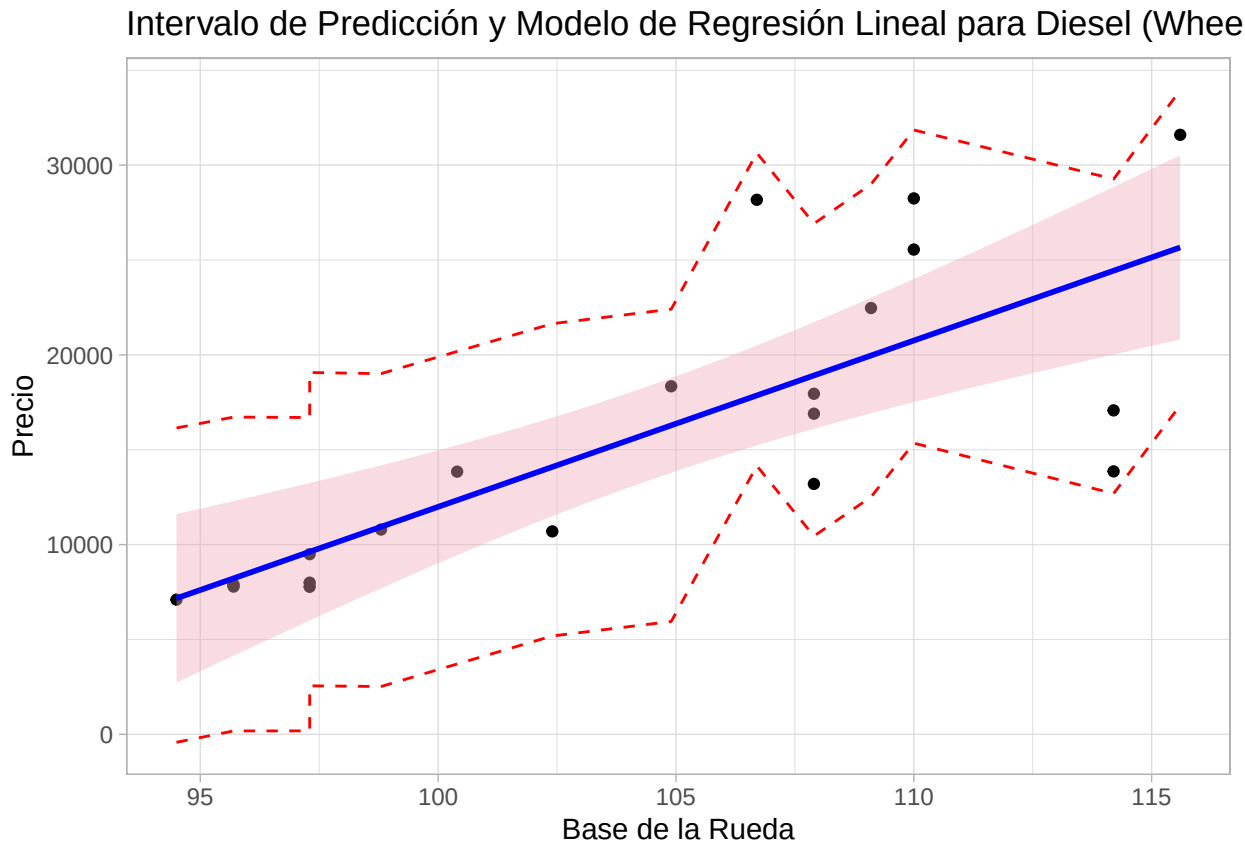
```
  ) +
  theme_light()
```




```
ggplot(datos_Diesel, aes(x = horsepower, y = price)) +
  geom_point() +
  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 = 0.96, col = "blue", fill = "pink2") +
  labs(
    title = "Intervalo de Predicción y Modelo de Regresión Lineal para Diesel",
    x = "Caballos de Fuerza",
    y = "Precio"
  ) +
  theme_light()
```




```
ggplot(datos_Diesel, aes(x = wheelbase, y = price)) +
  geom_point() +
  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 = 0.96, col = "blue", fill = "pink2") +
  labs(
    title = "Intervalo de Predicción y Modelo de Regresión Lineal para Diesel (Wheelbase)",
    x = "Base de la Rueda",
    y = "Precio"
  ) +
  theme_light()
```



Interpreta en el contexto del problema Como podemos observar, los datos no se ajustan a regresión lineal simple o múltiple, pues la dispersión de los datos y su variabilidad es caótica y no es fácil de predecir. Observamos cómo los valores reales de gas para con horsepower y wheelbase se salen de los intervalos de confianza del 96%, pues son valores que están muy alejados de las predicciones del modelo y de los intervalos donde puedes asegurar con un 96% de confianza que estarán los siguientes valores. Con diesel no se muestra esto, pero es muy probable que no se salgan de los intervalos del 96% de confianza, pues se tiene una muy pequeña cantidad de datos para esta categoría.

