Universidad Nacional de Colombia
Facultad de Ciencias

Departamento de Estadística

Análisis de Regresión para la predicción del precio de un automóvil

Docente:
Mario Enrique Arrieta Prieto

Grupo 7

Autores:

Andrés Mauricio Rico Parada	Estadística	aricop@unal.edu.co
Ander Steven Cristancho Sanchez	Estadística	acristanchos@unal.edu.co
John Anderson Guarín López	Estadística	jaguarinl@unal.edu.co
Juan David Carrascal Ibañez	Matemáticas	jdcarrascali@unal.edu.co

Noviembre 2022

1. Descripción

Este trabajo, de carácter académico, busca afianzar los conocimientos adquiridos en el curso Análisis de Regresión, aplicando algunas de las temáticas vistas en clase en un caso práctico. Es por ello que, para esta segunda parte se realizará un análisis de las diferentes fases que se va a abordar en el estudio del precio de un automóvil. La temática en cuestión consiste en el análisis del precio de los automóviles en Estados Unidos en función de las variables elegidas en la primera entrega. Para abordar esta temática, en esta segunda parte se realizará el análisis del modelo de regresión elegido para el logro de cada uno de los objetivos propuestos en la primera entrega. Por lo tanto, esta segunda entrega se divide en los siguientes items: Fase de identificación, Fase de estimación, Fase de validación, observaciones de alta palanca, atípicas e influyentes. Luego de realizar estas fases, se concluirá con el uso del modelo propuesto. A lo largo de este trabajo se darán respuesta a diferentes preguntas, referentes a cada uno de los ítems enunciados.

2. Fase de Identificación:

Antes de empezar con el desarrollo de esta fase, primero hay que recordar que el objetivo general de este trabajo consiste en estudiar los factores que influyen en el precio de un modelo de automóvil en el mercado estadounidense. Por ello, la variable de interés para efectos de este trabajo, corresponde al precio de un automóvil.

Para determinar las variables que inciden en el precio de un automóvil, se realizó una búsqueda bibliográfica, donde se mencionan las principales variables o factores que afectan el precio de un automóvil. En esta revisión de la literatura se identificaron las variables que exponen otros autores; con el fin de identificar si las variables propuestas en la primera entrega de este trabajo coinciden con las variables encontradas bajo el análisis de la literatura. A continuación se realiza una breve descripción de los textos encontrados y las variables identificadas, las cuales servirán como base para la elección de las variables que se tendrán en cuenta para la elección del modelo.

El primer texto, basado en la página de Minicooper de Perú, tiene como objetivo en identificar los factores que influyen en el precio de un auto (Fuente: `https://www.inchcapemotors.com.pe/mini/preciosautosnuevos`). En dicho artículo, destacan las variables **Tipo de motor, rendimiento, uso de combustible, existencia de frenos especiales (ABS) y otros sistemas**. De estas variables enunciadas, coinciden el Tipo de motor, rendimiento y el uso del combustible.

La segunda fuente, basada en un artículo referente al mercado automotriz ecuatoriano, destacan los factores que determinan el precio de un automóvil, al igual que el texto anterior; sin embargo, estos factores son enunciados de acuerdo al crecimiento de la demanda de autos en la pandemia (Fuente: `https://www.primicias.ec/nota_comercial/autos/garage/talleres/la-demanda-de-vehiculos-usados-crecio-en-la-pandemia/#gsc.tab=0`). Los factores que destacan son: **Marca, tipo de vehículo, dimensiones del vehículo y motor**. De estas variables mencionadas, se tienen en común la marca, el tipo de vehículo y las dimensiones del motor.

La tercera fuente encontrada hace referencia a los factores externos e internos que afectan el precio de un automóvil. Esta fuente es tomada de la página de Money Crashers. (Fuente: `https://www.moneycrashers.com/factors-affect-used-cars-resale-value/`). LAs variables a resaltar en este artículo son las siguientes: **Factores económicos, marca del vehículo, forma del vehículo y clase del vehículo, kilometraje, tipo de transmisión, condición exterior y condición interior**. Exceptuando los factores económicos, la condición exterior y condición interior del automóvil, las demás variables coinciden con las que se tienen en cuenta para efectos de este trabajo.

En el cuarto artíulo, publicado por CarSellZone (una empresa dedicada a la venta de automóviles), destacan los siguientes factores que afectan el precio de un automóvil: **Kilometraje, condición interior y exterior, marca del vehículo y modelo, condición mecánica, historial de mantenimiento, historial de accidentes** (Fuente: `https://carsellzone.com/blog/detail/factors-affect-car-price`). Cabe aclarar que, aunque estos factores se identificaron en el artículo para carros usados, se tienen variables en común, tales como kilometraje, marca del vehículo y modelo.

