Regresión Lineal Simple

Equipo X

12 May, 2023

Contents

1 Introducción

1.1 Problema de interés: Análisis del precio de los vehículos [Raúl]

1.1.1 Breve explicación de la base de datos "Scrap price" [Raúl]

1.2 ¿Por qué usar una regresión lineal? [Raúl]

2 Marco teórico [AOKI]

2.1 Conceptos básicos

2.2 Supuestos del modelo

2.3 Método de selección de variables y limitaciones del modelo

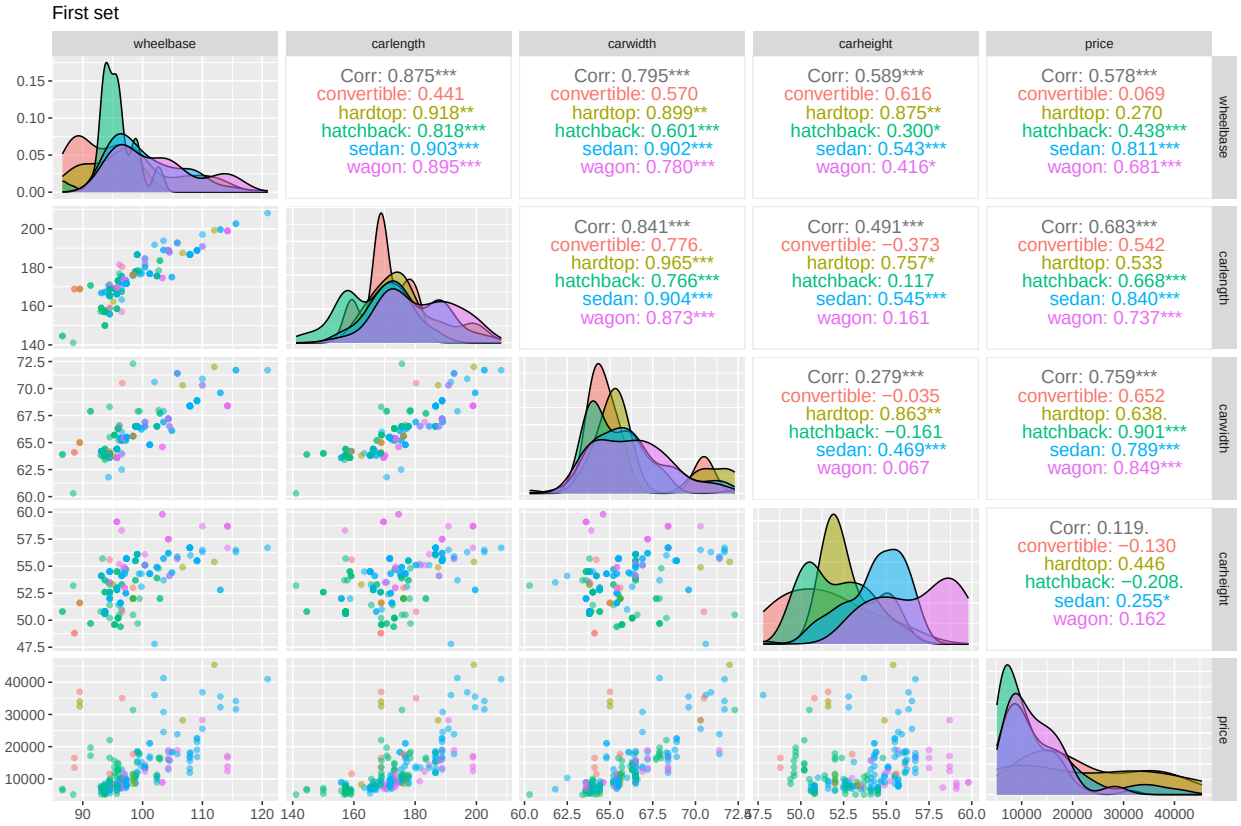
3 Análisis exploratorio de datos

3.1 Análisis de la base de datos

Aquí va filtros aplicados, estadígrafos, gráficas, limpieza y selección de variable depen- diente con justificación

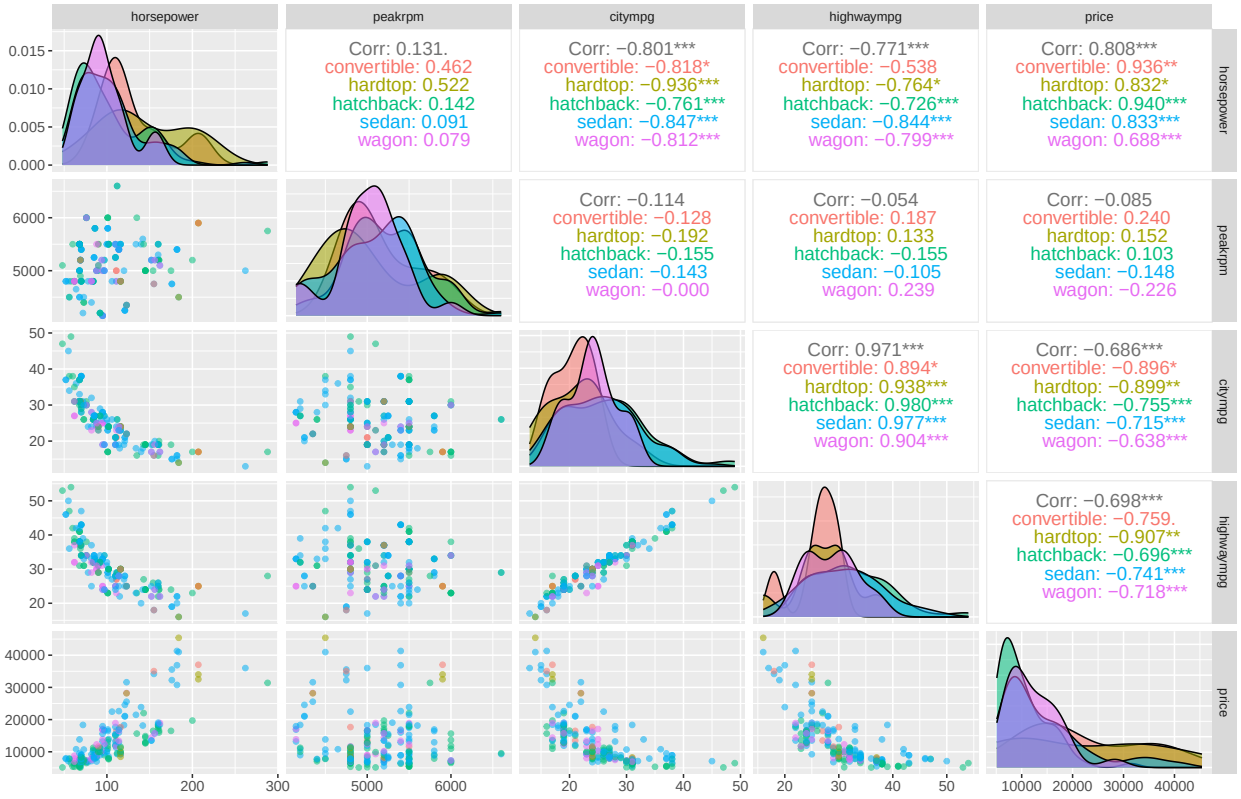
3.2 Selección de la variable explicativa