¿Propondrías una nueva agrupación de las variables a la empresa automovilísitica? Retoma todas las variables y haz un análisis estadístico muy leve (medias y correlación) de cómo crees que se deberían agrupar para analizarlas.

```
summary (M)
```

```
##    symboling         CarName            fueltype           carbody
##  Min.   :-2.0000   Length:205         Length:205         Length:205
##  1st Qu.: 0.0000   Class :character   Class :character   Class :character
##  Median : 1.0000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 0.8341
##  3rd Qu.: 2.0000
##  Max.   : 3.0000
```

```
##   drivewheel        enginelocation       wheelbase        carlength
##  Length:205         Length:205         Min.   : 86.60   Min.   :141.1
##  Class :character   Class :character   1st Qu.: 94.50   1st Qu.:166.3
##  Mode  :character   Mode  :character   Median : 97.00   Median :173.2
##                                        Mean   : 98.76   Mean   :174.0
##                                        3rd Qu.:102.40   3rd Qu.:183.1
##                                        Max.   :120.90   Max.   :208.1
##     carwidth        carheight       curbweight    enginetype
##  Min.   :60.30   Min.   :47.80   Min.   :1488   Length:205
##  1st Qu.:64.10   1st Qu.:52.00   1st Qu.:2145   Class :character
##  Median :65.50   Median :54.10   Median :2414   Mode  :character
##  Mean   :65.91   Mean   :53.72   Mean   :2556
##  3rd Qu.:66.90   3rd Qu.:55.50   3rd Qu.:2935
##  Max.   :72.30   Max.   :59.80   Max.   :4066
##  cylindernumber       enginesize        stroke      compressionratio
##  Length:205         Min.   : 61.0   Min.   :2.070   Min.   : 7.00
##  Class :character   1st Qu.: 97.0   1st Qu.:3.110   1st Qu.: 8.60
##  Mode  :character   Median :120.0   Median :3.290   Median : 9.00
##                     Mean   :126.9   Mean   :3.255   Mean   :10.14
##                     3rd Qu.:141.0   3rd Qu.:3.410   3rd Qu.: 9.40
##                     Max.   :326.0   Max.   :4.170   Max.   :23.00
##    horsepower       peakrpm        citympg        highwaympg        price
##  Min.   : 48.0   Min.   :4150   Min.   :13.00   Min.   :16.00   Min.   : 5118
##  1st Qu.: 70.0   1st Qu.:4800   1st Qu.:19.00   1st Qu.:25.00   1st Qu.: 7788
##  Median : 95.0   Median :5200   Median :24.00   Median :30.00   Median :10295
##  Mean   :104.1   Mean   :5125   Mean   :25.22   Mean   :30.75   Mean   :13277
##  3rd Qu.:116.0   3rd Qu.:5500   3rd Qu.:30.00   3rd Qu.:34.00   3rd Qu.:16503
##  Max.   :288.0   Max.   :6600   Max.   :49.00   Max.   :54.00   Max.   :45400
```

```
library(dplyr)
M3 <- M %>% select(carlength, carwidth, horsepower, enginesize, price)
head(M3)
```

```
##   carlength carwidth horsepower enginesize price
## 1     168.8     64.1        111        130 13495
## 2     168.8     64.1        111        130 16500
## 3     171.2     65.5        154        152 16500
## 4     176.6     66.2        102        109 13950
## 5     176.6     66.4        115        136 17450
## 6     177.3     66.3        110        136 15250
```

```
#carlength
summary(M3$carlength)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   141.1   166.3   173.2   174.0   183.1   208.1
```

```
print(IQR(M3$carlength))
```

```
## [1] 16.8
```

```r
#carwidth
summary(M3$carwidth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   60.30   64.10   65.50   65.91   66.90   72.30
```

```r
#carheight
summary(M3$horsepower)
```

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

```r
print(IQR(M3$horsepower))
```

```
## [1] 46
```

```r
#peakrpm
summary(M3$enginesize)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    61.0    97.0   120.0   126.9   141.0   326.0
```

```r
print(IQR(M3$enginesize))
```

```
## [1] 44
```

```r
#Price
summary(M3$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5118    7788   10295   13277   16503   45400
```

```r
print(IQR(M3$price))
```

```
## [1] 8715
```

```r
cor(M3)
```

```
##            carlength  carwidth horsepower enginesize     price
## carlength  1.0000000 0.8411183  0.5526230  0.6833599 0.6829200
## carwidth   0.8411183 1.0000000  0.6407321  0.7354334 0.7593253
## horsepower 0.5526230 0.6407321  1.0000000  0.8097687 0.8081388
## enginesize 0.6833599 0.7354334  0.8097687  1.0000000 0.8741448
## price      0.6829200 0.7593253  0.8081388  0.8741448 1.0000000
```

Observando los valores anteriores le propondría a la empresa utilizar las variables de carlength, carwidth, carheight y peakrpm, pues son relevantes como información de los automóviles y permite a la empresa realizar un análisis más completo. Además, pareciera que tienen mejores distribuciones, por lo que es posible que con estos se obtenga un modelo con un mejor ajuste.

Finalmente, podemos observar que estas cuatro variables tienen una buena correlación (ya sea moderada o fuerte) con la variable de price, por lo que son adecuadas y sería bueno realizar un análisis más complejo para observar qué tan relevantes realmente son.