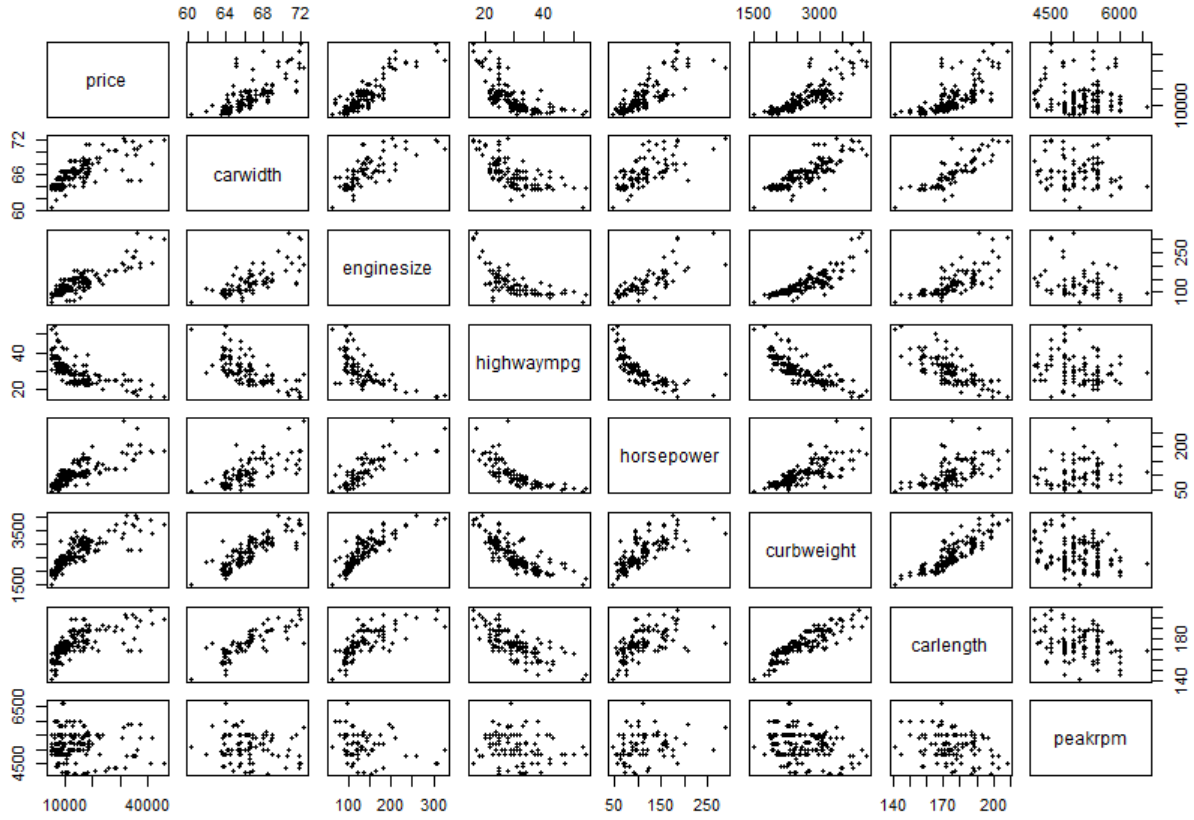
Finalmente, en el último artículo, tomado de una empresa del Reino Unido (cashpoint4cars) se destacan las siguientes variables, tanto para carros nuevos, como para carros usados: **Años, kilometraje, condición,**

número de dueños, historial de mantenimiento, marca, tipo de combustible, fiabilidad del modelo y modificaciones (Fuente: `https://www.cashpoint4cars.co.uk/blog/15-factors-affect-price-your-car/`). De estas variables, se tienen en común el kilometraje, marca y tipo de combustible.

Ahora bien, después de realizar el análisis de literatura correspondiente a este trabajo, se va a llevar a cabo todo el análisis descriptivo con las principales variables cuantitativas y cualitativas. Esto con el fin de determinar coeficientes de correlación, coeficientes de asociación, coeficientes parciales, entre otros.

Variables cuantitativas:

Primero veremos un scatterplot para cada par de variables cuantitativas.



Como se puede observar, en la primera fila, observamos que parece haber una fuerte relación entre el precio y las otras variables en la mayoría de las variables, menos en *peakrpm*. Es necesario ver los coeficientes de correlación de Pearson, Kendall, Spearman y Xi, para poder concluir qué tipo de relación corresponde, así como los coeficientes de correlación parciales, para ver realmente el aporte de cada variable explicando el precio de un automóvil.

Pearson	Price	Carwidth	Enginesize	Highwaympg	Horsepower	Curbweight	Carlength	Peakrpm
Price	1.00	0.76	0.87	-0.70	0.81	0.84	0.68	-0.09
Carwidth	0.76	1.00	0.74	-0.68	0.64	0.87	0.84	-0.22
Enginesize	0.87	0.74	1.00	-0.68	0.81	0.85	0.68	-0.24
highwaympg	-0.70	-0.68	-0.68	1.00	-0.77	-0.80	-0.70	-0.05
Horsepower	0.81	0.64	0.81	-0.77	1.00	0.75	0.55	0.13
Curbweight	0.84	0.87	0.85	-0.80	0.75	1.00	0.88	-0.27
Carlength	0.68	0.84	0.68	-0.70	0.55	0.88	1.00	-0.29
Peakrpm	-0.09	-0.22	-0.24	-0.05	0.13	-0.27	-0.29	1.00

Kendall	Price	Carwidth	Enginesize	Highwaympg	Horsepower	Curbweight	Carlength	Peakrpm
Price	1.00	0.64	0.66	-0.64	0.67	0.74	0.63	-0.04
Carwidth	0.64	1.00	0.59	-0.53	0.51	0.69	0.72	-0.14
Enginesize	0.66	0.59	1.00	-0.58	0.65	0.72	0.60	-0.19
Highwaympg	-0.64	-0.53	-0.58	1.00	-0.73	-0.67	-0.53	-0.04
Horsepower	0.67	0.51	0.65	-0.73	1.00	0.62	0.48	0.08
Curbweight	0.74	0.69	0.72	-0.67	0.62	1.00	0.72	-0.16
Carlength	0.63	0.72	0.60	-0.53	0.48	0.72	1.00	-0.18
Peakrpm	-0.04	-0.14	-0.19	-0.04	0.08	-0.16	-0.18	1.00

