Regresión Lineal Simple

Equipo X

14 May, 2023

Contents

1 Introducción **1**

1.1 Problema de interés: Análisis del precio de los vehículos 1

1.2 ¿Por qué usar una regresión lineal? 1

2 Marco teórico **2**

2.1 Conceptos básicos 2

2.2 Supuestos del modelo 2

2.3 Método de selección de variables y limitaciones del modelo 2

3 Análisis exploratorio de datos **3**

3.1 Análisis de la base de datos 3

3.2 Selección de la variable explicativa 3

4 Modelo de regresión lineal simple **6**

4.1 Parámetros del modelo 6

4.2 Análisis de residuales 7

4.3 Intervalo de confianza y predicción al 95% 10

5 Modelo de regresión lineal Múltiple **11**

5.1 Selección de los regresores 11

5.2 Análisis de residuales 13

6 Comparativo entre modelos y selección de un modelo **15**

7 Conclusiones **16**

8 Bibliografía **16**

1 Introducción

La globalización causó una revolución en el mercado automotriz. La apertura al comercio y la inversión internacional creó mercados más competitivos. Los consumidores ya no tenían que resignarse a consumir únicamente manufacturas nacionales, sino que podían acceder a una amplia gama de proveedores para un mismo tipo de producto o servicio. Un caso emblemático es el sector automotriz que, a pesar de ser una industria altamente protegida, no pudo evitar la penetración de competidores extranjeros. Hoy en día los diferentes fabricantes de automóviles compiten para colocar su producto y extraer mejores márgenes. Bajo este entorno altamente competitivo es necesario que los tomadores de decisiones hagan uso de herramientas precisas que ayuden a encontrar áreas de oportunidad y ventajas comparativas. Las herramientas econométricas se presentan como una manera de asistir a los tomadores de decisiones para hacer predicciones precisas y afrontar los desafíos de la industria.

Los modelos econométricos que predominan en la literatura se centran principalmente en evaluar factores externos al diseño automotriz como determinantes del precio. No obstante, estos estudios resultan anacrónicos ante el entorno competitivo actual. La entrada de más competidores y por ende la reducción de la cuota de mercado hace que controlar los factores externos a la producción sea cada vez más complejo, inasequible y costoso. Al mismo tiempo, tal competitividad crea fuertes incentivos por reducir los costos de producción de manera que el precio final al consumidor sea el menor posible, sin sacrificar márgenes atractivos.

La innovación y el desarrollo tecnológico ayudan con esta tarea, pero se debe ser eficiente en el uso de recursos. En este trabajo se propone el uso de un modelo de regresión linear simple (MRLS) en contraposción con un módelo de regresión lineal multiple (MRLM), para identificar el componente automotriz que tenga un mayor efecto sobre en el precio. De esta forma, la alta gerencia tendrá acceso a herramientas con base en la evidencia empírica que ayude a determinar la asignación eficiente de recursos para la investigación y desarrollo (I+D). Proponemos que un uso eficiente de I+D ayudará a abaratar costos y reducir el precio de venta, sin sacrificar la competitividad en el mercado.

1.1 Problema de interés: Análisis del precio de los vehículos

1.1.1 Breve explicación de la base de datos “Scrap price”

1.2 ¿Por qué usar una regresión lineal?

2 Marco teórico

2.1 Conceptos básicos

El MRLS es una técnica para modelar la relación lineal entre dos variables. De manera general, el MRLS se define en la siguiente ecuación: y=BO + B1X1 + U Donde y es la variable dependiente y x la variable independiente, mientras que la variable u, denominada como la perturbación estocástica, son variables aleatorias no observables que representa a todos los factores distintos de x que afectan a y. En tanto BO y B1, representan respectivamente el coeficiente del intercepto y el coeficiente de la pendiente. El coeficiente de la pendiente es el interés principal del análisis econométrico, pues mide el efecto de la variable independiente sobre la variable independiente mantenido todos los demás factores constantes. En este caso en particular, omitiremos el análisis del coeficiente del intercepto.

En el caso en particular que se explora en este reporte, la variable independiente y será el precio del automóvil, y la dependiente x será determinada durante el proceso de selección de variable.

2.2 Supuestos del modelo

El MRLS depende de seis supuestos elementales, que deberemos verificar para determinar La validez del modelo; En primer lugar, asumimos que la relación entre Xi y Yi es lineal en parámetros; en segundo lugar, se asume que no existe una dependencia lineal entre los errores y la variable dependiente o, dicho de otra manera, Cov(Ui, Xi) = 0; en tercer lugar, se supone que el valor promedio de u en la población es igual a cero, por lo que E(U) = 0, debido a esto, dejamos BO en el modelo; en cuarto lugar, se da por supuesto la homocedasticidad, por lo que Var[Ui] = (SIGMA)ˆ2 ;en quinto lugar, el número de las observaciones debe ser mayor que el de los parámetros; y por último, los valores de xi no debieron ser todos iguales ni existen observaciones atípicas.

2.3 Método de selección de variables y limitaciones del modelo

Dado que la base de datos en la que se basó este reporte tiene observaciones y variable relativamente limitadas (se analizaron 205 modelos de automovil y se capturaron 23 variables diferentes) se utilizó un método de selección de variables basado en un análisis gráfico de los datos.

El MRLS es una herramienta que ayuda a orientar a los tomadores de decisiones, más no es capaz de detallar relaciones complejas. Principalmente, se encuentra limitado por explicar la relación que existe entre una sola característica del auto y el precio, cuando los determinantes de este último son multivariados. Al no poder incorporar más información el modelo propuesto podría caer en la redundancia al crear otro tipo de ineficiencias derivadas del análisis de una relación simple. Esto dado que el modelo podría informar decisiones de inversión de recursos sin considerar relaciones más complejas.