Para la selección de la variable explicativa eliminamos las variables de caracter, y dejamos las variables numéricas. Dividimos las variables en 3 grupos para poder analizar la correlación de las variables respecto el precio:



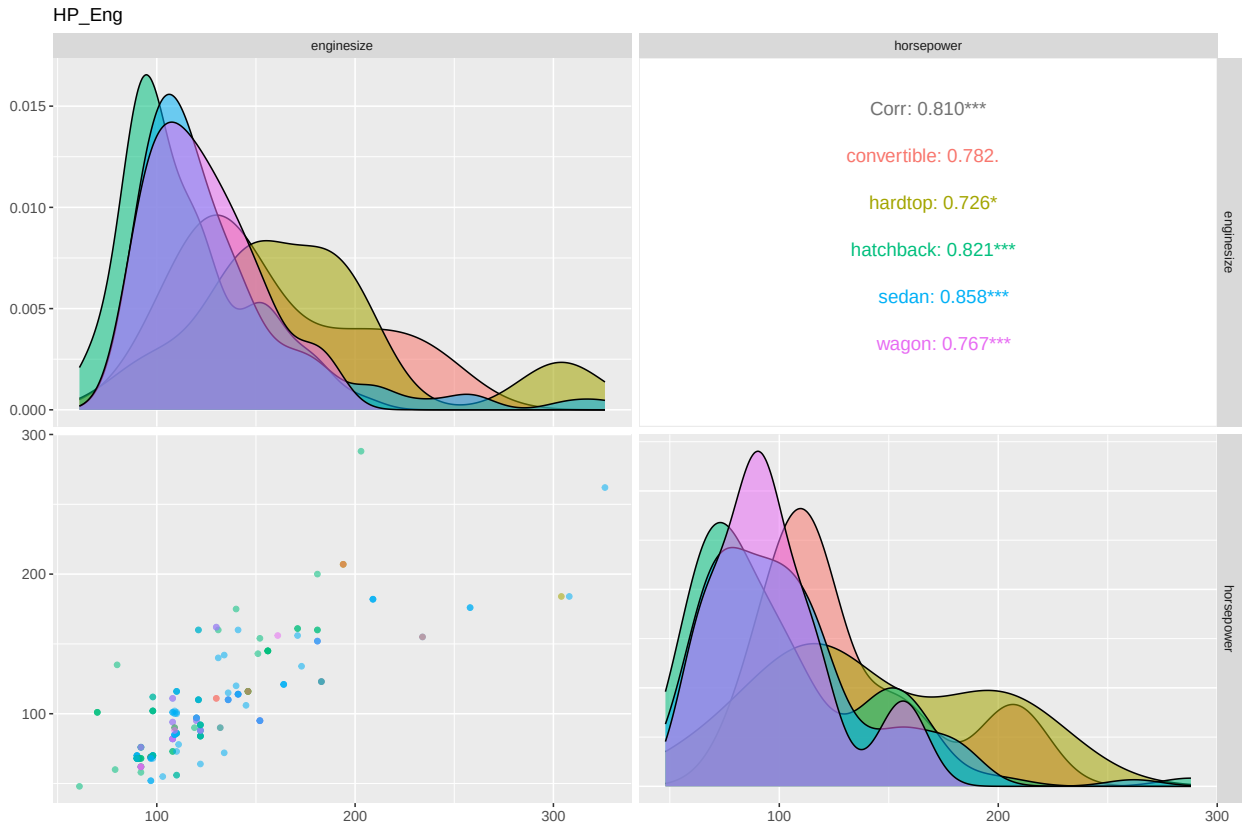
Second set

Third set



Cómo se puede observar las variables con las correlaciones más altas son:

Horsepower, curbweight, enginesize con 0.808, 0.835 y 0.874 respectivamente. No obstante, la variable que más sentido hace para elegir para explicar el motor es “enginesize”, ya que además de tener la correlación más alta respecto al precio, podemos eliminar “horse power” porque tiene una multicolienalidad imperfecta de 0.810 con la variable que elegimos.



4 Modelo de regresión lineal simple

4.1 Parámetros del modelo

En un modelo de regresión, los parámetros son los valores que se ajustan al conjunto de datos para crear la mejor línea o curva que represente la relación entre la variable independiente y la variable dependiente.

En una regresión lineal simple, los parámetros son la pendiente y la intersección en el eje. La pendiente representa el cambio en la variable dependiente por cada cambio unitario en la variable independiente, mientras que la intersección en el eje y representa el valor de la variable dependiente cuando la variable independiente es igual a cero.

El objetivo de un modelo de regresión es encontrar los valores óptimos de los parámetros que minimicen la diferencia entre las predicciones del modelo y los valores reales de la variable dependiente en el conjunto de datos.

```
Call:
lm(formula = price ~ enginesize, data = train.base)

Residuals:
     Min       1Q   Median       3Q      Max
-10487.4  -2314.5   -566.6   1664.7  14439.1

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8156.662   1022.821  -7.975 4.59e-13 ***
enginesize    167.620      7.497  22.359  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4024 on 142 degrees of freedom
Multiple R-squared:  0.7788,    Adjusted R-squared:  0.7772
F-statistic: 499.9 on 1 and 142 DF,  p-value: < 2.2e-16
```

Los valores de los parámetros son B0= -8156.662 y B1= 167.62 con unos errores estándar de 1022.821 y 7.497 respectivamente.

Podemos observar que para B0 y B1 el P value < |t value | por lo tanto B0 y B1 son significativas con un nivel de confianza de 1

Tabla ANOVA

```
Analysis of Variance Table

Response: price
            Df     Sum Sq    Mean Sq F value    Pr(>F)
enginesize   1 8096361095 8096361095  499.94 < 2.2e-16 ***
Residuals  142 2299624940   16194542
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.2 Análisis de residuales

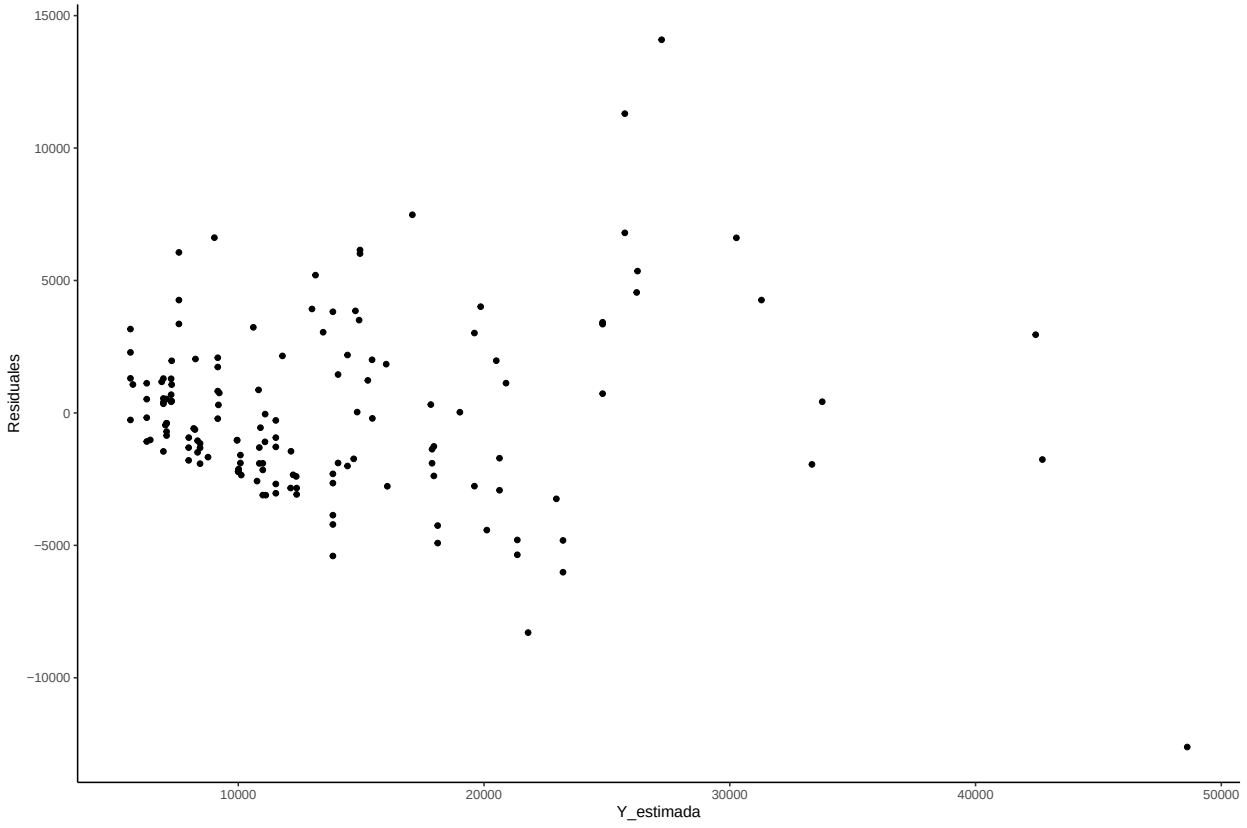
4.2.1 Comprobación de la linealidad de la Fn de regresión

Comprobamos con la R^2, en este caso los errores se acercan un 77% a nuestra recta de regresión lo que nos dice que sí hay linealidad en ella, lo comprobamos sacando la R^2

nos da el .7787968 de R^2. Esto quiere decir que la variable X explica en un 77.88% a la variable dependiente Y.

4.2.2 Heterocedasticidad

Comprobamos heterocedisticidad (la varianza de los errores es constante), lo comprobamos con un gráfico comparando los residuales con las Y observadas ($\hat{y}$), para esto tenemos que hacer un DF con ambos vectores obtenidos de nuestro modelo



Se puede observar que no hay un patrón en sí en el gráfico, como una recta, con esto podemos asumir que hay heterocedisticidad.

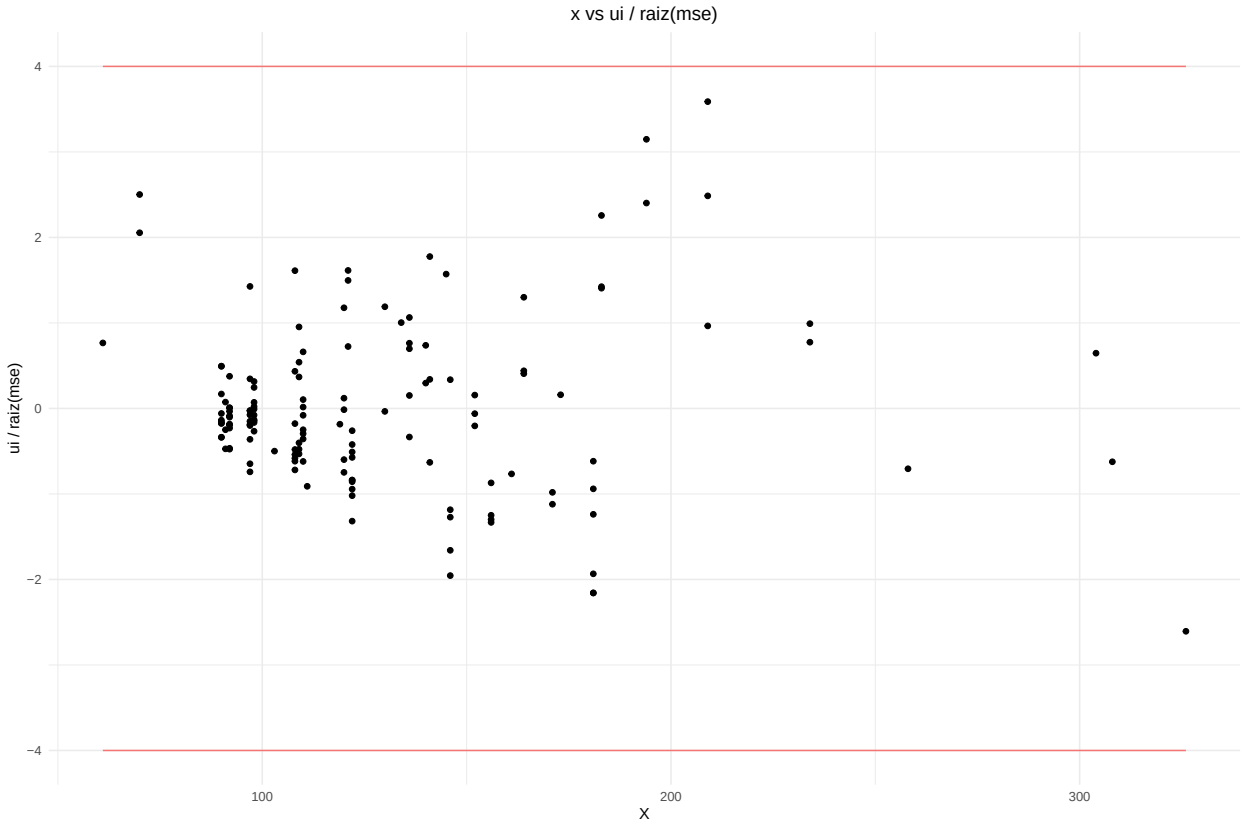
4.2.3 Independencia en los errores

No es una serie de tiempo - no aplica ya que los datos no llevan un orden y pueden cambiar de pocisión

4.2.4 Presencia de errores atípicos

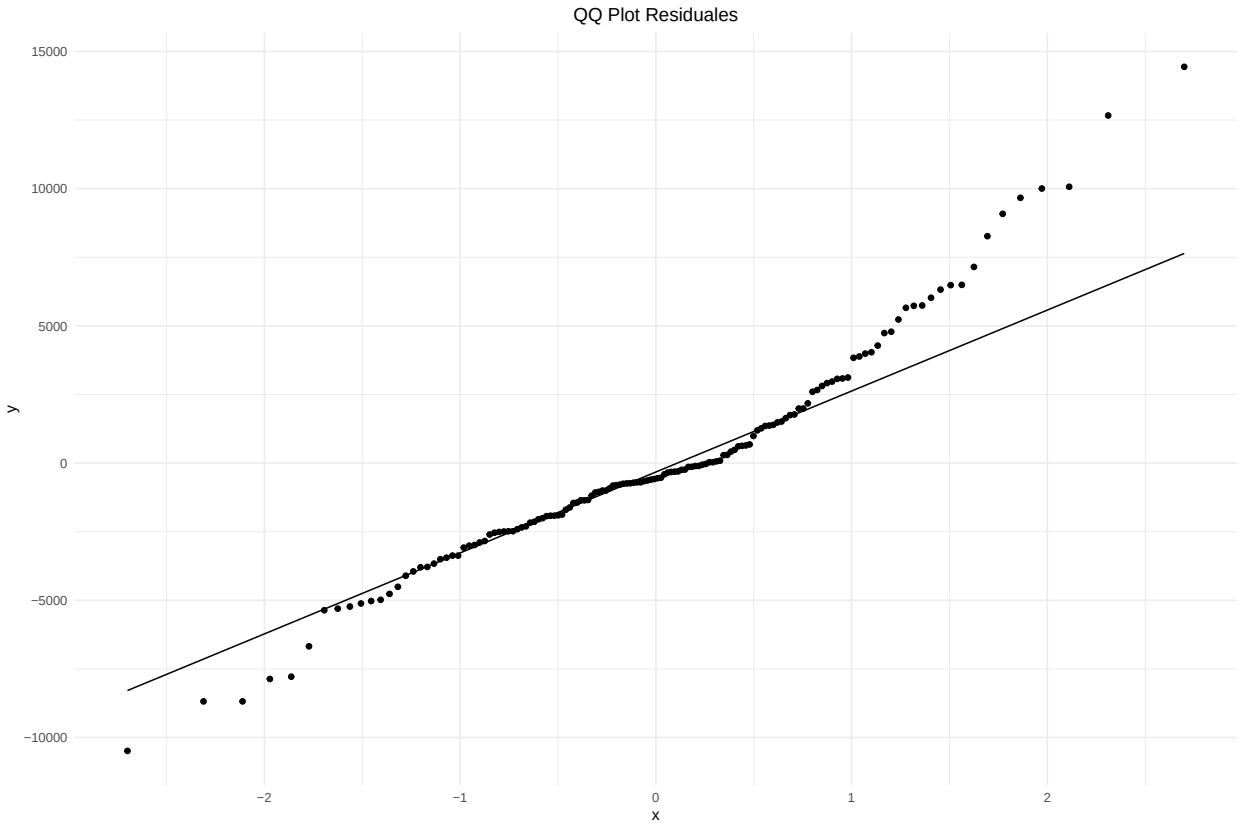
Esto se hace calculando la raíz de el cuadrado medio de la suma de cuadrados de los errores (MSE), el cual se obtiene de la tabala ANOVA de nuestros residuales el mean.

Ya con MSEˆ(1/2), podemos sacar la división de los residuales entre la raíz de MSE y compararlos con xi esas variables las metemos en un DF y graficamos las diferencias.



4.2.5 Verificar la normalidad en los errores

La QQ plot - esa se hace con los residuales sacamos los residuales de nuestro modelo

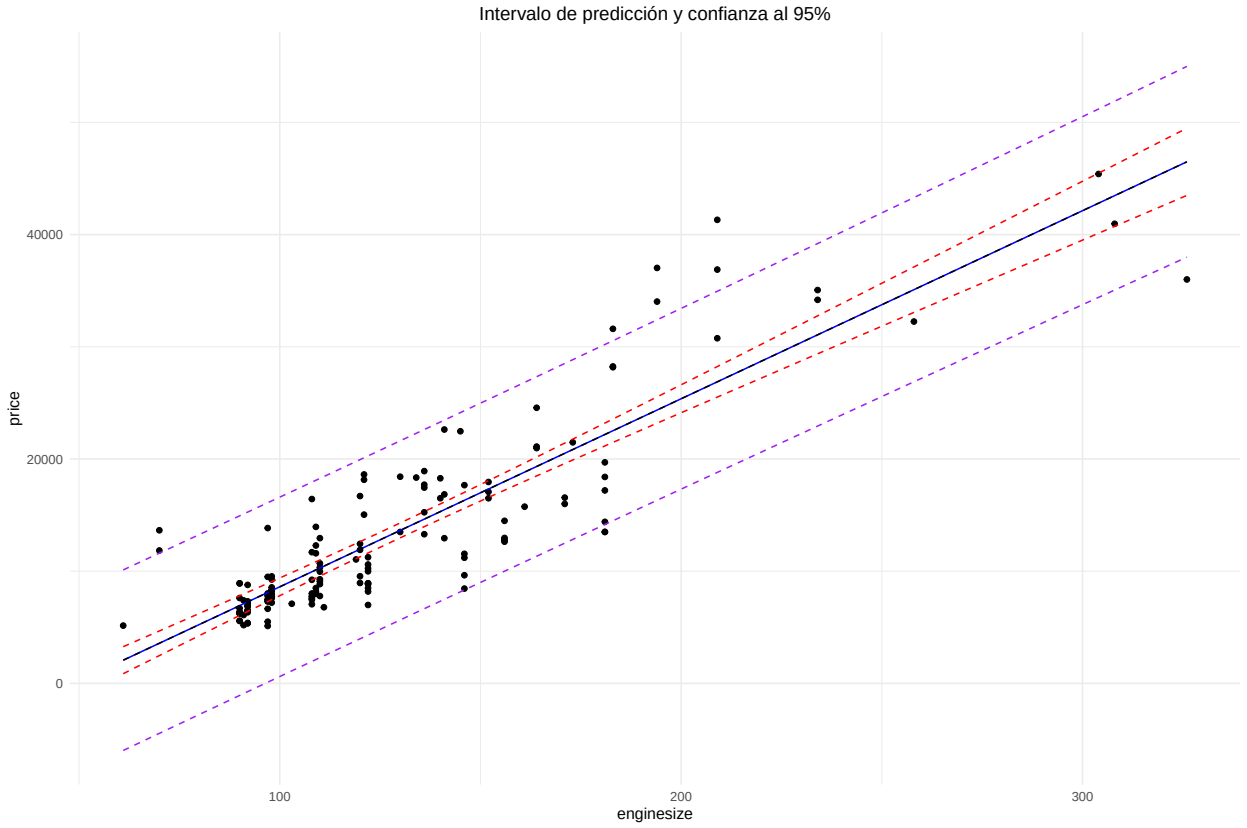


Rechazamos el supuesto de normalidad de los errores debido a las dos colas que muestra el gráfico de QQ plot. No obstante, ya que el modelo de regresión lineal simple ajustado es robusto ante el supuesto de normalidad podemos continuar usando esta variable explicativa.

4.3 Intervalo de confianza y predicción al 95%

Sacamos el intervalo de confianza de E[Y], esto lo hacemos para ver el intervalo en donde van a estar las siguientes E[Y / X], independientemente de la muestra.

En R usamos la fn Predict.lm la alimentamos con el modelo_1 con nuestra data de entrenamiento aplicando el intervalo de confianza a el nivel requerido, en este caso .95 [Explicar porque (AOKI)]



5 Modelo de regresión lineal Múltiple

5.1 Selección de los regresores

Utilizamos las variables numéricas nuevamente para poder tener un mejor control sobre las variables y utilizamos el método "Backward", este método es computacional pero debido a que hay una cantidad manejable de variables lo hicimos a mano, tomando en cuanta un nivel de significancia de .005

Iniciamos con el modelo tomando en cuenta todas las variables del modelo

```
Call:
lm(formula = price ~ ., data = train.base_m)

Residuals:
     Min       1Q   Median       3Q      Max
-10075.3  -1602.1    -93.4   1510.3  14041.4

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -4.490e+04  1.689e+04  -2.658 0.008891 **
wheelbase         1.778e+02  1.246e+02   1.427 0.156129
carlength        -9.900e+01  7.156e+01  -1.383 0.168998
carwidth          5.001e+02  2.797e+02   1.788 0.076196 .
carheight         3.331e+02  1.785e+02   1.867 0.064288 .
curbweight        1.692e+00  2.101e+00   0.805 0.422152
enginesize        1.071e+02  1.615e+01   6.635 8.63e-10 ***
boreratio        -1.883e+03  1.525e+03  -1.235 0.219197
stroke           -2.741e+03  1.012e+03  -2.708 0.007716 **
compressionratio  3.460e+02  1.013e+02   3.415 0.000859 ***
horsepower        2.037e+01  1.929e+01   1.056 0.293020
peakrpm           2.209e+00  7.389e-01   2.990 0.003354 **
citympg          -6.103e+02  2.268e+02  -2.690 0.008108 **
highwaympg        3.010e+02  1.863e+02   1.615 0.108715
carbodyhardtop   -5.151e+03  2.252e+03  -2.287 0.023886 *
carbodyhatchback -7.085e+03  2.018e+03  -3.511 0.000619 ***
carbodysedan     -6.150e+03  2.033e+03  -3.025 0.003012 **
carbodywagon     -8.352e+03  2.241e+03  -3.727 0.000292 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3100 on 126 degrees of freedom
Multiple R-squared:  0.8823,    Adjusted R-squared:  0.8664
F-statistic: 55.54 on 17 and 126 DF,  p-value: < 2.2e-16
```

Ya con esos datos podemos observar que los Pvalues son mayores a 0.005 en algunos regresores, por lo que vamos a ir quitando variables que no cumplen con la prueba Global de la regresión y corriendo la regresión nuevamente por cada modelo.

También utilizamos el Variance Inflation Factor (VIF) el cual nos da las correlaciones entre los regresores, y estamos buscando un VIF menor a 10 en todas las variables de nuestro modelo

En el reporte sólo vamos a poner el modelo sin las variables con Pvalue mayor a 0.005

```
Call:
lm(formula = price ~ . - citympg - curbweight - carlength - horsepower -
    boreratio - highwaympg - carheight - wheelbase - carbody,
    data = train.base_m)

Residuals:
     Min       1Q   Median       3Q      Max
-12613.9  -1922.8   -426.3   1759.9  14089.1

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -7.843e+04  1.218e+04  -6.440 1.84e-09 ***
carwidth         1.014e+03  1.887e+02   5.375 3.18e-07 ***
enginesize       1.443e+02  9.913e+00  14.558  < 2e-16 ***
stroke          -3.515e+03  1.038e+03  -3.387 0.000921 ***
compressionratio 2.258e+02  8.123e+01   2.779 0.006207 **
peakrpm          3.098e+00  6.428e-01   4.819 3.75e-06 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 3458 on 138 degrees of freedom
Multiple R-squared:  0.8395,    Adjusted R-squared:  0.8337
F-statistic: 144.4 on 5 and 138 DF,  p-value: < 2.2e-16
```

Como se puede observar, las variables ya tienen un VIF menos a 10

```
                vif(modelo.8)
carwidth             2.155313
enginesize           2.249104
stroke               1.070312
compressionratio     1.327599
peakrpm              1.313397
```

5.2 Análisis de residuales

5.2.1 Comprobación de la linealidad de la Fn de regresión Multiple

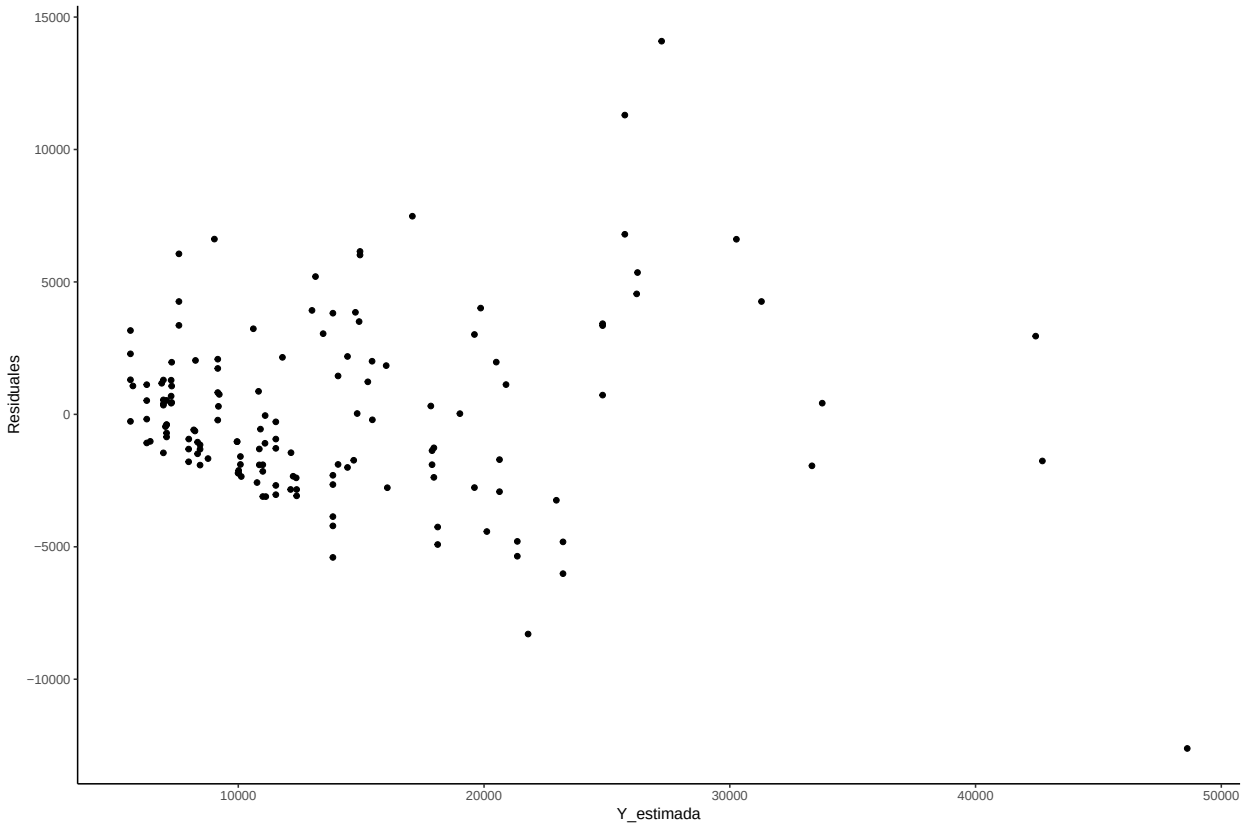
Lo checamos con la R^2 ajustada

```
[1] 0.839532
```

tenemos una R^2 ajustada que nos dice que sí es una regresión lineal.

5.2.2 Comprobamos heterocedasticidad

Esto es para poder ver si la varianza de los errores es constante.

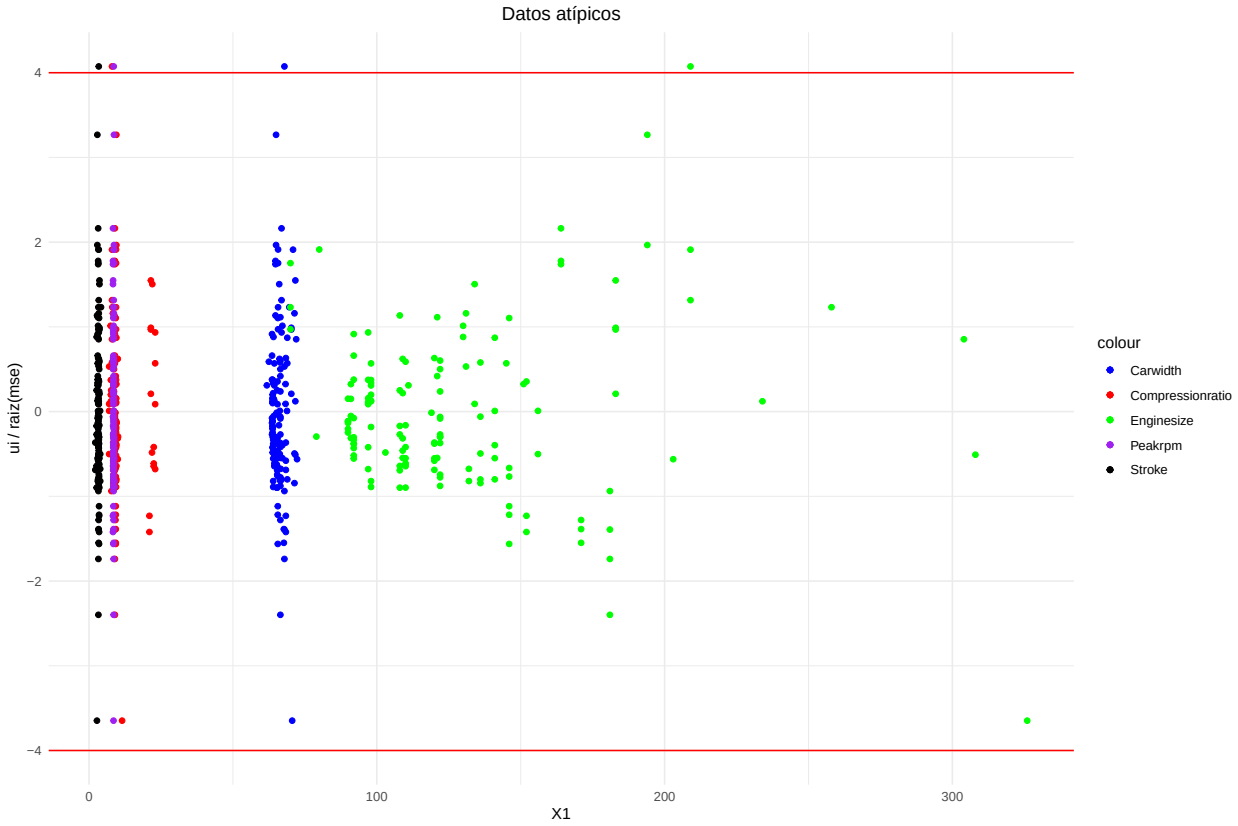


Se puede observar que no hay un patrón en sí en los errores, lo cual nos dice que sí hay heterocedasticidad

5.2.3 Independencia en los errores

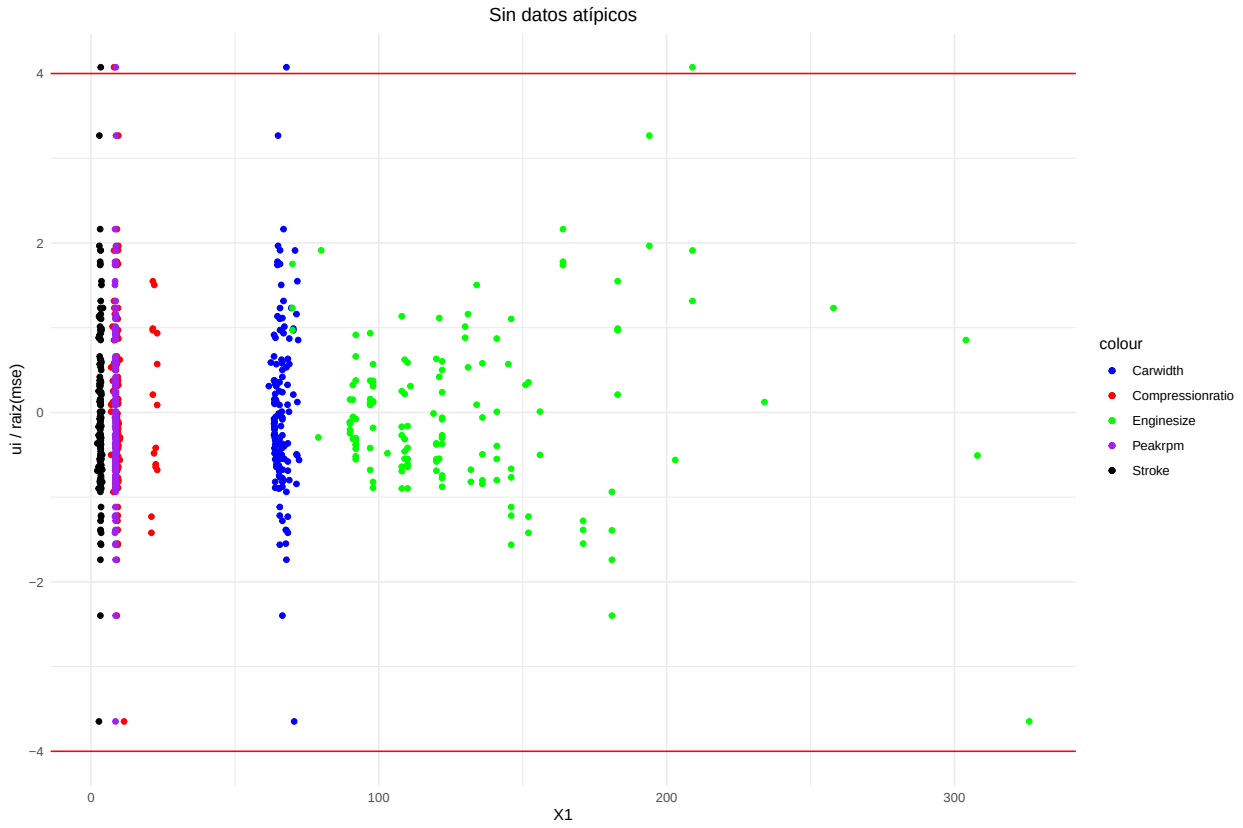
No es una serie de tiempo - no aplica ya que los datos no llevan un orden y pueden cambiar de pocisión

5.2.4 Presencia de errores atípicos



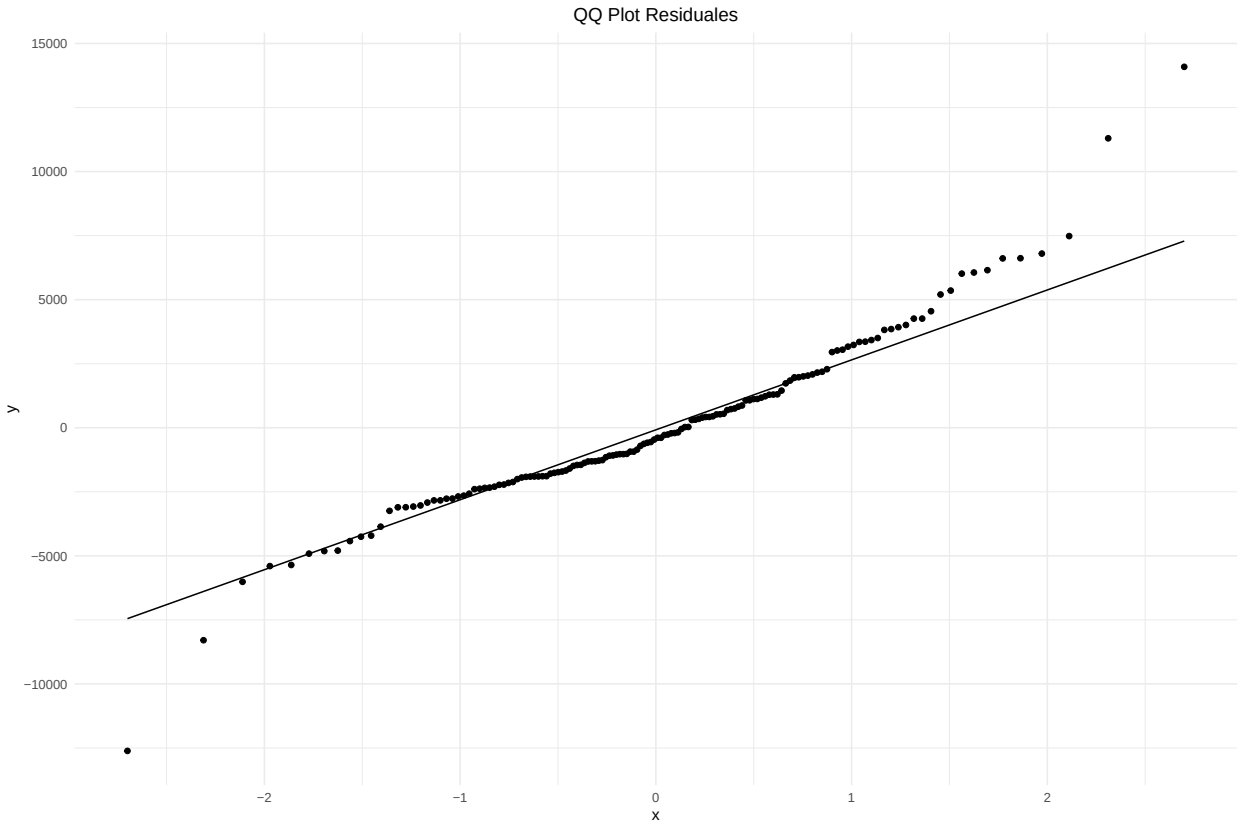
Se puede observar que hay errores atípicos, los cuales podemos eliminar debido a que este análisis no es una serie de tiempo, en nuestro DF es el renglón 28, por lo que lo eliminamos

Graficamos para confirmar que no hay datos atípicos ahora


Sin datos atípicos
ui / raiz(mse)
X1
4
2
0
−2
−4
0
100
200
300
colour
Carwidth
Compressionratio
Enginesize
Peakrpm
Stroke

5.2.5 Verificar la normalidad en los errores

La QQ plot - esa se hace con los residuales sacamos los residuales de nuestro modelo



Rechazamos el supuesto de normalidad de los errores debido a las dos colas que muestra el gráfico de QQ plot. No obstante, ya que el modelo de regresión lineal Multiple ajustado es robusto ante el supuesto de normalidad podemos continuar usando estas variable.

6 Comparativo entre modelos y selección de un modelo

```
El modelo de regresión lineal simple tiene una R^2 ajustada de  [1] 0.7772391
```

```
Y el MRLM tiene una R ajustada de  [1] 0.839532
```

Ya que estamos utilizando las R^2 ajustadas, las cuales si nos dejan comparar modelos, y vemos que R^2 del MRLM es mayor que la del MRLS por:

```
[1] 0.05647888
```

Lo cual nos lleva a elegir el MRLM que es el mejor ajustado para nuestro precio.

7 Conclusiones

8 Bibliografía