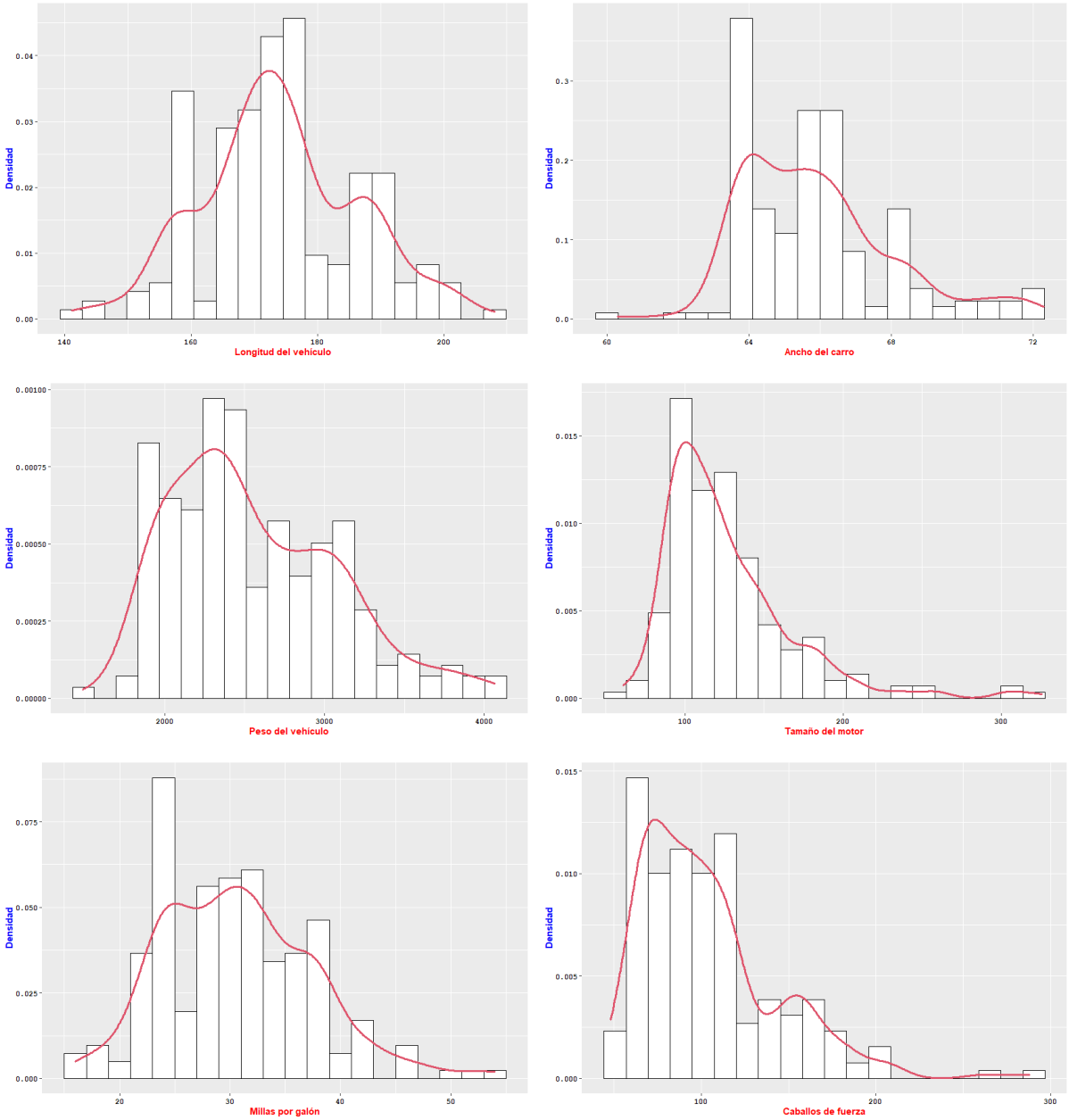
Spearman	Price	Carwidth	Enginesize	Highwaympg	Horsepower	Curbweight	Carlength	Peakrpm
Price	1.00	0.81	0.83	-0.82	0.85	0.91	0.80	-0.07
Carwidth	0.81	1.00	0.77	-0.70	0.69	0.86	0.89	-0.20
Enginesize	0.83	0.77	1.00	-0.72	0.82	0.88	0.78	-0.27
Highwaympg	-0.82	-0.70	-0.72	1.00	-0.89	-0.83	-0.70	-0.06
Horsepower	0.85	0.69	0.82	-0.89	1.00	0.81	0.66	0.11
Curbweight	0.91	0.86	0.88	-0.83	0.81	1.00	0.89	-0.24
Carlength	0.80	0.89	0.78	-0.70	0.66	0.89	1.00	-0.27
Peakrpm	-0.07	-0.20	-0.27	-0.06	0.11	-0.24	-0.27	1.00

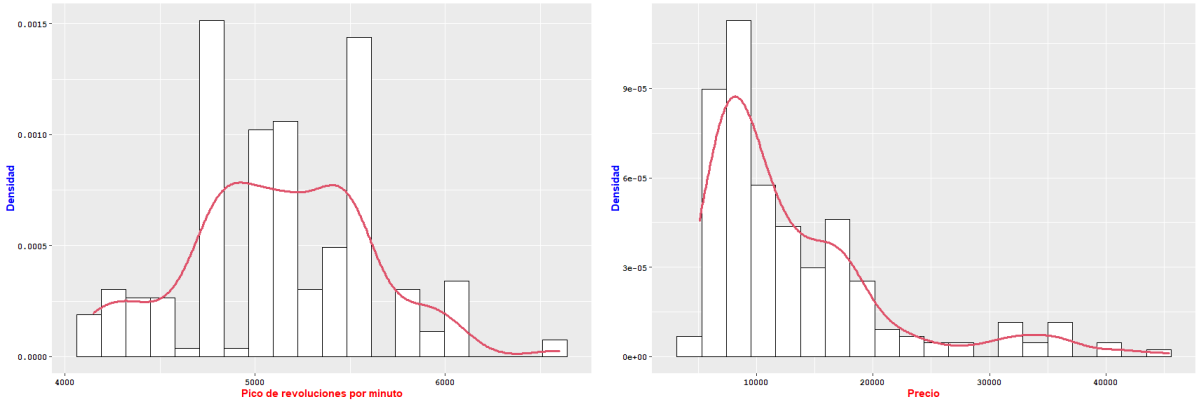
Coef. Xi	Price	Carwidth	Enginesize	Highwaympg	Horsepower	Curbweight	Carlength	Peakrpm
Price	0.99	0.49	0.54	0.50	0.54	0.62	0.47	0.07
Carwidth	0.59	0.99	0.61	0.43	0.49	0.64	0.73	0.21
Enginesize	0.59	0.57	0.99	0.56	0.59	0.67	0.57	0.36
Highwaympg	0.56	0.38	0.48	0.99	0.61	0.56	0.38	0.18
Horsepower	0.63	0.59	0.70	0.68	0.99	0.64	0.56	0.55
Curbweight	0.64	0.59	0.64	0.58	0.55	0.99	0.64	0.19
Carlength	0.58	0.80	0.67	0.51	0.50	0.71	0.99	0.36
Peakrpm	0.15	0.15	0.25	0.09	0.13	0.16	0.19	0.99

C. Parcial	Price	Carwidth	Enginesize	Highwaympg	Horsepower	Curbweight	Carlength	Peakrpm
Price	1.00	0.17	0.45	0.08	0.15	0.18	-0.04	0.21
Carwidth	0.17	1.00	-0.04	0.10	0.03	0.27	0.36	0.03
Enginesize	0.45	-0.04	1.00	0.11	0.44	0.21	-0.04	-0.39
Highwaympg	0.08	0.10	0.11	1.00	-0.34	-0.34	-0.19	-0.24
Horsepower	0.15	0.03	0.44	-0.34	1.00	0.17	-0.26	0.44
Curbweight	0.18	0.27	0.21	-0.34	0.17	1.00	0.48	-0.30
Carlength	-0.04	0.36	-0.04	-0.19	-0.26	0.48	1.00	-0.02
Peakrpm	0.21	0.03	-0.39	-0.24	0.44	-0.30	-0.02	1.00

En la primera columna de cada tabla encontramos la información más útil, pues corresponde al coeficiente de correlación del precio y otra variable. Los coeficientes de correlación de Pearson y Spearman resultan tener una magnitud muy grande, los de Kendall tienen una magnitud considerable pero no tan grande como los de Pearson y Spearman. El hecho de que los coeficientes de correlación de Spearman sean más grandes que los de Pearson significa las variables no se distribuyen normalmente o que la correlación de las variables parece ser monótona más no lineal, pero como los coeficientes de correlación de Kendall no son tan grandes como los de Pearson, resulta ser la primera opción.

