

Regresión Lineal Simple

Equipo X

25 abril, 2023

Contents

1	Introducción	1
1.1	Problema de interés: Análisis del precio de los vehículos [Raúl]	1
1.2	¿Por qué usar una regresión lineal? [Raúl]	1
2	Marco teórico [AOKI]	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

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

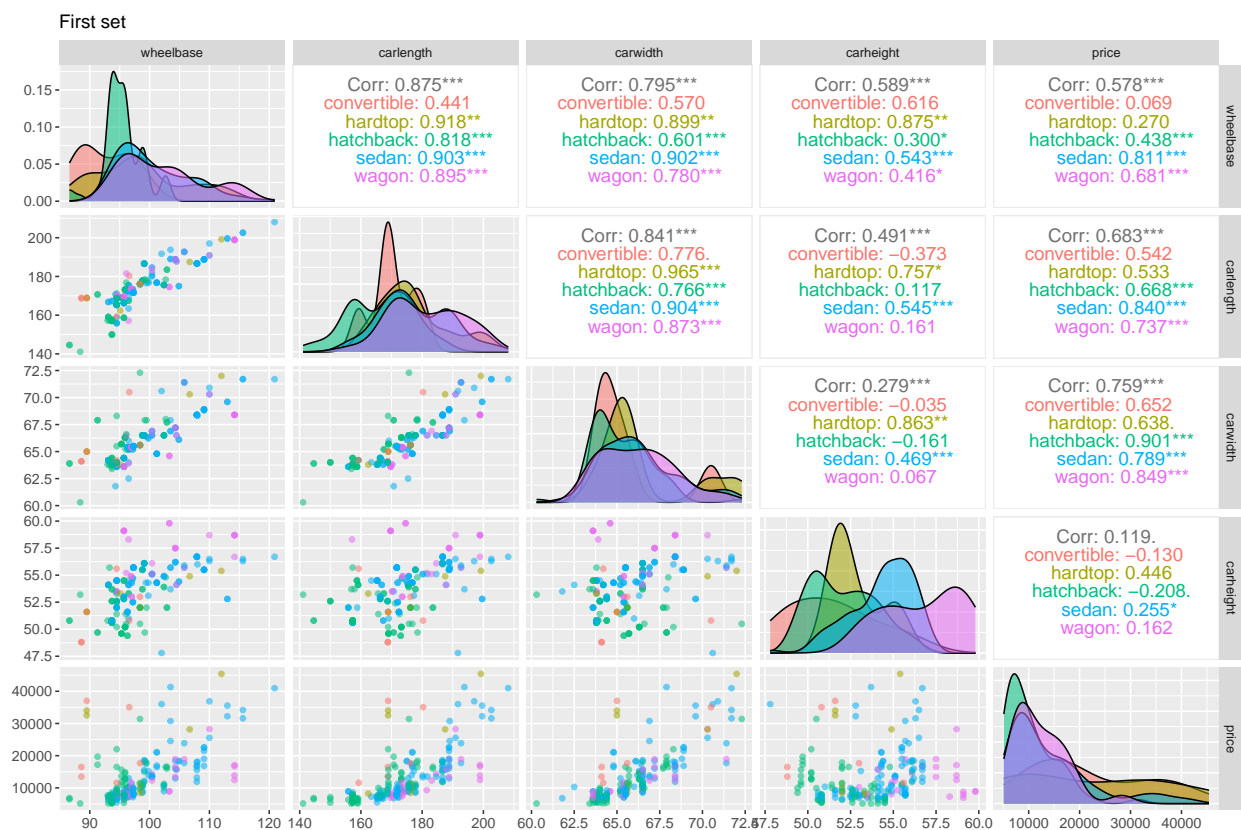
3 Análisis exploratorio de datos

3.1 Análisis de la base de datos

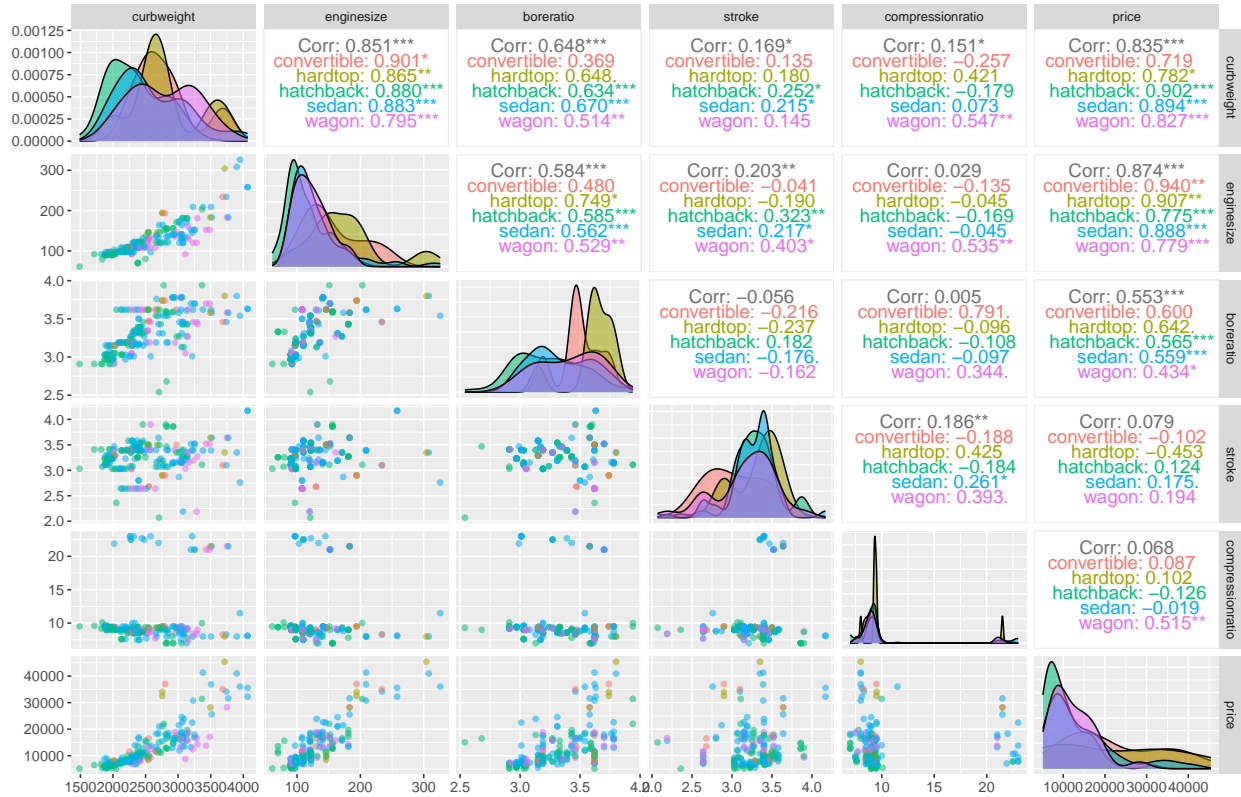
Aquí va filtros aplicados, estadígrafos, gráficas, limpieza y selección de variable dependiente con justificación

3.2 Selección de la variable explicativa

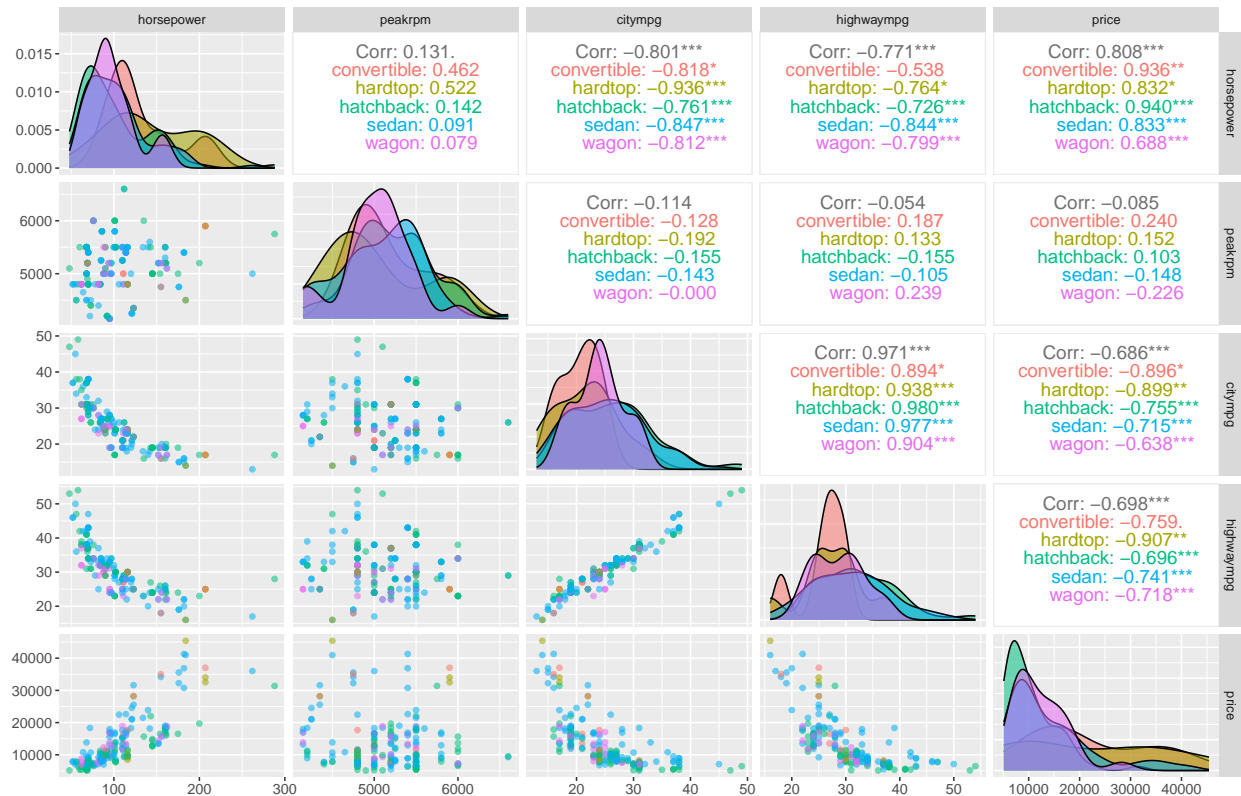
Para la selección de la variable explicativa eliminamos las variables de caracter, y dejamos las variables numéricas. Dividimos las variables en 3 grupos para poder analizar la correlación de las variables respecto el precio:



Second set

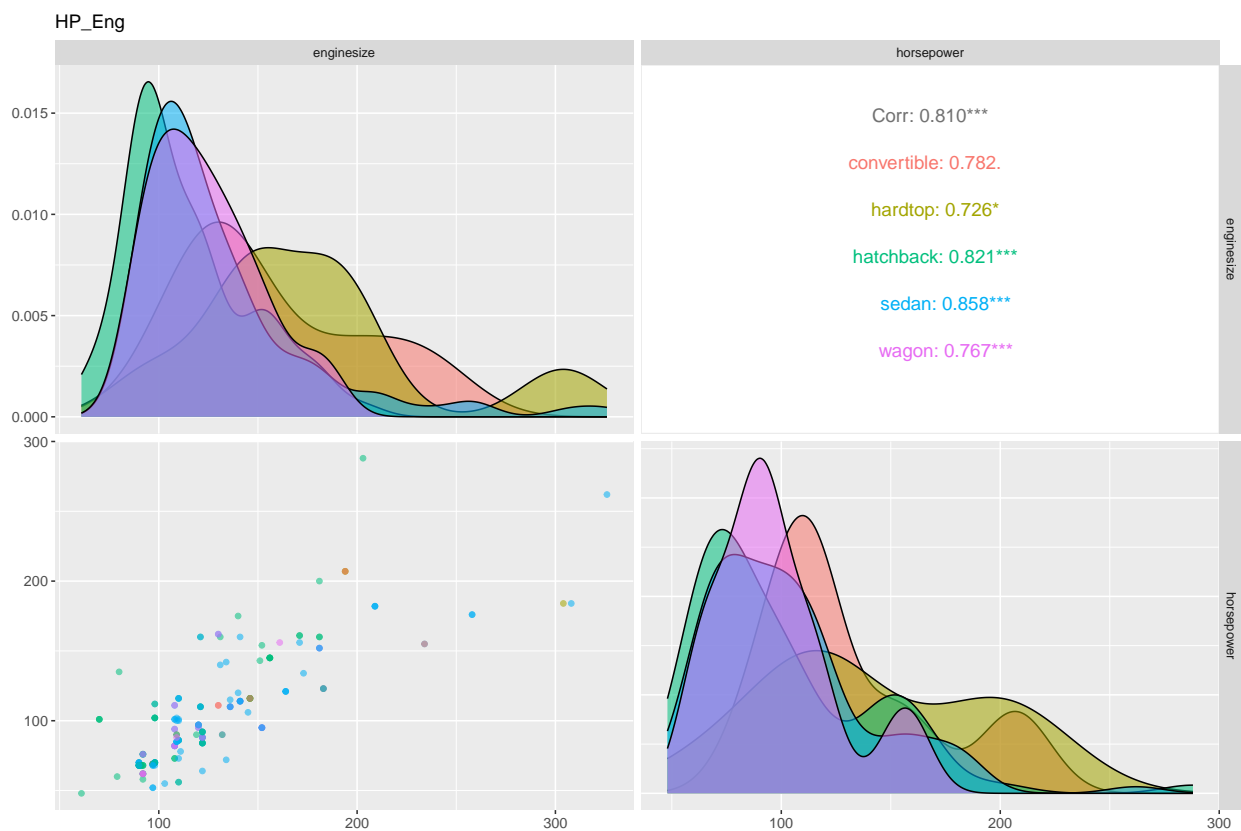


Third set



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 multicolinealidad imperfecta de 0.810 con la variable que elegimos.



4 Modelo de regresión lineal simple

4.1 Parámetros del modelo

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

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

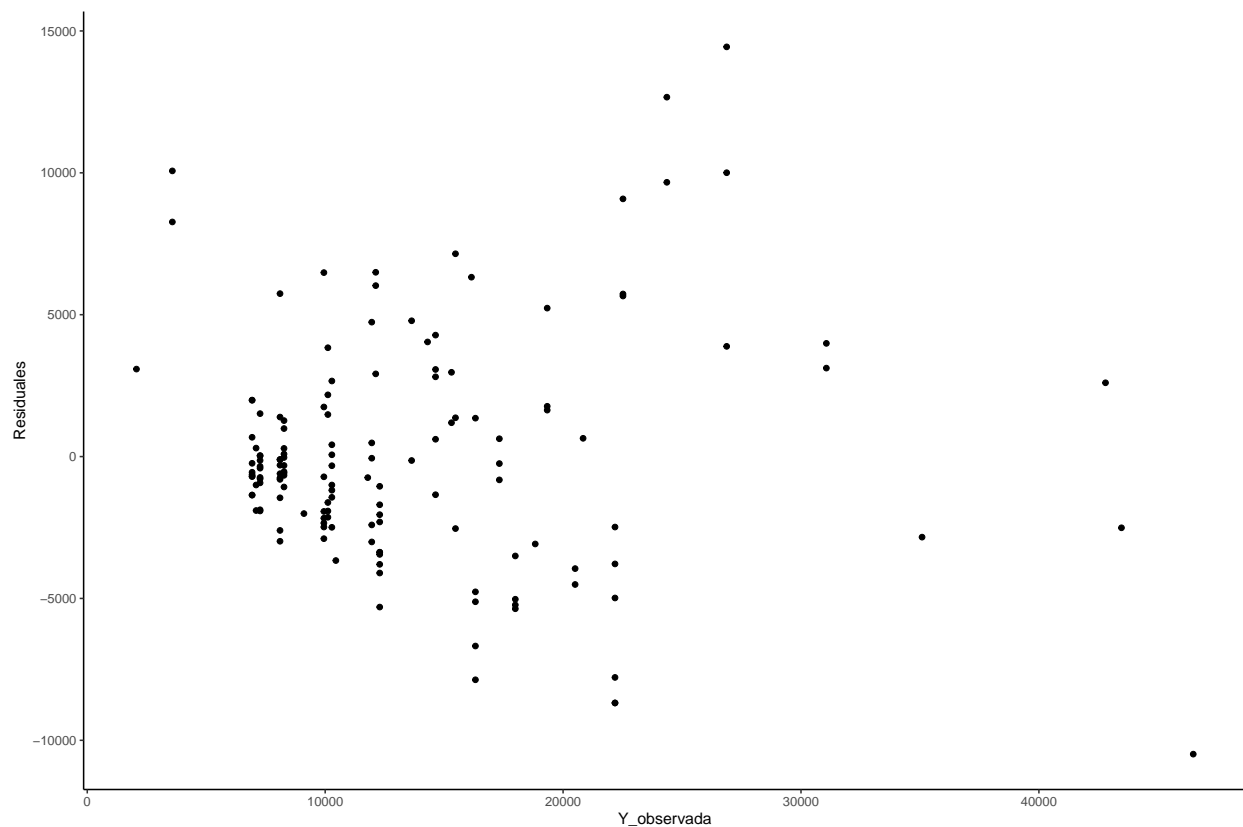
4.2.1 Comprobación de la linealidad de la Fn de regresión

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

4.2.2 Heterocedasticidad

Comprobamos heterocedasticidad (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 heterocedasticidad.

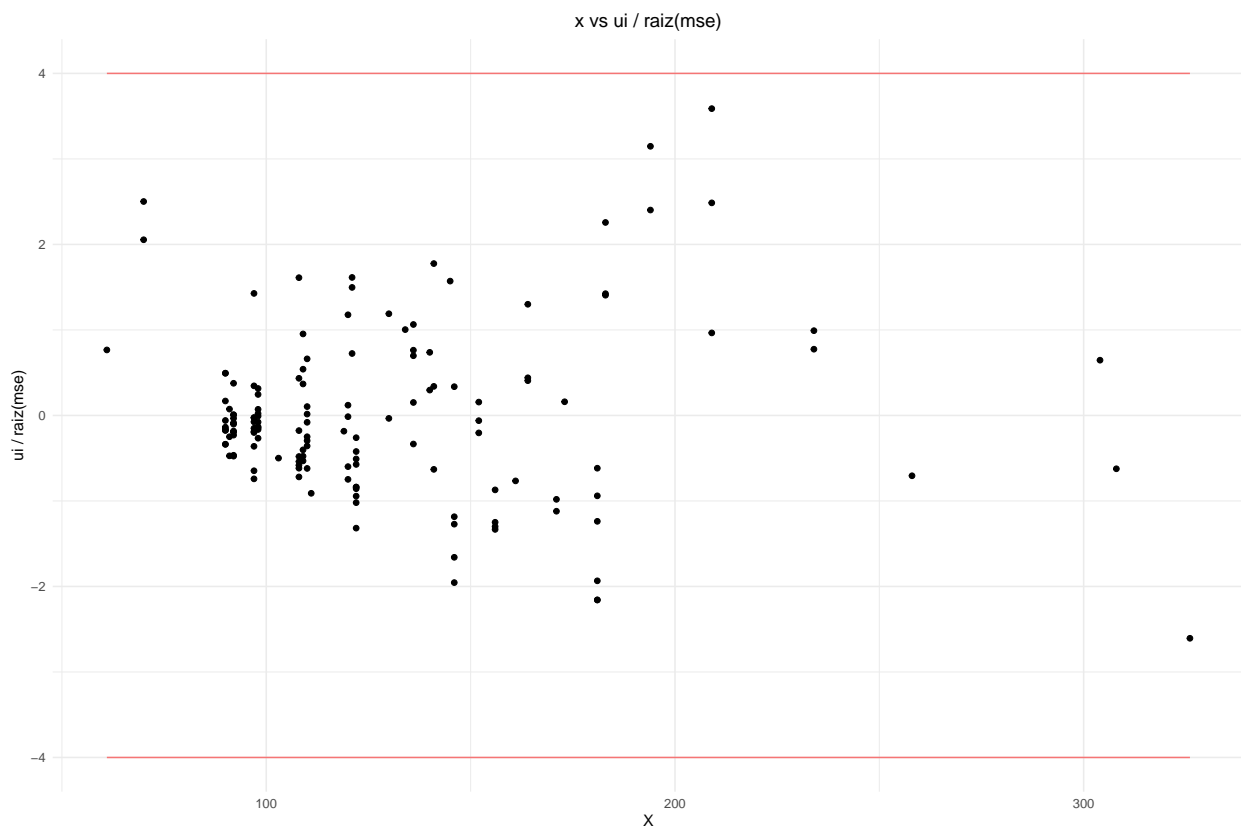
4.2.3 Independencia en los errores

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

4.2.4 Presencia de errores atípicos

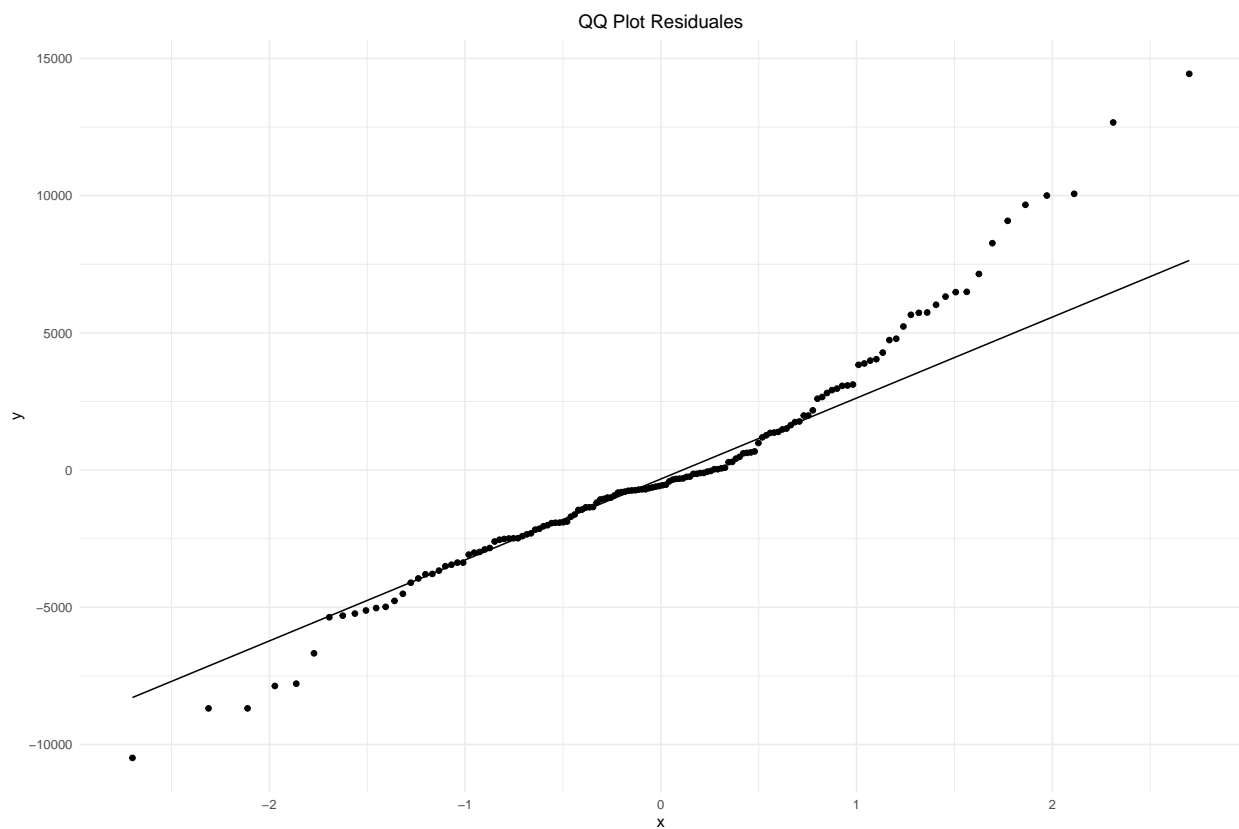
Esto se hace calculando la raíz de el cuadrado medio de la suma de cuadrados de los errores (MSE), el cual se obtiene de la tabala ANOVA de nuestros residuales el mean.

Ya con $MSE^{(1/2)}$, podemos sacar la división de los residuales entre la raíz de MSE y compararlos con x_i 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)]