3 Análisis exploratorio de datos

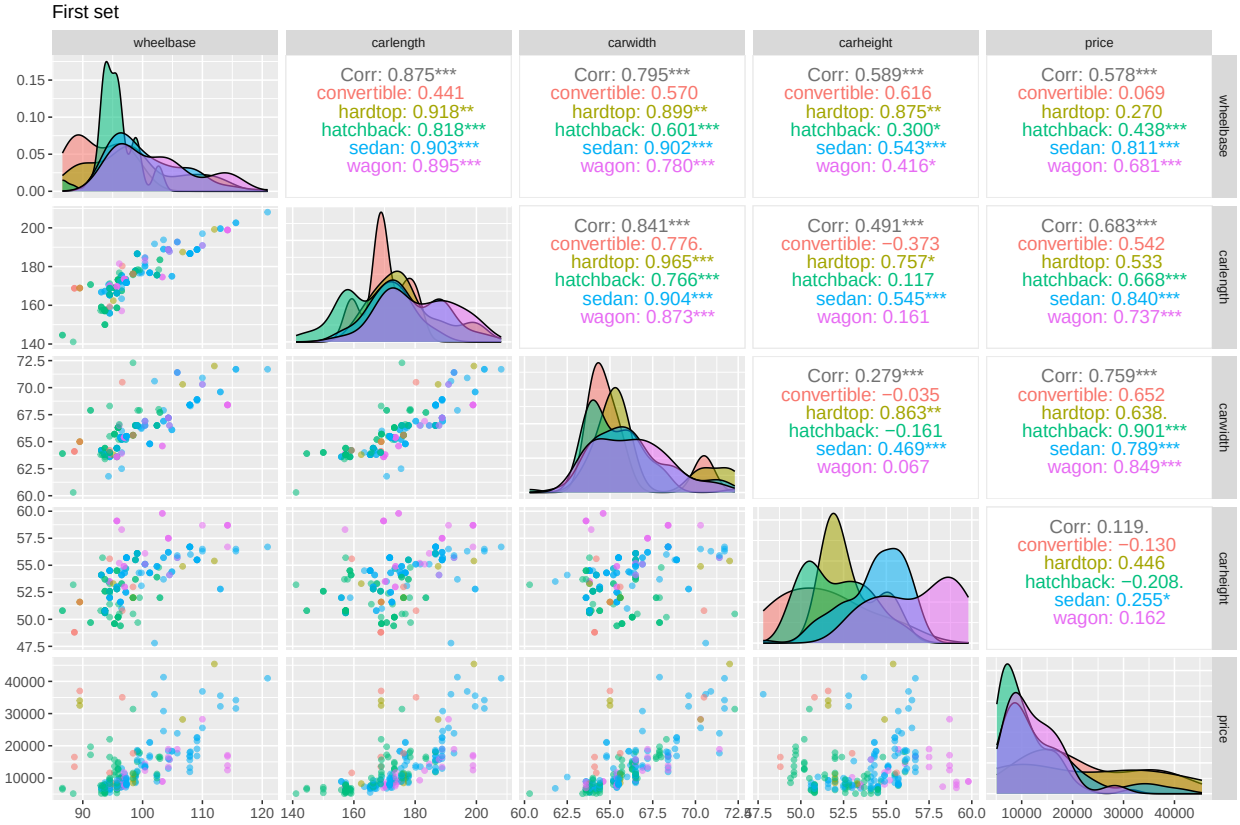
3.1 Análisis de la base de datos

Para constituir la base de datos se registraron 23 variables para 205 modelos de automóvil, así como sus precios, correspondientes a 22 fabricantes de autos. Todas las variables corresponden a características del auto, como el número de puertas, las medidas del modelo, etc.

Como primer filtro, se descartaron las variables no numéricas. Después, se dividía a las 13 variables restantes en tres lotes para los que se generó una matriz gráfica de la correlación de cada variable con el precio del modelo automovilístico y resto de las variables del lote. De ahí se prosiguió a un análisis gráfico en el que se determinó las variables con la mayor correlación lineal al precio de la unidad.

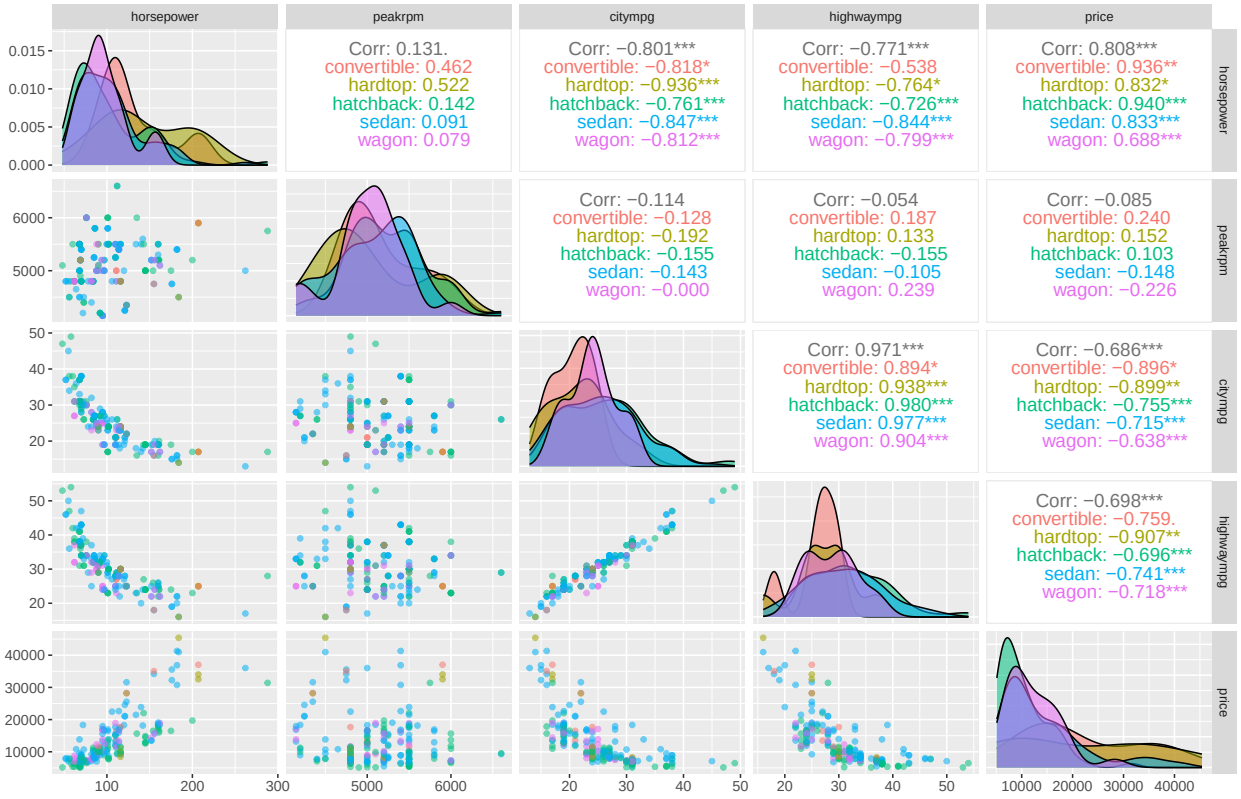
3.2 Selección de la variable explicativa

Dado el análisis gráfico se determinó que dos variables demostraban tener la mayor relación lineal: caballos de fuerza y tamaño del motor. Dado que los caballos de fuerza representan la energía que produce el motor, que a su vez está en función del tamaño del motor, se determinó que la variable dependiente deberá ser el tamaño del motor.



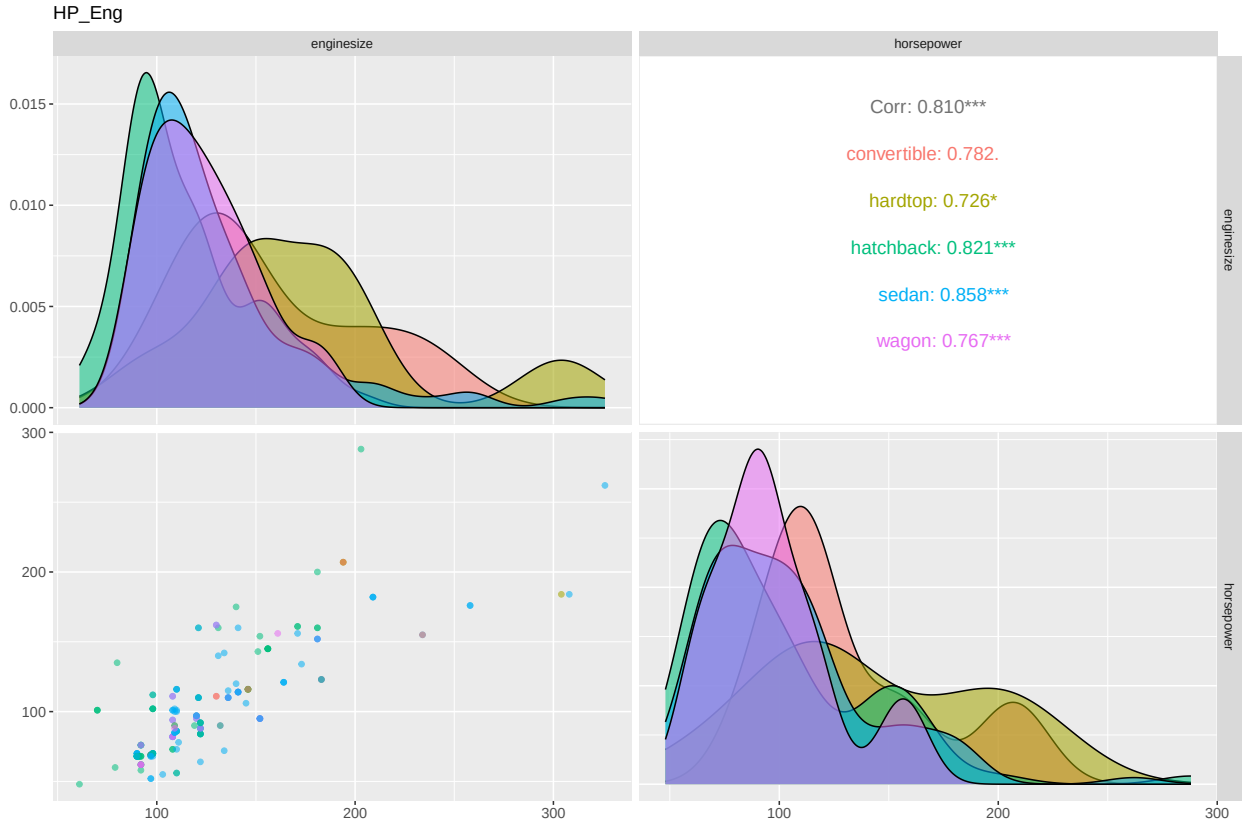
Second set

Third set



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

Horsepower, curbweight y enginesize con 0.808, 0.835 y 0.874 respectivamente. Aunque las tres tienen correlaciones muy altas, teóricamente hace más sentido escoger es “enginesize”, ya que tenemos la evidencia que el tamaño del motor influye en el precio de un auto, además de ser la variable con la correlación más alta respecto al precio. En este sentido, 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 sea la mejor representación posible de 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 lineal 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.

A continuación, presentamos los siguientes valores obtenidos, para describir lo explicado anteriormente:

```
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

A continuación, la 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

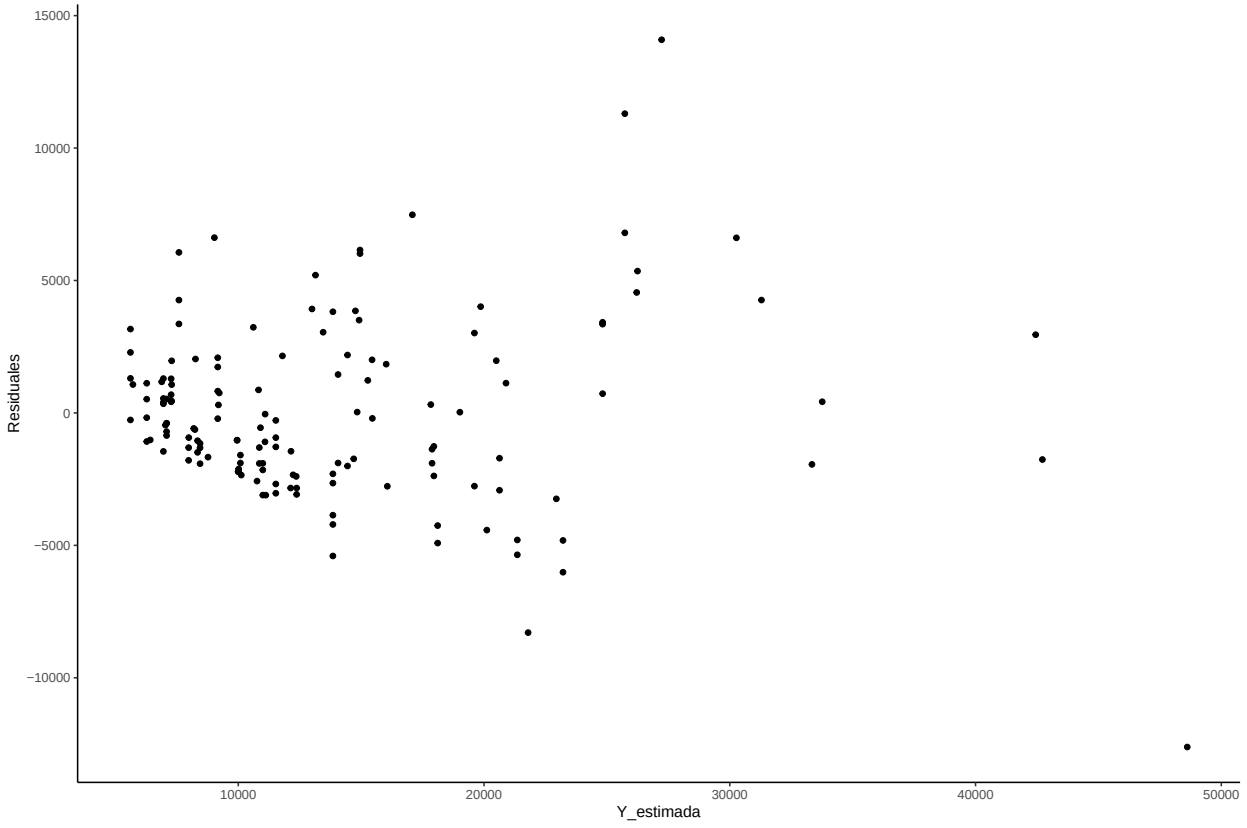
4.2.1 Comprobación de la linealidad de la Fn de regresión

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.

Al calcular la R^2, como anteriormente se mencionó, obtenemos que es .7787968 ; esto quiere decir que la variable X explica en un 77.88% a la variable dependiente Y.

4.2.2 Heterocedasticidad

Comprobamos heterocedisticidad (que la varianza de los errores sea constante), lo comprobamos con un gráfico, donde comparamos los residuales con las Y observadas (ŷ), para esto tenemos que hacer un DF con ambos vectores obtenidos de nuestro modelo



Dado el anterior gráfico, podemos observar que no existe un patrón notório en el gráfico, ya sea como la silueta de una recta, con esto podemos asumir que hay heterocedisticidad, por lo tanto la varianza de nuestros errores es constante.

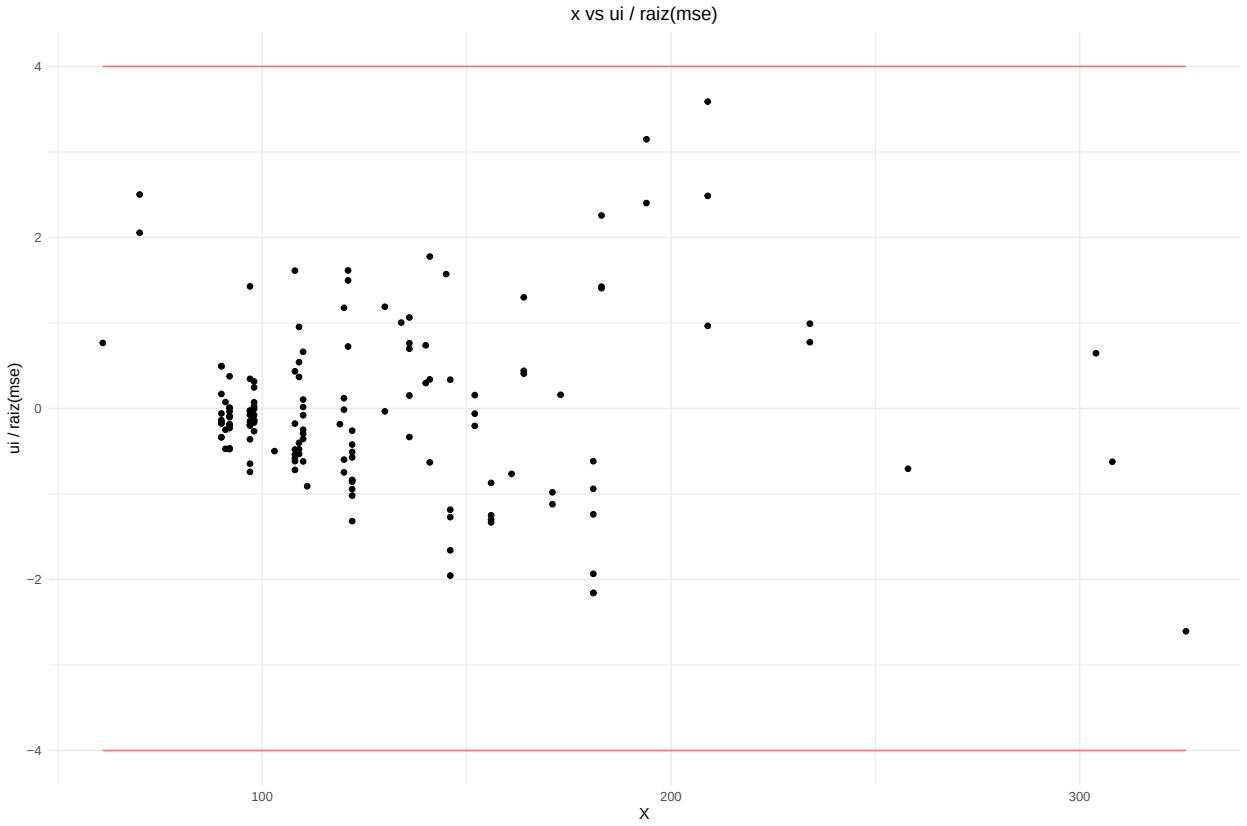
4.2.3 Independencia en los errores

Los datos usados en este módelo no corresponden a una serie de tiempo, por lo tanto no aplica, ya que los datos no llevan un orden y pueden cambiar de posición

4.2.4 Presencia de errores atípicos

Este resultado lo obtenemos de la siguiente forma: Obtenemos la suma de cuadrados de los errores (MSE), de la tabla ANOVA, después calculamos la raiz cuadrada del MSE.

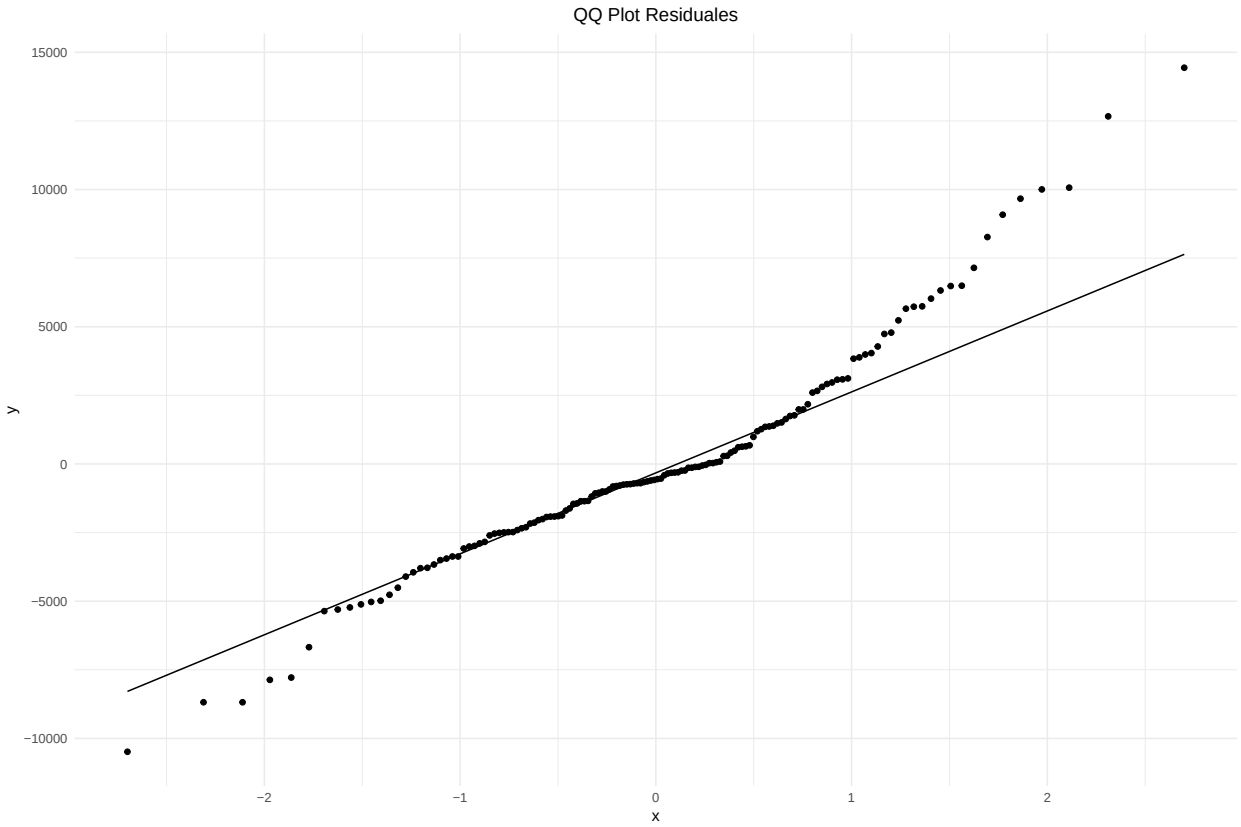
Ya con (MSE)ˆ(1/2), podemos obtener la división de los residuales entre la raíz de MSE y compararlos con nuestras x'i, estas variables las metemos en un DF y graficamos las diferencias.



Como podemos observar en el gráfico anterior no existe ningún dato atípico.

4.2.5 Verificar la normalidad en los errores

Usaremos la función QQ plot para verificar esto; esta función se realiza con los residuales obtenidos en nuestro modelo.

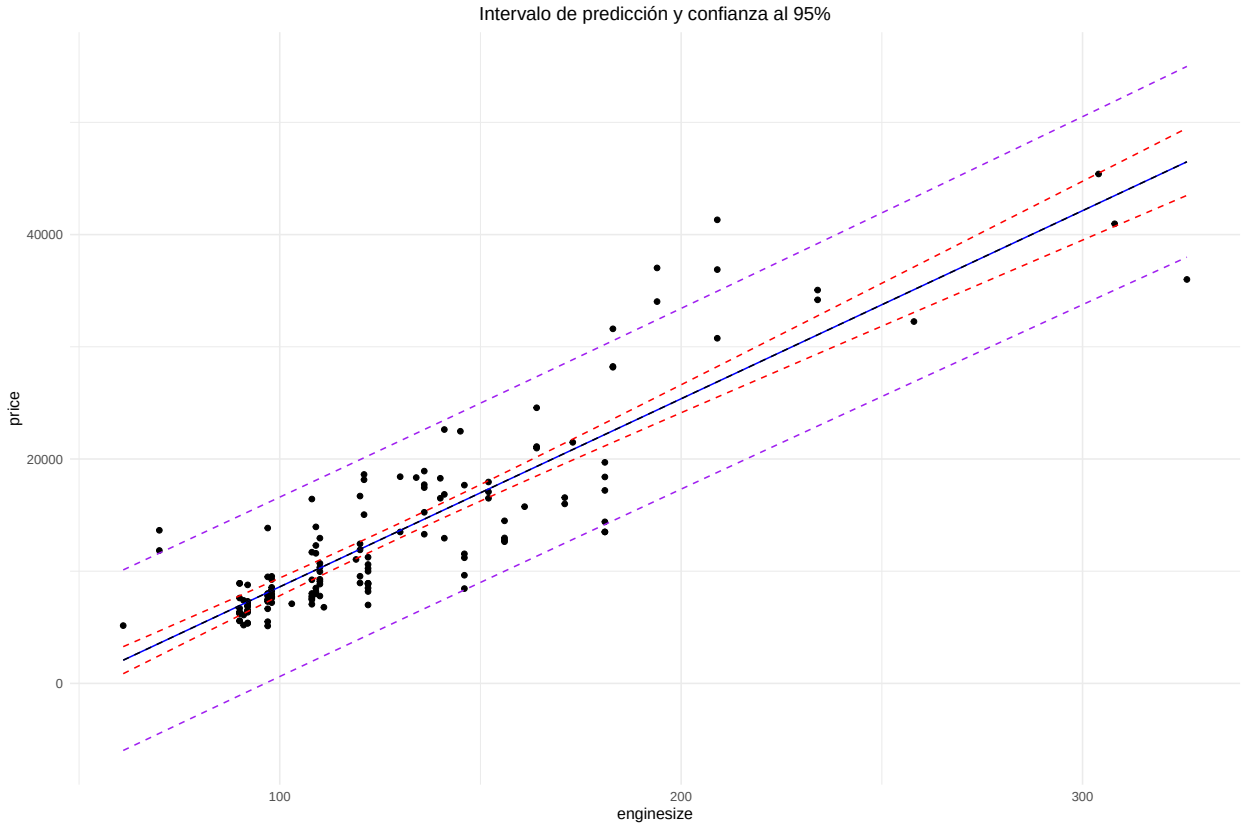


Rechazamos el supuesto de normalidad de los errores debido a las dos colas que muestra el gráfico de QQ plot. No obstante, ya que el modelo de regresión lineal simple ajustado es robusto ante el supuesto de normalidad podemos continuar usando esta variable explicativa.

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



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 regresoras 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 cuenta un nivel de significancia de .05

Iniciamos el modelo tomando en cuenta todas las variables numericas.

```
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
```

Con los datos que obtuvimos anteriormente, podemos observar que algunos de los P-values de las regresoras son mayores a 0.05, por lo que iremos depurando las variables que no cumplen con la prueba Global de la regresión, e iremos corriendo la regresión nuevamente por cada modelo.

También utilizamos el Variance Inflation Factor (VIF) el cual nos muestra las correlaciones que existen entre los regresores, para esta prueba estamos buscando que el 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 0.05.

```
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
```

Observando los resultados obtenidos y los requisitos expuestos, estas son las regresoras que cuentan con un VIF menor 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

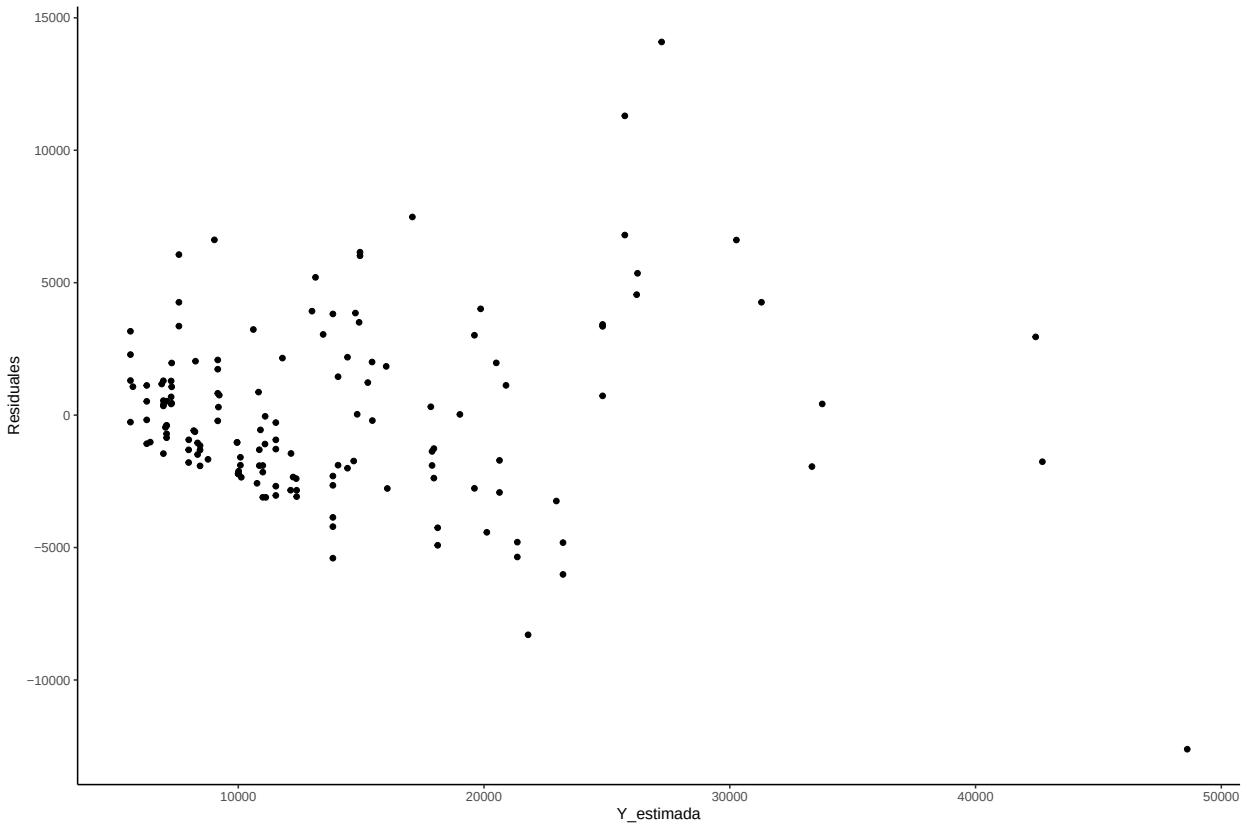
Ahora comprobamos nuestros resultados usando la R^2 ajustada, una vez que la calculamos, obtenemos lo siguiente:

```
[1] 0.839532
```

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

5.2.2 Comprobamos heterocedasticidad

Nuevamente, en el modelo de regresión lineal multiple, debemos comporbar que la varianza de los errores sea constante, comprobamos que se cumple la heterocedasticidad.

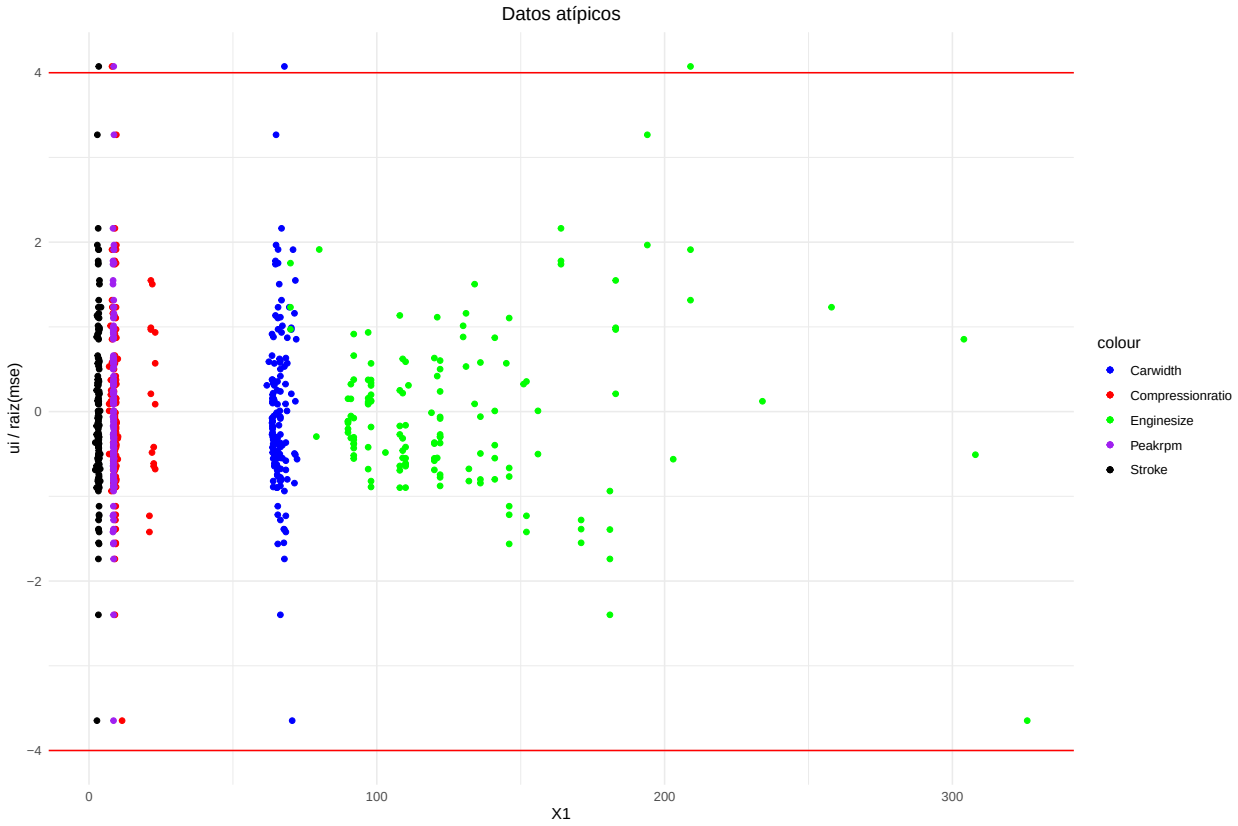


Como podemos observar en el gráfico anterior, no se nota que exista un patrón evidente en los errores, lo cual verifica que sí hay heterocedasticidad.

5.2.3 Independencia en los errores

Los datos usados en este módelo no corresponden a una serie de tiempo, por lo tanto no aplica, ya que los datos no llevan un orden y pueden cambiar de posición.

5.2.4 Presencia de errores atípicos

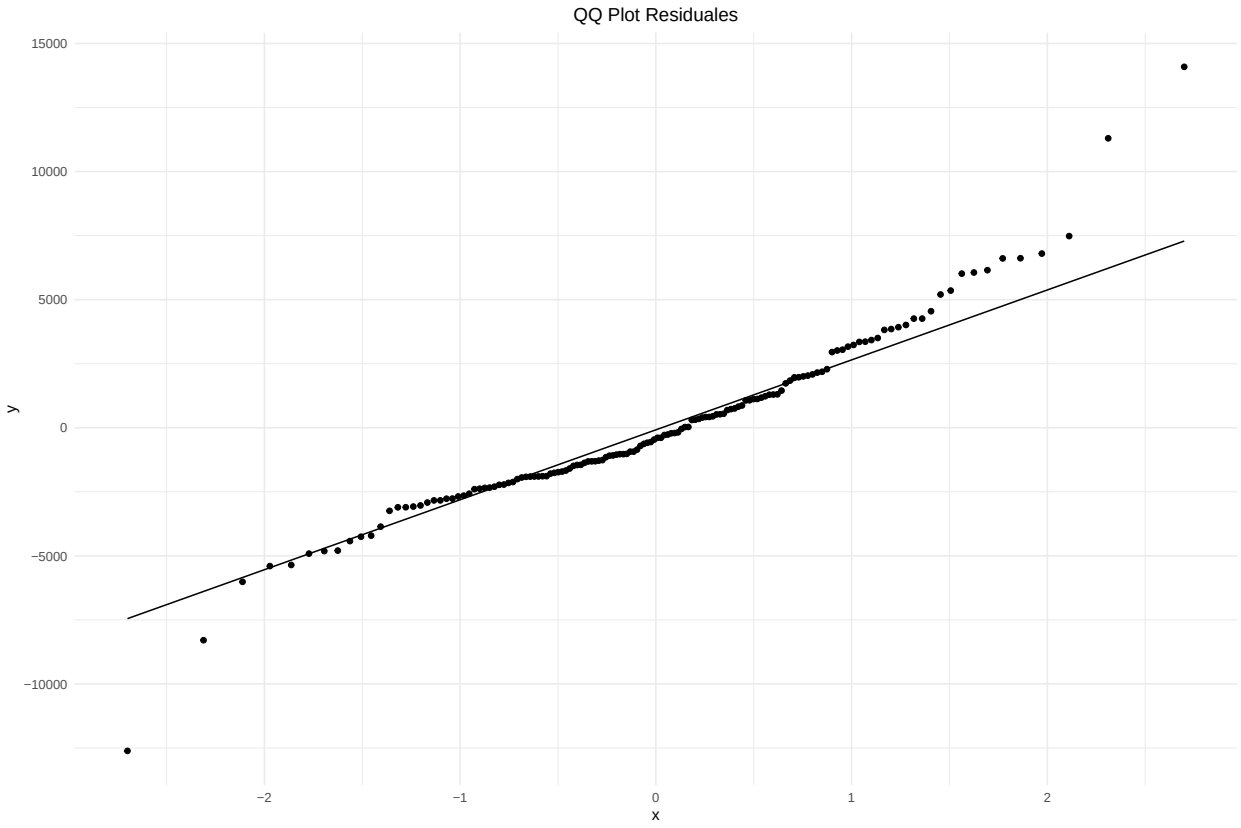


Observando este gráfico, podemos notar que hay errores atípicos, los cuales podemos eliminar debido a que este análisis no es una serie de tiempo, en nuestro DF es el renglón 28, por lo que lo eliminamos

Una vez que eliminamos estos datos, podemos proseguir con las siguientes verificaciones.

5.2.5 Verificar la normalidad en los errores

Para verificar este punto, usamos la función QQ plot; esta se realiza con 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

A partir de la evidencia recolectada en los analisís que se realizaron en este trabajo, podemos concluir con seguridad que el Modelo de Regresión Lineal Multiple es mejor para este problema y explica de mejor manera el compartamiento del precio en los coches.

Esta conclusión se puede explicar desde dos maneras; la primera, de una forma matemática, en este sentido sabemos que la R^2 (ajustada) será mejor en un modelo multiple a comparación de un modelo simple. Este razonamiento lo obtenemos a partir de la formula de la R^2 (ajustada), como sabemos a mayor número de regresoras, el valor de la R^2 (ajustada) será mayor al del modelo lineal; en segundo lugar, de forma teórica, sabemos que en muy pocos bienes, si no es que en ninguno, el precio depende únicamente de un solo factor. Por lo que, es de esperarse que un modelo acierte en mayor medida al precio de un coche si se consideran más factores. En el caso especifico de los coches, el precio se ve afectado por muchos factores. Dentro del mercado, existen muchos modelos que son iguales en casí todos los factores, con excepción de un par y la diferencia en precios se explica por esas mínimas diferencias. De igual forma, es una industra que no se diferencia por un solo factor, cada marca, cada modelo, tiene sus caracteristicas particulares y diferencias entre sus diferentes factores.

Dado los puntos anteriores, se nota una clara mejora entre usar el modelo de regresión lineal multiple, al usar el modelo de regresion lineal simple.

8 Bibliografía

Gatto, L. (2017, 10 noviembre). Data analysis and R programming. https://lgatto.github.io/2017_11_09_Rcourse_Jena/manipulating-and-analyzing-data.html

Hernández, F. (2020, 30 octubre). 12 Pruebas de independencia de los errores | Modelos de Regresión con R. https://fhernanb.github.io/libro_regresion/indep.html

Chatterjee, S., & Price, B. (1980). Regression Analysis by Example. Journal of the American Statistical Association, 75(372), 1039.

Dalgaard, P. (2008). Introductory Statistics with R. En Springer eBooks.