Veamos como se distribuyen los datos para ver si en una primera instancia, es cierto que su distribución no es normal.





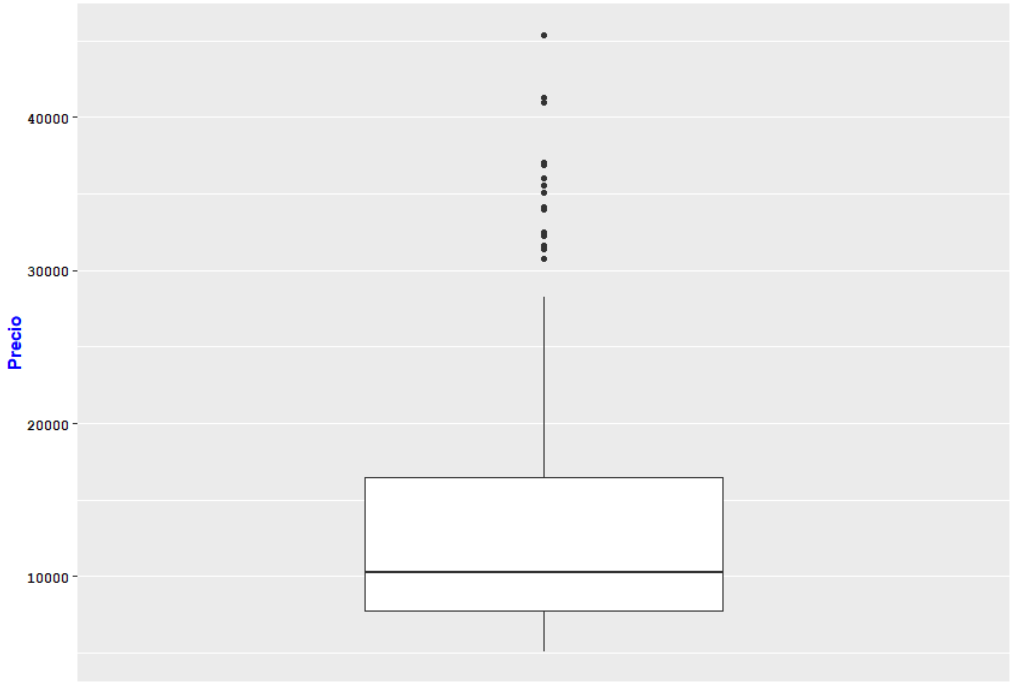
Como se suponía, parece que los datos no se distribuyen de forma normal, lo que hace que el coeficiente de correlación de Spearman sea mayor al de Pearson, pues se usa específicamente cuando los datos no están distribuidos normalmente o con variables cualitativas ordinales.

Además, los coeficientes de correlación Xi no son tan grandes, como los de Pearson o Spearman, por lo que no parece que haya una relación funcional no lineal entre las variables asociadas.

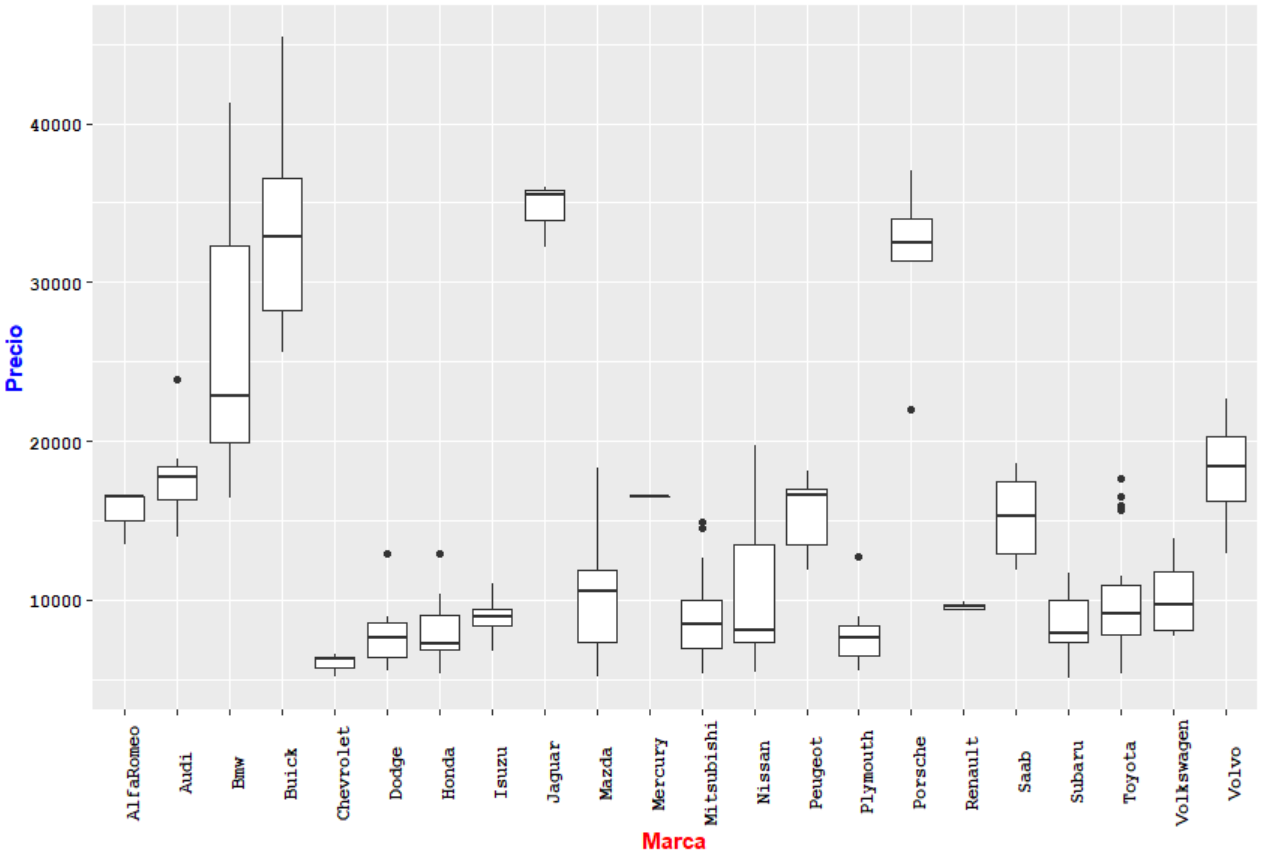
Cabe resaltar que la correlación de algunas variables es muy alta, lo que puede llevar a problemas de multicolinealidad, por lo que la tabla de coeficientes de correlaciones parciales nos dicen cuales variables realmente aportan explicando el precio. En este caso, aunque la correlación entre precio y las demás variables es generalmente alta, la correlación parcial entre precio y variables como *highwaympg*, *wheelbase* o *citympg* es casi nula, por lo que realmente no aportan mucho si tenemos en cuenta variables como *carwidth*, *enginesize* y *horsepower*.

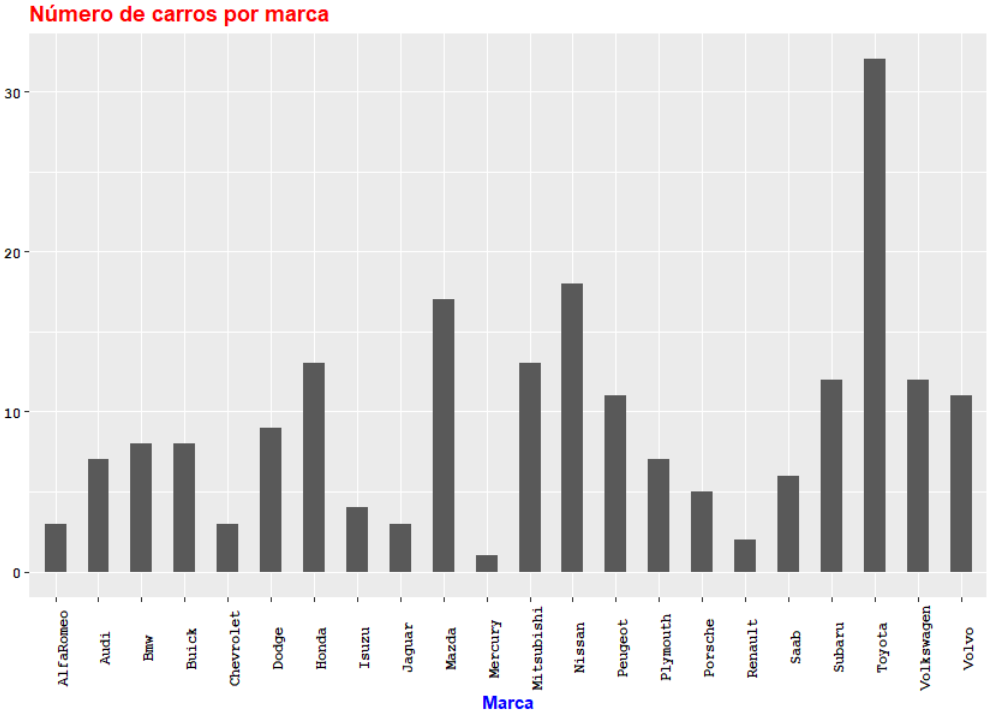
Variables cualitativas:

Veamos qué tanto se relacionan las variables cualitativas con el precio. Para esto, lo haremos con boxplots, pero primero, también veamos un boxplot de la variable precio.

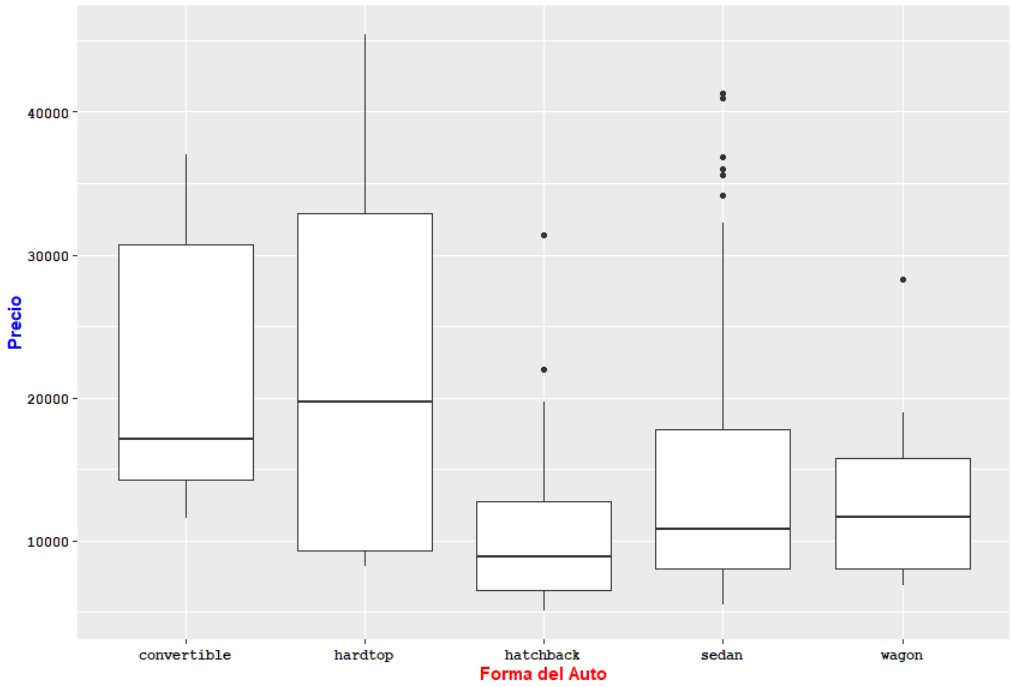


Como vemos, hay algunos precios que son atípicos, concretamente, hay carros que son mucho más caros que la mayoría de la muestra. Esto es importante pues servirá luego para ver si hay datos con alta influencia, concretamente, outliers.

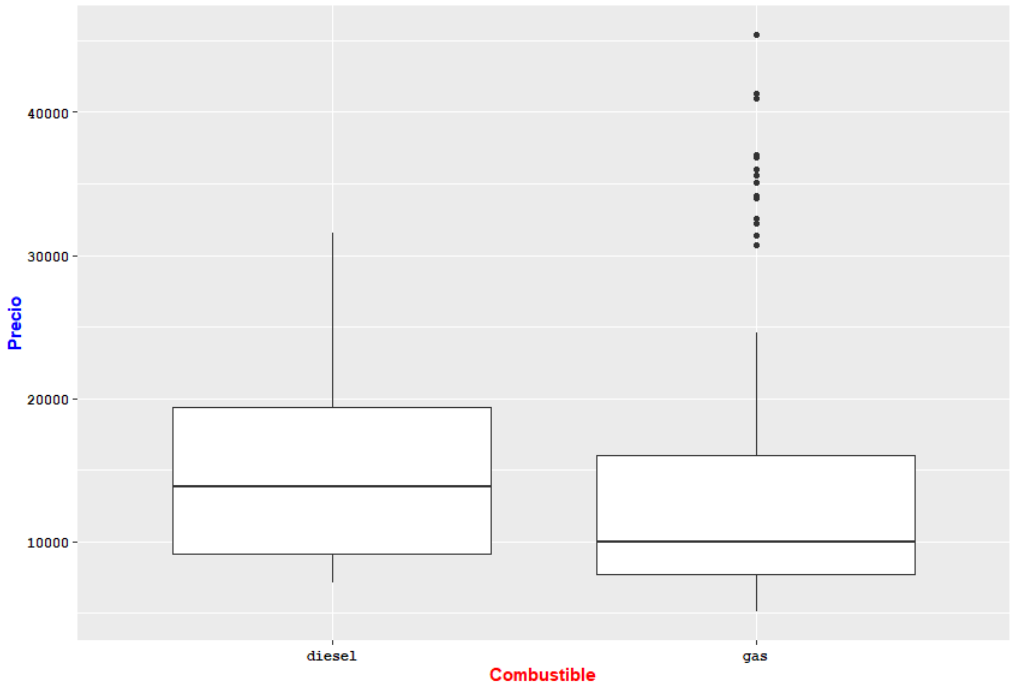




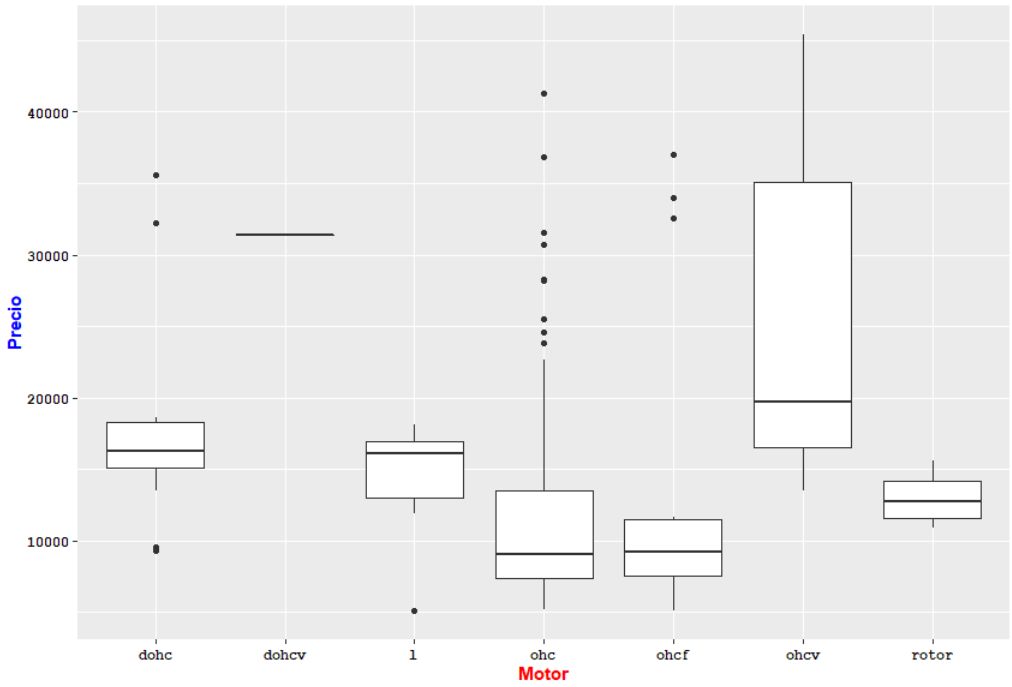
Primero hay que resaltar que algunas marcas tienen muy pocas observaciones, por esta razón apenas se aprecian en el boxplot. Hay unas cuantas observaciones atípicas en algunas marcas. Pese a que las cajas de algunas marcas están en el mismo rango, otras están tomando un rango de precios distinto y muy notorio, por lo que puede que algunas marcas si tengan un impacto respecto al precio del automóvil.



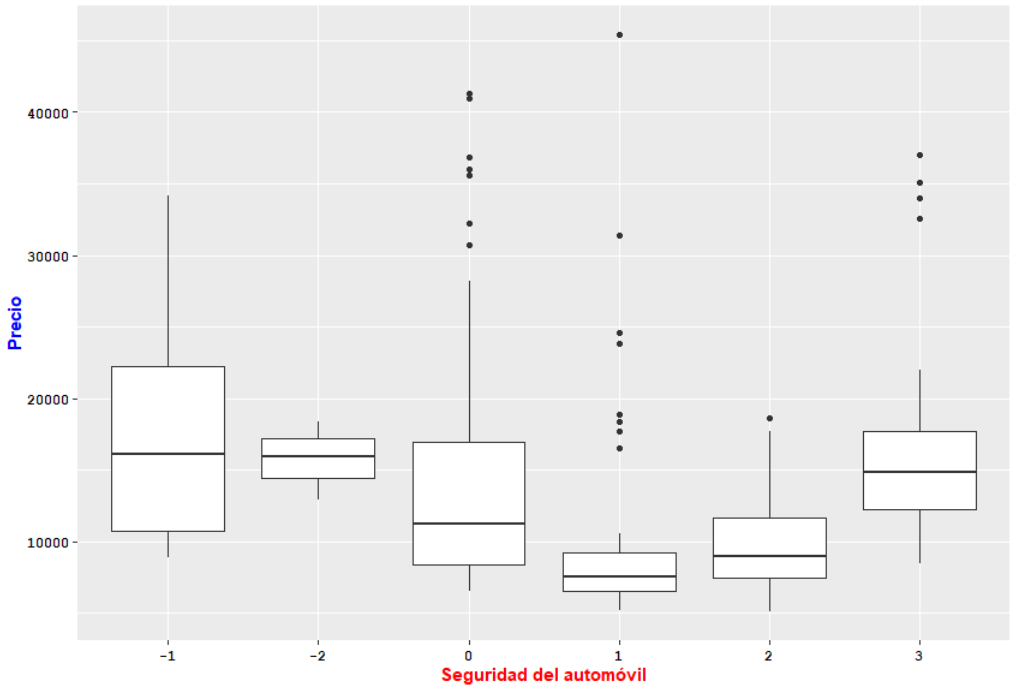
Sólo hay datos atípicos en los carros *hatchback*, *sedan* y uno en *wagon*. Las medias de los carros convertibles están cerca y las de *hatchback*, *sedan* y *wagon* también están relativamente cerca, pero las cajas tienen distintos tamaños, por lo que puede que haya una relación entre el precio y la forma del auto. El número de observaciones por marca se muestra en el siguiente gráfico.



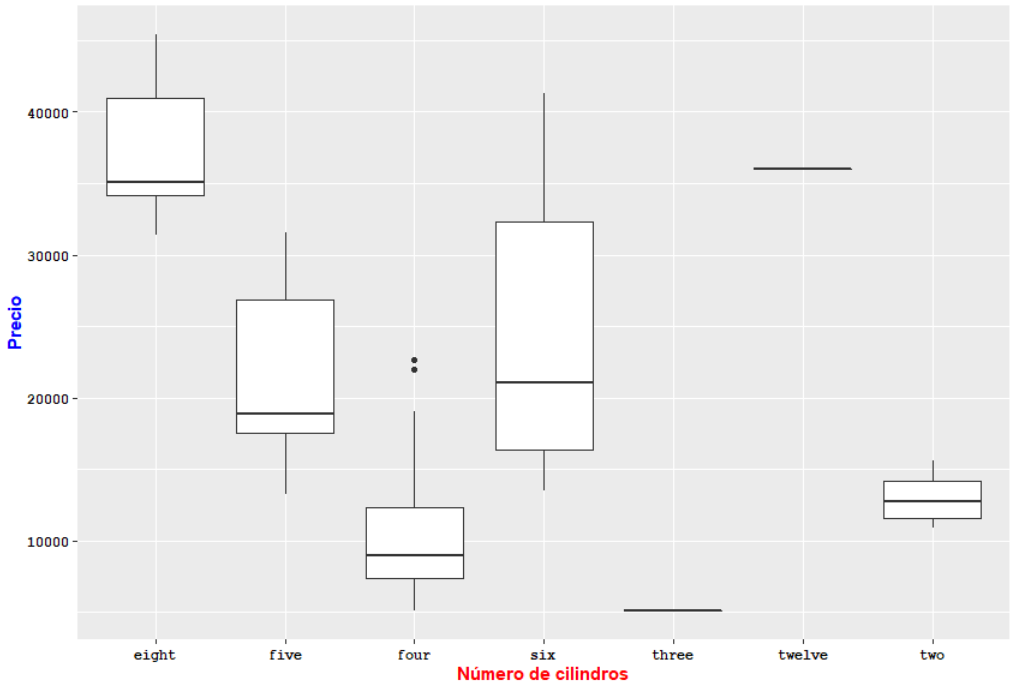
Hay atipicidad sólo en los carros que usan gasolina, y es notoria. Ahora bien, las medias y las cajas no son muy distintas, por lo que en una primera instancia, podríamos decir que no hay una relación entre el precio del automóvil y el combustible que usa, o que en caso de que la haya, no afecta enormemente en el precio.



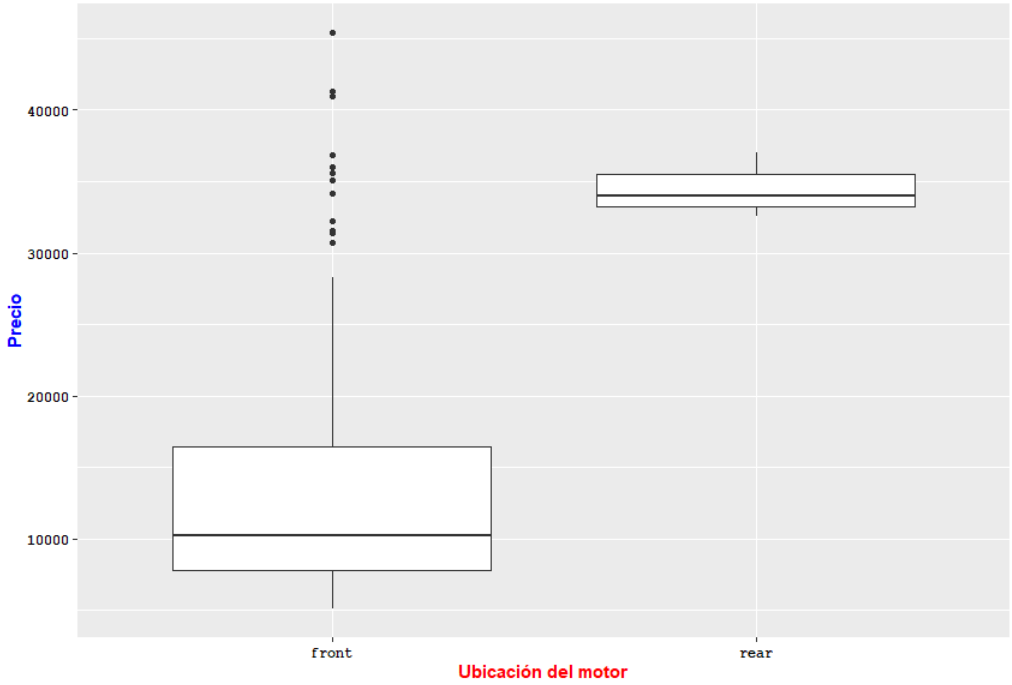
En este caso, hay datos atípicos en carros con motor *dohc, l, ohc* y *ohcf*, hay muy pocas observaciones para *dohcv*; cabe resaltar que hay atipicidad notoria en los carros con motor *ohc*. Las cajas no parecen estar muy alejadas del mismo rango de valores, excepto por el tipo de motor *ohcv*, la cual está en un rango de precios distintos, por lo que puede haber una relación entre este tipo de motor y el precio.



Hay atipicidad en los carros con un nivel 0, 1 y 3 de seguridad (los cuales corresponden a los carros con un nivel medio-alto de seguridad), la cual es notoria en los carros con nivel 0 y 1 de seguridad.



Apenas hay carros con 3 y 12 cilindros, por lo que pueden ocasionar problemas. De ahí las cajas son muy distintas y toman intervalos también muy distintos, por lo que es una variable que se debe considerar en el modelo.



Las dos cajas son muy distintas, parece que la ubicación del motor influye mucho en el precio del vehículo.

Visto esto, podemos decir que en la muestra hay unos cuantos datos atípicos, los cuales tendrán que ser tomados en cuenta por si llegan a ser datos influyentes en nuestro modelo.

En las fuentes mencionadas, las principales variables que se usan son la marca, forma del vehículo, tamaño del motor. En nuestras diez variables iniciales no está incluida la marca. Sin embargo, luego de la

búsqueda bibliográfica, se llegó a la conclusión de que no incluirla puede generar problemas con el modelo, como efectos de confusión, por lo que se cambió la variable *"Aspiration"* por la variable *"Marca"*. Además, agregamos el ancho del automóvil y el número de cilindros, que tampoco estaban en nuestras 10 variables iniciales, puesto que creemos que pueden aportar información pertinente explicando el precio. Para el uso de variables dummys en el modelo se tuvieron en cuenta las que tuvieran por lo menos cuatro observaciones en el respectivo factor.

3. Fase de Estimación/Identificación:

Después de hacer una serie de pruebas para encontrar un modelo que superara de la mejor manera los supuestos de regresión lineal clásica, optamos por elegir uno con la siguiente estructura donde x_i con $i \in \{\ 3, .., 27\}$ representa alguna variable dummy de la categoría seleccionada:

$$\begin{cases} Y_k = \mu_k + e_k \\ \mu_k = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 x_{k3} \cdots + \beta_{27} x_{k27} \\ e_k \sim N(0, \sigma^2) \\ e_1, e_2, ..., e_k \quad \text{independientes} \end{cases}$$

Notamos a las variables x_{k1} y x_{k2} como el ancho del automovil y el tamaño del motor respectivamente. Cabe aclarar que preferimos perder un poco de interpretabilidad al introducir el tamaño del motor al cuadrado ante aceptar la presencia de un patrón no explicado en los residuales,problema que se resuelve introduciendo esta variable
Procedemos a indicar como identificamos las variables dummys en nuestro modelo:

Variables base para la categoría marca:

$$x_{k3} = \begin{cases} 1, \text{Audi} \\ 0, \quad \text{e.o.c.} \end{cases} \qquad x_{k4} = \begin{cases} 1, \text{Bmw} \\ 0, \quad \text{e.o.c.} \end{cases} \qquad x_{k5} = \begin{cases} 1, \text{Buick} \\ 0, \quad \text{e.o.c.} \end{cases}$$

$$x_{k6} = \begin{cases} 1, \text{Honda} \\ 0, \quad \text{e.o.c.} \end{cases} \qquad x_{k7} = \begin{cases} 1, \text{Mazda} \\ 0, \quad \text{e.o.c.} \end{cases} \qquad x_{k8} = \begin{cases} 1, \text{Mitsubishi} \\ 0, \quad \text{e.o.c.} \end{cases}$$

$$x_{k9} = \begin{cases} 1, \text{Jaguar} \\ 0, \quad \text{e.o.c.} \end{cases} \qquad x_{k10} = \begin{cases} 1, \text{Nissan} \\ 0, \quad \text{e.o.c.} \end{cases} \qquad x_{k11} = \begin{cases} 1, \text{Peugeot} \\ 0, \quad \text{e.o.c.} \end{cases}$$

$$x_{k12} = \begin{cases} 1, \text{Plymouth} \\ 0, \quad \text{e.o.c.} \end{cases} \qquad x_{k13} = \begin{cases} 1, \text{Porsche} \\ 0, \quad \text{e.o.c.} \end{cases} \qquad x_{k14} = \begin{cases} 1, \text{Saab} \\ 0, \quad \text{e.o.c.} \end{cases}$$

$$x_{k15} = \begin{cases} 1, \text{Subaru} \\ 0, \quad \text{e.o.c.} \end{cases} \qquad x_{k16} = \begin{cases} 1, \text{Dodge} \\ 0, \quad \text{e.o.c.} \end{cases} \qquad x_{k17} = \begin{cases} 1, \text{VolksW.} \\ 0, \quad \text{e.o.c.} \end{cases}$$

$$x_{k18} = \begin{cases} 1, \text{Volvo} \\ 0, \quad \text{e.o.c.} \end{cases} \qquad x_{k19} = \begin{cases} 1, \text{Isuzu} \\ 0, \quad \text{e.o.c.} \end{cases}$$

Variables base para la categoría forma del vehículo:

$$x_{k20} = \begin{cases} 1, \text{Hatchback} \\ 0, \quad \text{e.o.c.} \end{cases} \qquad x_{k21} = \begin{cases} 1, \text{Wagon} \\ 0, \quad \text{e.o.c.} \end{cases}$$

$$x_{k22} = \begin{cases} 1, \text{Convertible} \\ 0, \quad \text{e.o.c.} \end{cases} \qquad x_{k23} = \begin{cases} 1, \text{Sedan} \\ 0, \quad \text{e.o.c.} \end{cases}$$

Variables base para la categoría número de cilindros:

$$x_{k24} = \begin{cases} 1, \text{Cuatro} \\ 0, \quad \text{e.o.c.} \end{cases} \qquad x_{k25} = \begin{cases} 1, \text{Seis} \\ 0, \quad \text{e.o.c.} \end{cases} \qquad x_{k26} = \begin{cases} 1, \text{Ocho} \\ 0, \quad \text{e.o.c.} \end{cases}$$

Variable de la categoría ubicación del motor:

$$x_{k27} = \begin{cases} 1, \text{Atrás} \\ 0, \quad \text{e.o.c.} \end{cases}$$

Estimación de los parámetros

Parámetro	Estimación	Std. