Actividad_Integradora2

Rodolfo Jesús Cruz Rebollar

2024-09-06

```
# Importar base de datos

autos = read.csv("precios_autos.csv")

head(autos)
```

```
##   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
```

```
# Wheelbase, fueltype, horsepower (primer grupo de variables a analizar)

grupo1 = autos[, c("wheelbase", "fueltype", "horsepower", "price")]

head(grupo1)
```

```
##   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
```

1. Exploración de la base de datos

```
# Medidas cuantitativas

summary(grupo1)
```

```
##    wheelbase        fueltype          horsepower        price
##  Min.   : 86.60   Length:205         Min.   : 48.0   Min.   : 5118
##  1st Qu.: 94.50   Class :character   1st Qu.: 70.0   1st Qu.: 7788
##  Median : 97.00   Mode  :character   Median : 95.0   Median :10295
##  Mean   : 98.76                      Mean   :104.1   Mean   :13277
##  3rd Qu.:102.40                      3rd Qu.:116.0   3rd Qu.:16503
##  Max.   :120.90                      Max.   :288.0   Max.   :45400
```

```
# Frecuencia para la varible categorica de tipo de gasolina

table(grupo1[, "fueltype"])
```

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

```
# Matriz de correlación entre wheelbase, horsepower y precio del
automóvil

cor(grupo1[, c("wheelbase", "horsepower", "price")])
```

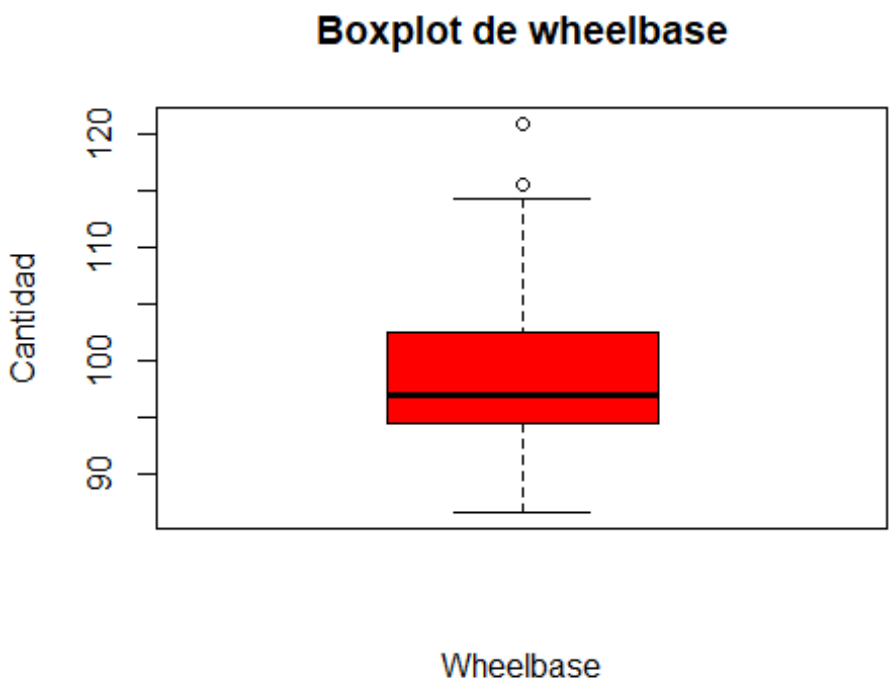
```
##            wheelbase horsepower     price
## wheelbase  1.0000000  0.3532945 0.5778156
```

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

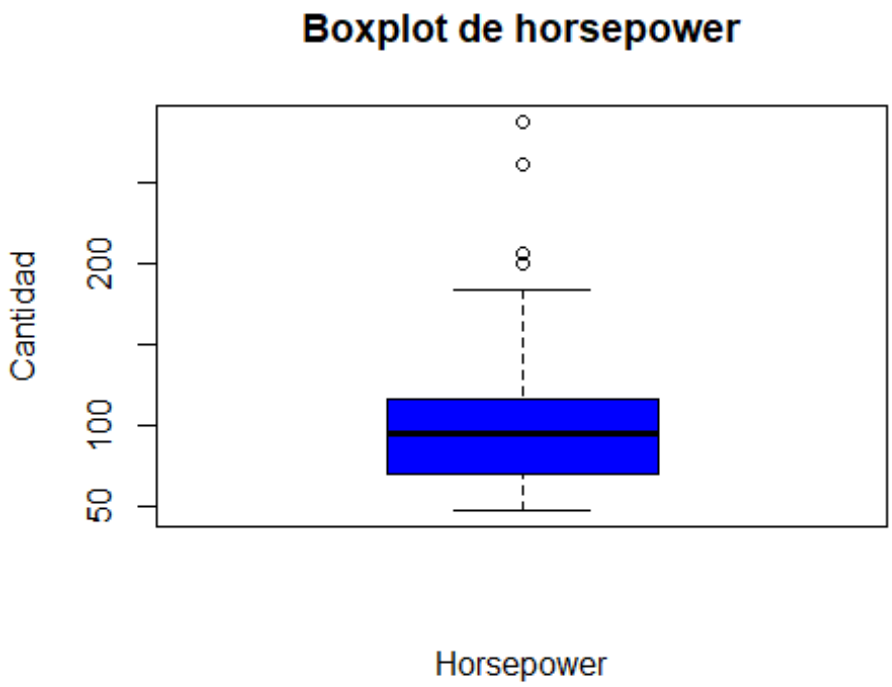
En la matriz de correlación se observa que ambas variables numéricas del grupo 1 presentan un nivel de correlación de 0.3532 entre ellas, por lo cual se puede afirmar que a pesar del bajo nivel de correlación, dicha cantidad de la misma es positiva, por lo cual eso indica que conforme aumenta una variable también lo hace la otra, por lo tanto, se concluye que el grado de correlación entre wheelbase y horsepower es bajo, no obstante, a mayor valor de una, la otra también incrementa su valor. Además de lo anterior, también cabe mencionar que en cuanto a la correlación de la variable wheelbase con el precio de los automóviles, éste es de 0.57781, mientras que la correlación existente entre horsepower y price es de 0.8081, por lo cual se puede afirmar que la cantidad de caballos de fuerza de un automóvil se correlaciona fuertemente con su precio, por lo cual además la correlación positiva indica que a mayor cantidad de caballos de fuerza de un automóvil, mayor será el precio del mismo.

Explorar datos con herramientas de visualización

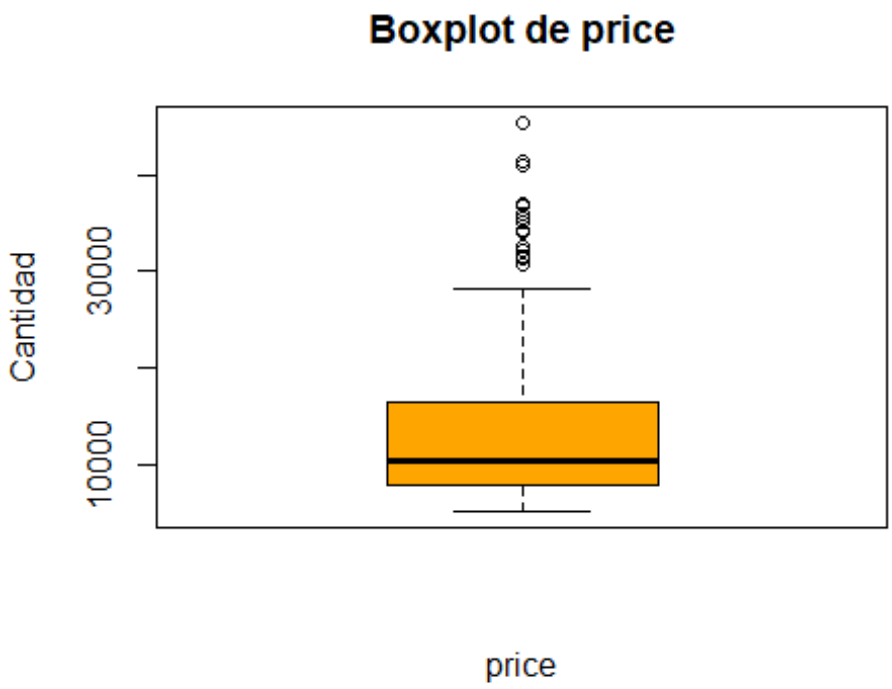
```
# Boxplot de la variable  del grupo 1 wheelbase

boxplot(grupo1["wheelbase"], col = "red", main="Boxplot de wheelbase",
xlab="Wheelbase",
        ylab="Cantidad")
```



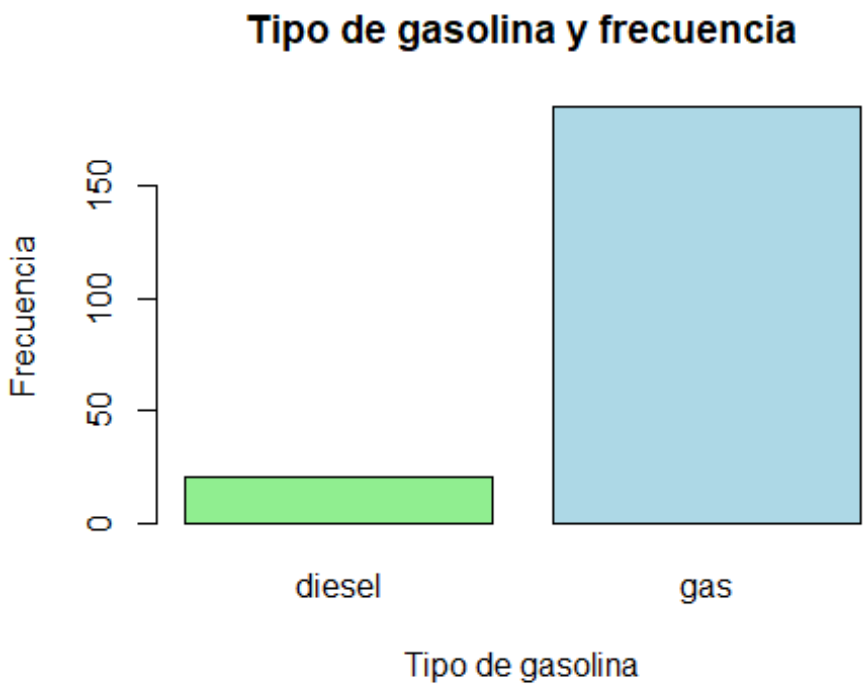
```r
# Boxplot de la variable  del grupo 1 horsepower

boxplot(grupo1["horsepower"], col = "blue", main="Boxplot de horsepower",
        xlab="Horsepower", ylab="Cantidad")
```




```r
# Boxplot de la variable  del grupo 1 price

boxplot(grupo1["price"], col = "orange", main="Boxplot de price",
xlab="price",
        ylab="Cantidad")
```




```r
# Gráfico de barras por tipo de gasolina (variable categórica del grupo 1)

barplot(table(grupo1$fueltype), col = c("lightgreen", "lightblue"),
        xlab = "Tipo de gasolina", ylab = "Frecuencia",
        main = "Tipo de gasolina y frecuencia")
```



2. Modelación y verificación del modelo

```
# modelo 1: precio a partir de la distancia entre ejes del automóvil
(wheelbase)

wb.model = lm(price ~ wheelbase, data = grupo1)

# mostrar modelo 1

wb.model

##
## Call:
## lm(formula = price ~ wheelbase, data = grupo1)
##
## Coefficients:
## (Intercept)    wheelbase
##    -62426.6        766.6

# Modelo 2: price a partir de horsepower

hp.model = lm(price ~ horsepower, data = grupo1)

# Mostrar modelo 2

hp.model
```

```
##
## Call:
## lm(formula = price ~ horsepower, data = grupo1)
##
## Coefficients:
## (Intercept)   horsepower
##     -3721.8        163.3
```

Ecuaciones de los modelos propuestos

- Price vs wheelbase:

```
# ecuacion del modelo de price vs wheelbase

cat("Modelo 1: ", "price = ", wb.model$coefficients[1], " + ",
    wb.model$coefficients[2], "* wheelbase")

## Modelo 1:  price =  -62426.65  +  766.5652 * wheelbase
```

- Price vs horsepower:

```
# ecuacion del modelo de price vs horsepower

cat("Modelo 2: ", "price = ", hp.model$coefficients[1], " + ",
    hp.model$coefficients[2], "* horsepower")

## Modelo 2:  price =  -3721.761  +  163.2631 * horsepower
```

Modelación y verificación de los modelos

Significancia de modelos

Hipótesis a verificar:

$H_0: $ el modelo no es estadísticamente significativo.

$H_1: $ el modelo sí es estadísticamente significativo.

```
# Para analizar significancia del modelo 1 se procederá a desplegar el
resumen del modelo

summary(wb.model)

##
## Call:
## lm(formula = price ~ wheelbase, data = grupo1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12675  -3364  -1956   1264  30847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -62426.7     7519.0  -8.303 1.42e-14 ***
## wheelbase      766.6       76.0  10.087  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6536 on 203 degrees of freedom
## Multiple R-squared:  0.3339, Adjusted R-squared:  0.3306
## F-statistic: 101.7 on 1 and 203 DF,  p-value: < 2.2e-16
```

```r
# Calcular la media de la variable independiente del modelo (wheelbase)

mean_wb = mean(grupo1$wheelbase)

# Desviación estándar de datos de wheelbase

sigma_wb = sd(grupo1$wheelbase) / sqrt(nrow(grupo1))

# Calcular valor frontera para probar la hipótesis con el modelo 1

x1_b = qnorm(0.04, mean_wb, sigma_wb)

z1_frontera = abs((x1_b - mean_wb) / sigma_wb)

cat("Valor frontera de la hipótesis para el modelo 1: ", z1_frontera)
```

```
## Valor frontera de la hipótesis para el modelo 1:  1.750686
```

```r
# Resumen del modelo 2 para analizar su nivel de significancia

summary(hp.model)
```

```
##
## Call:
## lm(formula = price ~ horsepower, data = grupo1)
##
## 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
```

```r
# Calcular la media de la variable independiente del modelo (horsepower)

mean_hp = mean(grupo1$horsepower)

# Desviación estándar de datos de horsepower

sigma_hp = sd(grupo1$horsepower) / sqrt(nrow(grupo1))

# Calcular valor frontera para probar la hipótesis con el modelo 2

x2_b = qnorm(0.04, mean_hp, sigma_hp)

z2_frontera = abs((x2_b - mean_hp) / sigma_hp)

cat("Valor frontera de la hipótesis para el modelo 2: ", z2_frontera)

## Valor frontera de la hipótesis para el modelo 2:  1.750686
```

Valor frontera para la hipótesis probada: 1.7506861

En cuanto a la significancia del modelo 1, se puede apreciar que éste modelo tiene un valor p inferior a 2.2e-16, lo cual es mucho menor que 0.04 y eso significa que se rechaza H_0, por lo que se puede concluir que el modelo 1 del precio en función de la variable wheelbase sí es estadísticamente significativo. Por otro lado, en cuanto al modelo 2 se puede observar que tiene un p valor menor a 2.2e-16, lo cual al ser menor que 0.04, se rechaza H_0, lo cual implica que el modelo 2 del precio en función de la variable horsepower sí es estadísticamente significativo.

Significancia de β_i

Hipótesis para β_0:

$H_0: \beta_0 = 0$ (β_0 no es significativo).

$H_1: \beta_0 \neq 0$ (β_0 sí es significativo).

Hipótesis para β_1:

$H_0: \beta_1 = 0$ (β_1 no es significativo).

$H_1: \beta_1 \neq 0$ (β_1 sí es significativo).

```r
# summary de ambos modelos

summary(wb.model) # modelo 1

##
## Call:
## lm(formula = price ~ wheelbase, data = grupo1)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -12675  -3364  -1956   1264  30847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -62426.7     7519.0  -8.303 1.42e-14 ***
## wheelbase      766.6       76.0  10.087  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6536 on 203 degrees of freedom
## Multiple R-squared:  0.3339, Adjusted R-squared:  0.3306
## F-statistic: 101.7 on 1 and 203 DF,  p-value: < 2.2e-16
```

```
summary(hp.model) # modelo 2
```

```
##
## Call:
## lm(formula = price ~ horsepower, data = grupo1)
##
## 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
```

Para el caso del modelo 1, se observa que el coeficiente β_0 tiene un p valor de 1.42e-14, lo cual al ser menor que 0.04, se rechaza H_0, por lo que β_0 sí es significativo. Además para el caso del modelo 2, se aprecia que el coeficiente β_0 posee un p valor igual a 8.78e-05, lo cual es menor a 0.04 y provoca el rechazo de H_0 e indica que β_i sí es estadisticamente significativo.

Para el caso del modelo 1, se observa que el coeficiente β_1 tiene un p valor menor a 2e-16, lo cual al ser menor que 0.04, se rechaza H_0, por lo que β_1 sí es significativo. Además para el caso del modelo 2, se aprecia que el coeficiente β_1 posee un p valor menor a 2e-16, lo cual es menor a 0.04 y provoca el rechazo de H_0 e indica que β_1 sí es estadisticamente significativo.

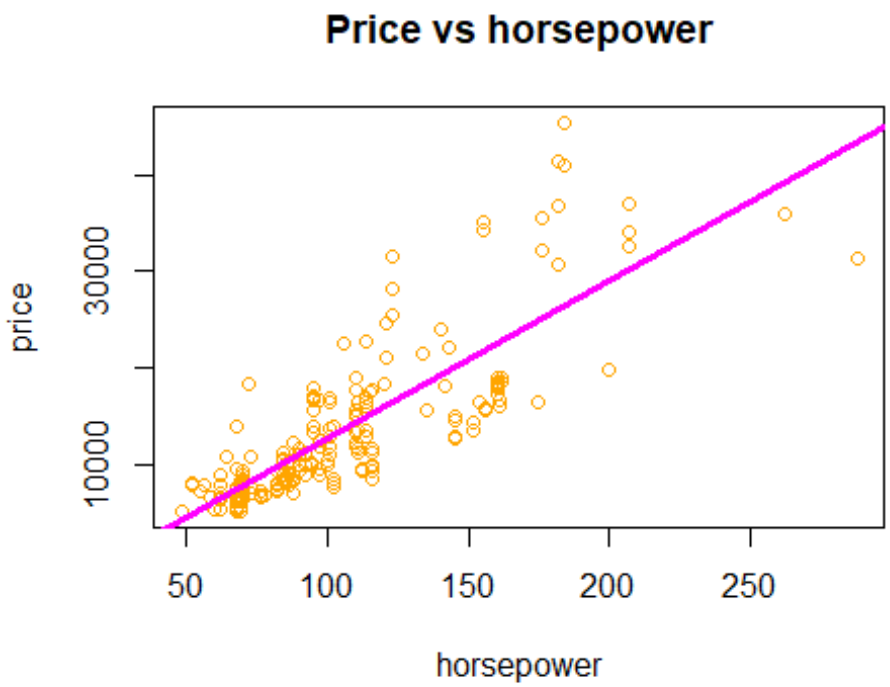
Además, para el modelo 1, el coeficiente de determinación ajustado es de 0.3306, mientras que para el modelo 2, dicho coeficiente tiene un valor de 0.6514, lo cual indica que el modelo 1 es capaz de explicar el 33.06% de la variabilidad total de los

datos, mientras que el modelo 2 es capaz de explicar el 65.14% de dicha variabilidad total, por lo que el modelo 2 explica un 32.08% más de variabilidad que el modelo 1.

Gráfica de dispersión por pares y recta de mejor ajuste

```
# Gráfico de dispersión de wheelbase contra price (modelo 1)

plot(grupo1$wheelbase, grupo1$price, col = "red",
     main = "Price vs wheelbase", xlab = "wheelbase", ylab = "price")

# Graficar recta del modelo de regresión 1 (wheelbase vs price)

abline(wb.model, col = "blue", lwd = 3)
```


```
# Gráfico de dipsersión de horsepower contra price

plot(grupo1$horsepower, grupo1$price, col = "orange",
     main = "Price vs horsepower", xlab = "horsepower", ylab = "price")

abline(hp.model, col = "magenta", lwd = 3)
```



En términos generales, en cuanto a todos los análisis realizados previamente, es posible concluir que en el contexto del problema a resolver, cabe mencionar que el hecho de que el modelo 1 (precio a partir de wheelbase) tenga un coeficiente de determinación R^2 mayormente pequeño (0.3306), significa que en realidad no es posible predecir con suficiente grado de confiabilidad el precio de los automóviles en función de la distancia entre sus ejes (wheelbase), mientras que los precios predichos en función de la cantidad de caballos de fuerza (horsepower) serán más confiables y precisos que aquellos basados en wheelbase, lo cual se respalda por el hecho de que en los gráficos de dipsersión y de las rectas de ambos modelos, se evidencia que los datos en el gráfico de horsepower vs price se ubican más cerca de la recta del modelo, mientras que por el contrario, en el gráfico de wheelbase vs price, dichos datos se encuentran más alejados de dicha recta principal, lo cual a su vez evidencia que en el caso de wheelbase y price, estas variables no tienen una correlación lineal fuerte entre ellas, mientras que en el caso de horsepower y price, éstas variables sí tienen correlación lineal considerablemente fuerte entre ellas, lo cual se evidenció en la matriz de correlación realizada previamente, donde wheelbase tiene correlación de 0.5778 con price, mientras que horsepower tiene correlación de 0.8081 con price, confirmando así que la asociación lineal más acertada es la de horsepower vs price.

Validez de modelos propuestos

Normalidad de residuos

Hipótesis:

H_0: los residuos del modelo provienen de una población normal.

H_1: los residuos del modelo no provienen de una población normal.

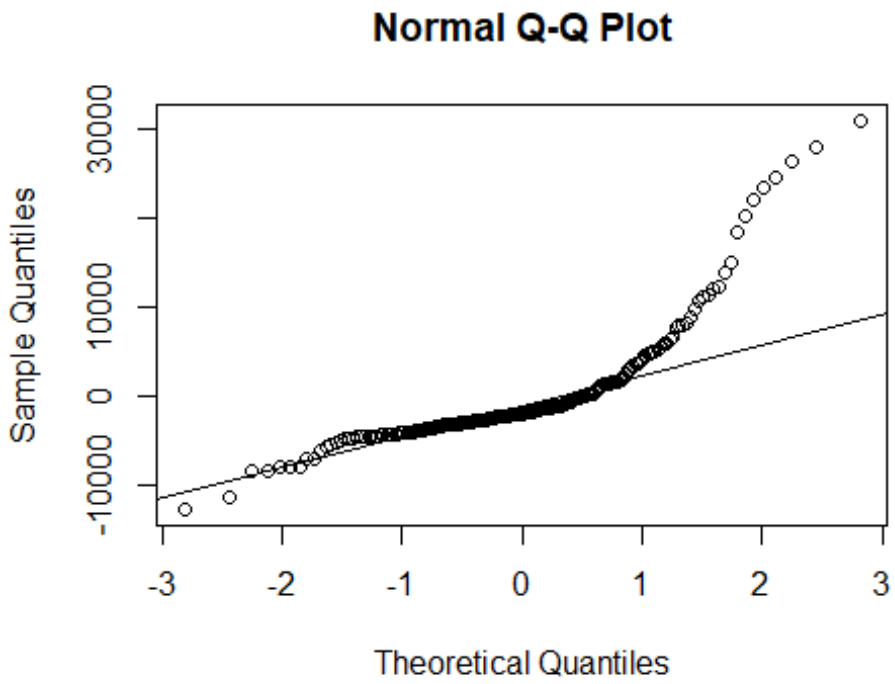
```
# Libreria para tests de normalidad

library(nortest)
```

Modelo 1: wheelbase vs price

```
# Anderson Darling test para normalidad de residuos del modelo 1

ad.test(wb.model$residuals)
```

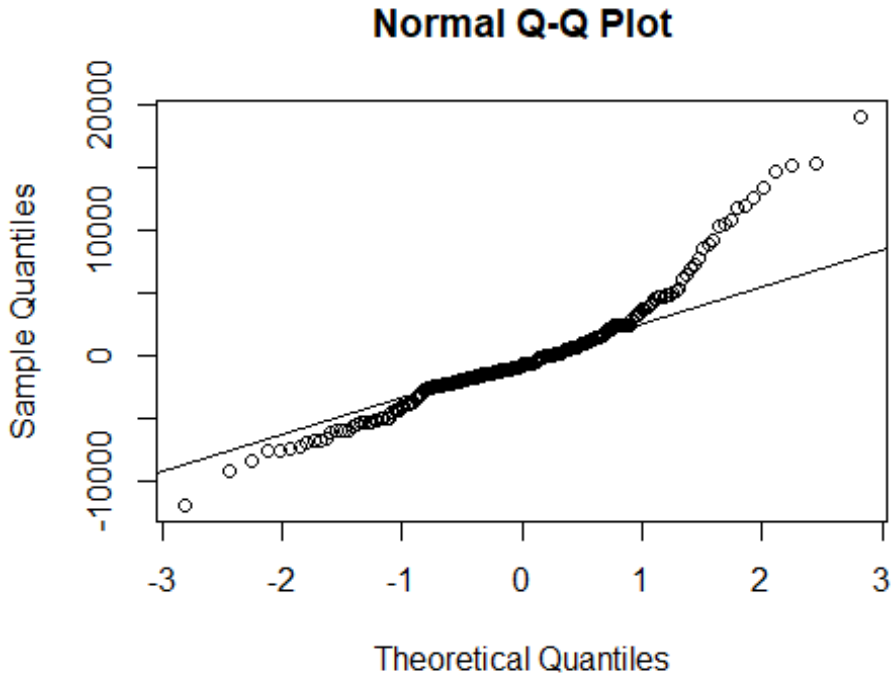
```
##
##  Anderson-Darling normality test
##
## data:  wb.model$residuals
## A = 15.605, p-value < 2.2e-16
```

```
# QQ-plot de los residuos del modelo 1

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



Modelo 2: horsepower vs price

```r
# Anderson Darling test para normalidad de residuos del modelo 2

ad.test(hp.model$residuals)

##
##  Anderson-Darling normality test
##
## data:  hp.model$residuals
## A = 4.8029, p-value = 6.267e-12

# QQ-plot de los residuos del modelo 1

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



Dado que se observa que en el caso de ambos modelos, el p valor del test de Anderson Darling es menor a 0.04, en concreto, menor a 2.2e-16 en el caso del modelo 1, e igual a 6.267e-12 en el caso del modelo 2, por lo cual, para ambos modelos se rechaza H_0, lo cual significa que los datos usados para la creación de dichos modelos no provienen de una población normal. Además de lo anterior, en el QQ-plot para cada modelo, se evidencia que los percentiles graficados se alejan de la recta de normalidad en los extremos de dicha recta, evidenciando el alejamiento de los datos respecto a la distribución normal que también se logra evidenciar en los tests de normalidad de Anderson Darling realizados con anterioridad.

Verificación de media 0

Hipótesis:

H_0: $\mu_R = 0$ (media de los residuos es igual a 0)

H_1: $\mu_R \neq 0$ (media de los residuos no es igual a 0)

```
# verificacion de media 0 para residuos del modelo 1

t.test(wb.model$residuals)

##
##  One Sample t-test
##
## data:  wb.model$residuals
## t = -1.4537e-16, df = 204, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -897.8812  897.8812
## sample estimates:
##    mean of x
## -6.62007e-14

# verificacion de media 0 para residuos del modelo 2

t.test(hp.model$residuals)

##
##  One Sample t-test
##
## data:  hp.model$residuals
## t = -2.2725e-16, df = 204, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -647.9614  647.9614
## sample estimates:
##     mean of x
## -7.468281e-14
```

En los resultados de los t tests se observa que el p valor resultante de las pruebas para ambos modelos es igual a 1, lo cual al ser mayor que 0.04, se tiene evidencia estadística suficiente para no rechazar H_0 en el caso de ambos modelos, motivo por el cual, para ambos modelos es posible afirmar que la media de sus residuos es igual a 0.

Homcedasticidad, linealidad, independencia

Homocedasticidad:

Hipótesis a probar:

H_0: la varianza de los residuos es constante.

H_1: la varianza de los residuos no es constante.

Independencia:

H_0: Los residuos no están correlacionados (sí son independientes).

H_1: Los residuos sí están correlacionados (no son independientes).

Linealidad:

H_0: No hay términos omitidos que indiquen linealidad.

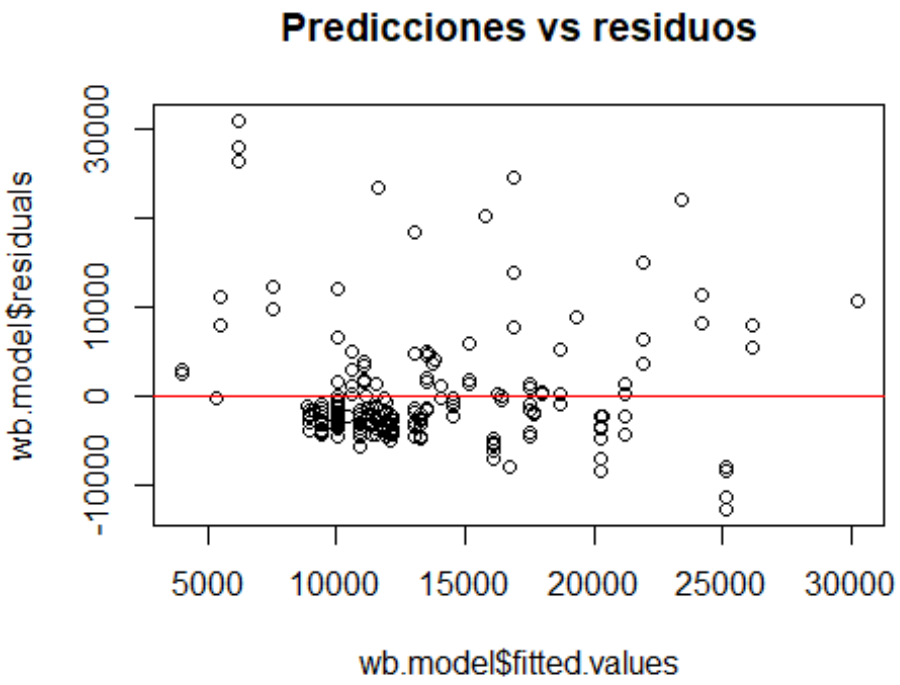
H_1: Hay una especificación errónea en el modelo que indica no linealidad.

```
# Biblioteca para pruebas de homocedasticidad e independencia

library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

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

Modelo 1: wheelbase vs price

Homocedasticidad

```
# Gráfica de predicciones vs residuos modelo 1

plot(wb.model$fitted.values, wb.model$residuals, main = "Predicciones vs
residuos")

# Trazar recta horiozntal en la media 0 de los residuos

abline(h = 0, col = "red")
```




```
# Prueba de White para validar homocedasticidad

gqtest(wb.model)

##
##  Goldfeld-Quandt test
##
## data:  wb.model
## GQ = 1.1934, df1 = 101, df2 = 100, p-value = 0.1886
## alternative hypothesis: variance increases from segment 1 to 2
```

De acuerdo al gráfico de predicciones vs residuos del modelo, se observa que los residuos se encuentran dispersos a lo largo del gráfico de una manera aproximadamente similar, dado que no se observa que la varianza de los residuos comience a incrementar o disminuir de forma evidente conforme incrementan las predicciones, sino que los residuos permanecen sin evidenciar una tendencia específica en su varianza a medida que varían las predicciones del modelo, por lo que en base al gráfico se puede afirmar que la varianza de los residuos es mayormente constante, además, de acuerdo con la prueba de homocedasticidad de White, se aprecia que el p valor del test es igual a 0.1886, lo cual al ser mayor que 0.04, se tiene evidencia estadística para no rechazar H_0, lo cual implica que los residuos del modelo tienen varianza constante, es decir, presentan homocedasticidad entre ellos.

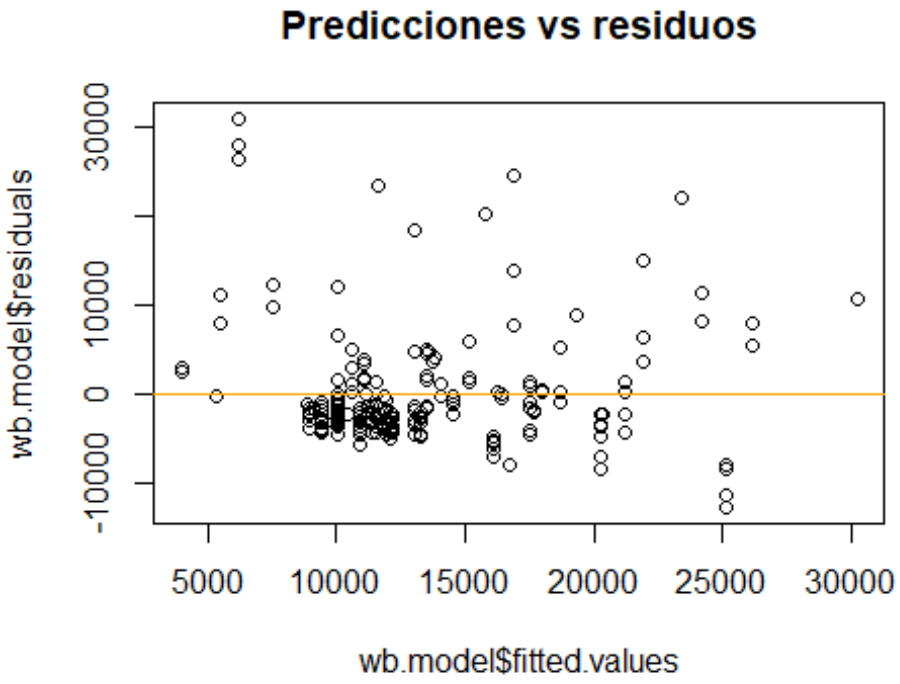
Independencia de residuos

```
# Realizar test de independencia de Durbin Watson para verificar
independencia de residuos
```

```r
dwtest(wb.model)
```

```
##
##  Durbin-Watson test
##
## data:  wb.model
## DW = 0.56645, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
# Graficar predicciones del modelo vs sus residuos para observar posibles tendencias

plot(wb.model$fitted.values, wb.model$residuals, main = "Predicciones vs residuos")

abline(h = 0, col = "orange")
```



En el gráfico de predicciones vs residuos del modelo 1, se aprecia que todos los residuos en el gráfico se encuentran distribuidos de forma que no siguen una tendencia o patrón en particular, sino que más bien se distribuyen de forma mayormente aleatoria a lo largo del gráfico, además de esto, en el resultado del test de independencia de Durbin Watson se puede apreciar que el p valor del test es inferior a 2.2e-16, lo cual al ser menor que 0.04, se tiene evidencia estadística para rechazar H_0, lo cual significa que los residuos del modelo 1 sí están correlacionados, es decir, que no son independientes entre sí.

Linealidad

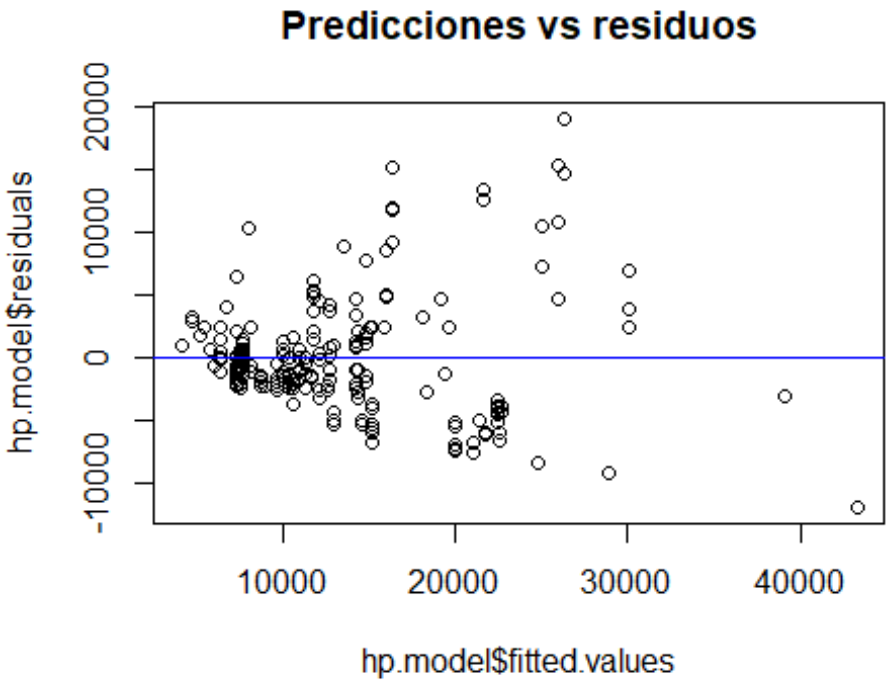
```
# Aplicar prueba de RESET de Ramsey para comprobar linealidad en los
datos del modelo

resettest(wb.model)

##
##  RESET test
##
## data:  wb.model
## RESET = 10.65, df1 = 2, df2 = 201, p-value = 4.016e-05
```

De acuerdo con la prueba de RESET de Ramsey para linealidad, se observa que el p valor resultante de la misma es igual a 4.016e-05, lo cual es menor que 0.04, por lo que se tiene suficiente evidencia estadística para rechazar H_0, motivo por el cual, se puede afirmar que hay una especificación errónea en el modelo que indica no linealidad, en otras palabras, que la naturaleza de los datos a partir de los cuales se generó dicho modelo no siguen un patrón lineal, motivo por el cual, la linealidad no está presente en los datos del modelo.

Modelo 2: horsepower vs price

Homocedasticidad

```
# Gráfica de predicciones vs residuos modelo 2

plot(hp.model$fitted.values, hp.model$residuals, main = "Predicciones vs
residuos")

# Trazar recta horizontal en la media 0 de los residuos

abline(h = 0, col = "blue")
```



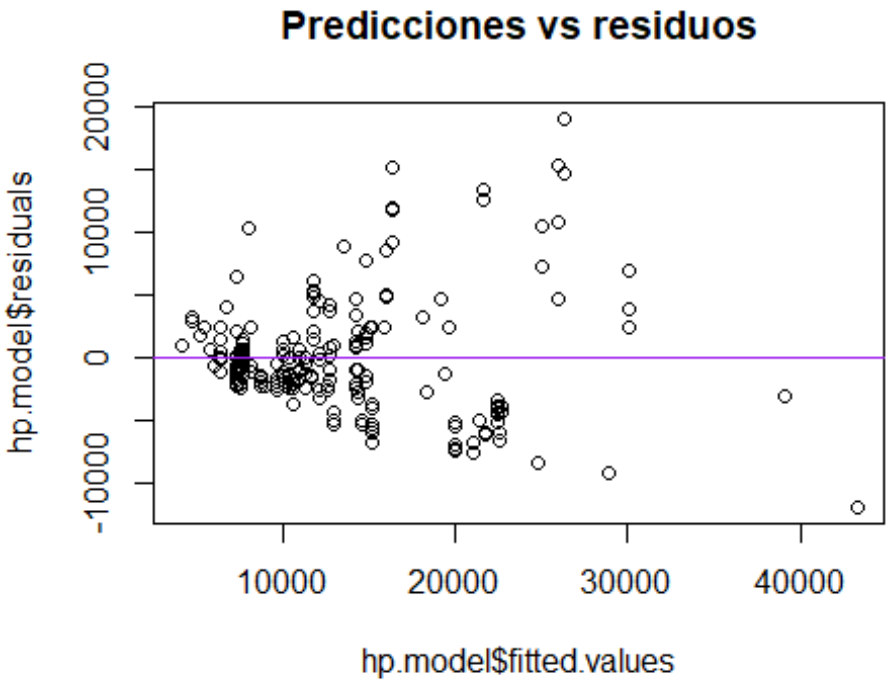

```
# Prueba de White para validar homocedasticidad

gqtest(hp.model)

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

En general se observa en el gráfico de predicciones contra residuos del modelo 2, que los residuos del modelo se encuentran dispersos por todo el gráfico de una forma mayormente similar en el sentido de que no evidencian alguna tendencia o patrón específico, por lo cual se puede afirmar que la varianza de los residuos se mantiene todo el tiempo casi igual con cambios mínimos, además, de acuerdo con los resultados del test de homocedasticidad de White para el caso del modelo 2, se aprecia que el p valor de la prueba es de 1, lo cual al ser mayor que la significancia de 0.04, se tiene evidencia estadística suficiente para no rechazar H_0, lo cual indica que los residuos del modelo 2 tienen varianza constante (homocedasticidad).

Independencia de residuos

```
# Realizar test de independencia de Durbin Watson para verificar
independencia de residuos

dwtest(hp.model)
```

```
##
##  Durbin-Watson test
##
## data:  hp.model
## DW = 0.79229, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0

# Graficar predicciones del modelo vs sus residuos para observar posibles tendencias

plot(hp.model$fitted.values, hp.model$residuals, main = "Predicciones vs residuos")

abline(h = 0, col = "purple")
```



De acuerdo con el gráfico de predicciones vs residuos se aprecia que en dicho gráfico, los residuos graficados se encuentran dispersos a lo largo de toda el área del gráfico sin evidenciar algún patrón o tendencia en particular a medida que cambian o varían las predicciones, lo cual sugiere que los residuos del modelo 2 son independientes entre sí, no obstante, de acuerdo con la prueba de independencia de Durbin Watson, es posible observar que el valor p del test es inferior a 2.2e-16, por lo que se tiene suficiente evidencia estadística para rechazar H_0 para independencia, lo cual significa que los residuos del modelo 2 sí están correlacionados y por tanto no son independientes.

Linealidad

```r
# Aplicar prueba de RESET de Ramsey para comprobar linealidad en los
datos del modelo 2

resettest(hp.model)

##
##  RESET test
##
## data:  hp.model
## RESET = 5.1766, df1 = 2, df2 = 201, p-value = 0.006424
```

En el test de RESET de Ramsey para linealidad, se puede apreciar que el p valor del test es igual a 0.006424, lo cual es menor que 0.04, por lo cual se tiene suficiente evidencia estadística para rechazar H_0, lo cual implica que existe una especificación errónea en el modelo 2 que indica no linealidad entre sus datos, lo cual quiere decir que los datos del modelo no presentan un patrón o tendencia de tipo lineal.

Interpretación de los análisis realizados

Después de todo el análisis anterior, es posible concluir hasta este punto que ambos de los modelos de regresión lineal propuestos presentan cierta correlación con los datos originales para predecir el precio de los automóviles, no obstante, es importante señalar que ambos modelos no cumplen con al menos 1 de los supuestos necesarios para garantizar su óptima capacidad predictiva y por tanto la calidad y confiabilidad de sus predicciones, por lo que a pesar de que en particular el modelo 2 sea capaz de explicar el mayor porcentaje posible de la variabilidad total de los datos, aún así será necesario considerar que dicho modelo no cumple con el supuesto de linealidad, además de aquel referente a la normalidad de residuos, por lo cual a pesar de tomar en cuenta al modelo 2 como al "mejor" modelo entre ambos analizados, será necesario tomar en cuenta que sus predicciones no serán totalmente confiables, por lo cual será necesario implementar otra clase de modelos no lineales para verificar si éstos se adaptan de mejor forma a los datos iniciales.

Conclusión final sobre mejor modelo de regresión lineal

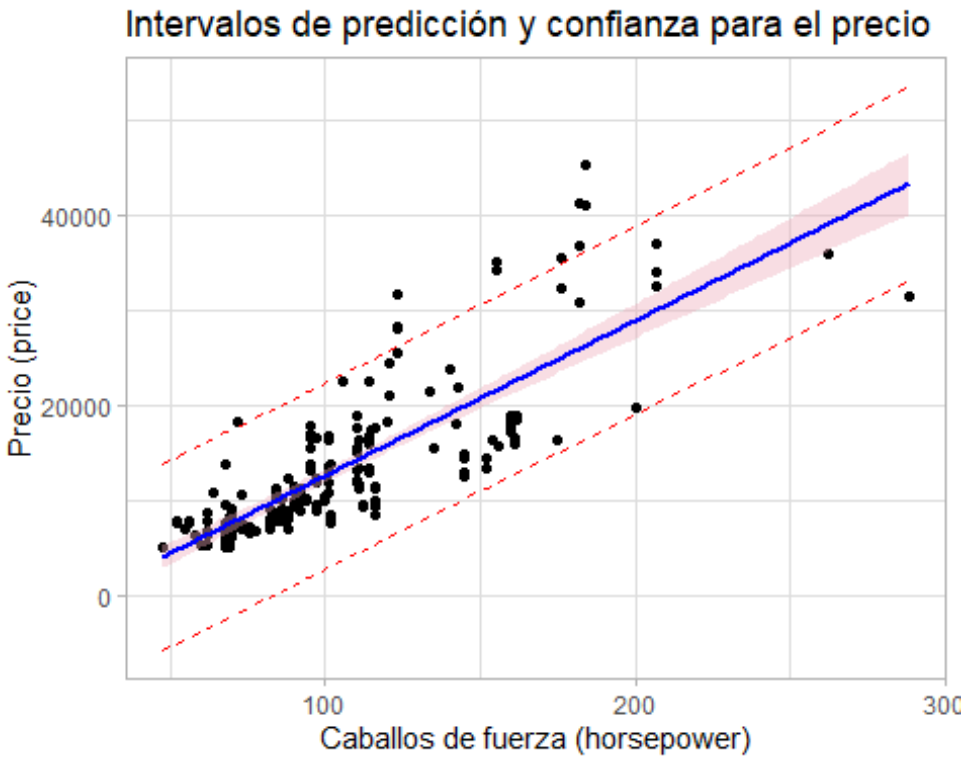
Finalmente, a manera de conclusión de todo el análisis previo, es posible afirmar que el mejor modelo para representar a los datos originales es el modelo 2, debido a que a pesar de que ambos modelos tanto el 1 como el 2 no cumplan con los supuestos de normalidad, independencia y linealidad, el modelo 2 de horsepower contra precio consigue explicar el mayor porcentaje de la variabilidad total de los datos originales, siendo éste mismo del 65.14%, contra un 33.06% de variabilidad explicada por el modelo 1 (wheelbase vs price), por lo cual, el modelo 2 resulta ser ligeramente mejor que el modelo 1, es decir que la mejor variable para predecir el precio de los automóviles es la cantidad de caballos de fuerza (horsepower) que tengan en lugar de la distancia entre sus ejes (wheelbase), por lo que en última instancia, se elige el modelo 2 (hp.model) como el mejor. Además, la variable que influye más en el precio del auto es la cantidad de caballos de fuerza que tenga el auto, dado que a mayor cantidad de caballos de fuerza, significa que el auto tiene un motor más potente y

como consecuencia también presenta un rendimiento de mejor calidad al momento de conducirlo en las autopistas, por lo que en general el automóvil será de una mayor calidad, lo que conduce a que su precio aumente.

Intervalos de predicción y confianza

Intervalos de confianza y predicción para variable precio

```
# Librería ggplot2 para graficos personalizados

library(ggplot2)

# Graficar intervalos de predicción para el precio de los autos en
función de los
# caballos de fuerza (horsepower)

suppressWarnings({

Ip = predict(object = hp.model, interval = "prediction", level = 0.96)

datos = cbind(grupo1[, -2], Ip)

ggplot(datos, aes(x = grupo1$horsepower, y = grupo1$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 = "Intervalos de predicción y confianza para el precio",
  x = "Caballos de fuerza (horsepower)",
  y = "Precio (price)"
) +
theme_light()})
```



Nota: los intervalos de confianza para el precio están representados por una pequeña área en color rosado que rodea a la recta color azul que representa al modelo de regresión, mientras que de forma similar, los intervalos de predicción también para el precio se representan en forma de líneas punteadas en color rojo, mismas que forman un área más grande que el área en color rosado que representa a los intervalos de confianza, lo cual simboliza que las predicciones derivadas del modelo tienen un margen de variación considerablemente mayor al de la media de la variable precio, lo cual es el caso ideal, es decir, que los intervalos de predicción deben ser mayores que los de confianza.

Separación de datos por categoría gas de tipo de combustible (fueltype)

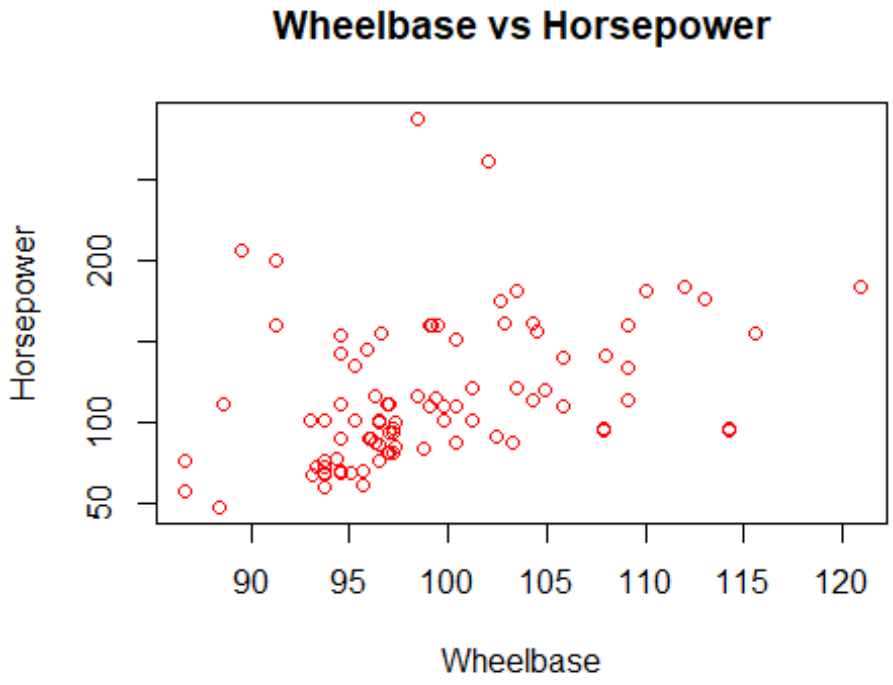
```
# Seleccionar categoría más relevante de la variable categoríca del grupo
fueltype: gas,
# debido a que posee la mayor frecuencia absoluta (se repite una mayor
cantidad de veces),
# lo cual podrá ayudar a saber con precisión el patrón o tendencia que
sigan éste tipo de
# datos específicamente y que podría estar influyendo de forma
significativa en el precio
# de los autos

# Separar los datos principales en otro subconjunto en base a la
categoría gas de la
# variable categórica fueltype

datos_gas = grupo1[grupo1$fueltype == "gas", ]
```

```r
head(datos_gas)
```

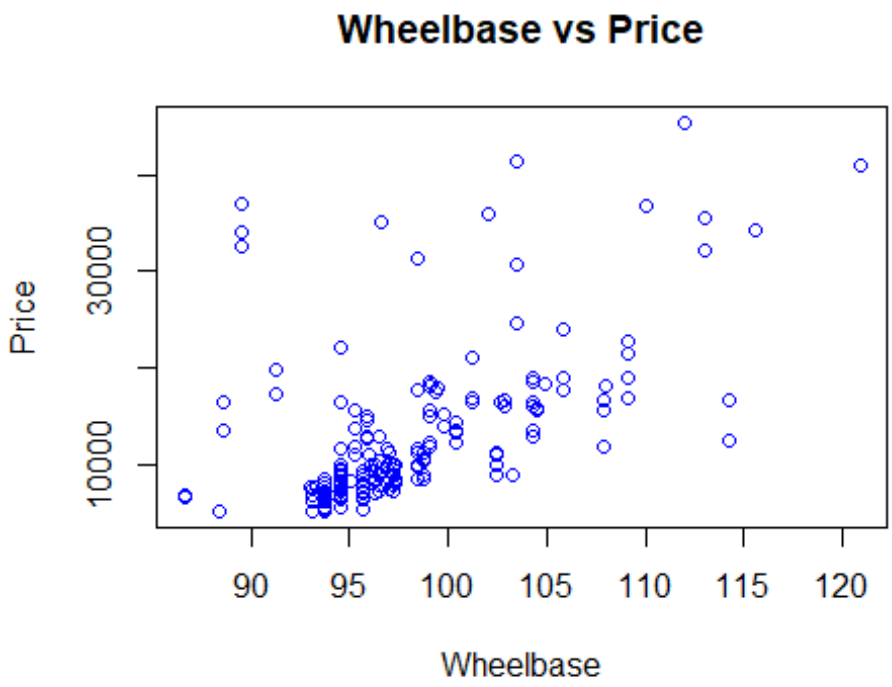
```
##   wheelbase fueltype horsepower price
## 1      88.6      gas        111 13495
## 2      88.6      gas        111 16500
## 3      94.5      gas        154 16500
## 4      99.8      gas        102 13950
## 5      99.4      gas        115 17450
## 6      99.8      gas        110 15250
```

Gráficas por pares de variables numéricas

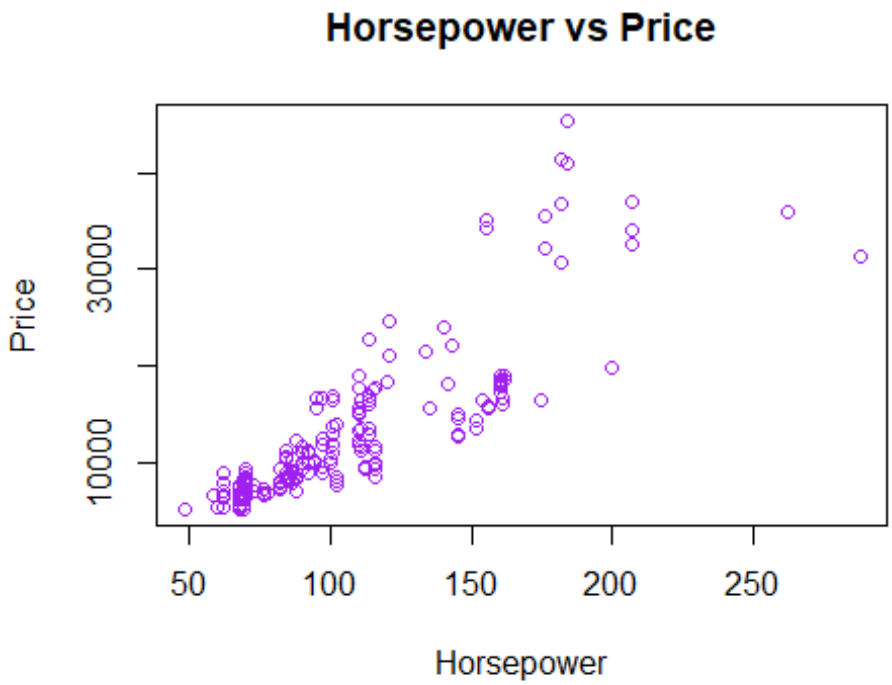
```r
# Gráfico de wheelbase contra horsepower

plot(datos_gas$wheelbase, datos_gas$horsepower, main = "Wheelbase vs
Horsepower",
     xlab = "Wheelbase", ylab = "Horsepower", col = "red")
```




```r
# Gráfico de wheelbase contra price

plot(datos_gas$wheelbase, datos_gas$price, main = "Wheelbase vs Price",
     xlab = "Wheelbase", ylab = "Price", col = "blue")
```




```r
# Gráfico de horsepower vs price

plot(datos_gas$horsepower, datos_gas$price, main = "Horsepower vs Price",
     xlab = "Horsepower", ylab = "Price", col = "purple")
```



Interpretación en el contexto del problema

En términos generales, es posible observar que al graficar los intervalos de predicción y de confianza para la variable precio en base al mejor modelo de regresión seleccionado, ciertos puntos en el gráfico se salen de las líneas punteadas en color rojo (intervalos de predicción), lo cual indica que dichos datos pueden ser posibles datos atípicos que es necesario analizar con mayor detenimiento para determinar su posible grado de influencia en el ajuste del modelo seleccionado a los datos originales del precio de los autos, además lo mencionado anteriormente evidencia que en realidad el "mejor" modelo obtenido no se termina de ajustar adecuadamente a los datos en cuestión, motivo por el cual, será necesario a futuro probar modelos predictivos de otros tipos, por ejemplo no lineales para evaluar el rendimiento de los mismos y comprobar si la precisión y confiabilidad de sus predicciones resulta ser significativamente mayor a las de las predicciones derivadas de un modelo de regresión específicamente lineal.

Más allá

¿Propondrías una nueva agrupación de las variables a la empresa automovilística?

Tomando en cuenta todo el análisis realizado con anterioridad, sí será necesario proponer una nueva agrupación de variables a la empresa automovilística, principalmente debido a que en las pruebas de validación de los supuestos realizadas, se obtuvo como resultado que ambos modelos generados a partir del grupo inicial de

variables establecido no cumplieron con todos los supuestos de un modelo de regresión lineal, ya que dichos modelos no cumplieron entre esos supuestos, la normalidad de sus residuos, la independencia de los mismos, además de la linealidad de los datos con los que fueron creados, motivo por el cual, eso conduce a que las predicciones derivadas de ambos modelos propuestos no sean lo suficientemente confiables, por lo que en caso de usar las variables en el grupo establecidas por la empresa, los precios predichos de los automóviles no serán confiables y por consiguiente, eso puede conducir a que la empresa experimente a corto, mediano o largo plazo, pérdidas monetarias importantes, lo que reduciría su capacidad para adquirir nueva materia prima para seguir con todos sus procesos internos de fabricación de automóviles, conduciendo en última instancia a que la empresa entre en un estado de bancarrota, pudiendo conducir a su cierre definitivo.

Retoma todas las variables y haz un análisis estadístico muy leve (medias y correlación) de cómo crees que se deberían agrupar para analizarlas.

Finalmente, una agrupación alternativa de variables para intentar mejorar los modelos predictivos del precio de los autos, contempla a las variables: caballos de fuerza (horsepower), tamaño del motor del auto (enginesize), además del lanzamiento del pistón (stroke), por lo cual, a continuación se procederá a realizar un análisis estadístico leve para ilustrar los aspectos esenciales de la naturaleza de estas nuevas variables propuestas:

Cálculo de media por variable

```
# Calcular la media por cada variable nueva propuesta anteriormente

medias_nuevas = c(horsepower = mean(autos$horsepower),
                  enginesize = mean(autos$enginesize),
                  stroke = mean(autos$stroke))

print("Las medias por cada variable nueva son:")

## [1] "Las medias por cada variable nueva son:"

medias_nuevas

## horsepower enginesize     stroke
## 104.117073 126.907317   3.255415
```

Cálculo de niveles de correlación lineal entre las nuevas variables

```
# Cálculo de la matriz de correlación entre las nuevas variables
propuestas

correlations_new = data.frame(cor(cbind(autos$horsepower,
autos$enginesize,
                                        autos$stroke, autos$price)))

# Establecer nombres de las columnas y filas para el dataframe de
```

```r
correlaciones entre las
# nuevas variables propuestas

colnames(correlations_new) = c("Horsepower", "EngineSize",
                               "Stroke", "Price") # nombres de columnas

row.names(correlations_new) = c("Horsepower", "EngineSize",
                                "Stroke", "Price") # nombres de filas

correlations_new
```

```
##            Horsepower EngineSize     Stroke      Price
## Horsepower 1.00000000  0.8097687 0.08093954 0.80813882
## EngineSize 0.80976865  1.0000000 0.20312859 0.87414480
## Stroke     0.08093954  0.2031286 1.00000000 0.07944308
## Price      0.80813882  0.8741448 0.07944308 1.00000000
```

Por último, en la matriz de correlación realizada considerando las nuevas variables propuestas anteriormente junto con la variable de respuesta (precio), es posible observar que la mayor correlación de todas es incluso mayor que la correlación máxima obtenida con las variables anteriores, dado que con las variables previas la máxima correlación existente era de 0.8081, mientras que con las nuevas variables, dicha correlación máxima asciende a 0.8741, poe lo cual sí existe una mejoría al cambiar el agrupamiento de variables propuestas inicialmente por la empresa, por lo cual será necesario generar nuevos modelos de regresión lineal para predecir el precio de los autos, principalmente entre las variables: precio en función del tamaño del motor (EngineSize), además de contruir otro modelo que contemple a las variables precio en función de la cantidad de caballos de fuerza que ya se realizó previamente, por lo cual, lo nuevo en esta ocasión es generar otro modelo de regresión lineal, para predecir el precio de los automóviles en función del tamaño de su motor, lo cual podrá dar predicciones más precisas y confiables que predecir el precio en función de los caballos de fuerza, por lo cual se corre menos riesgo de que la empresa automotriz comience a tener pérdidas monetarias significativas que puedan afectar su capacidad de producción.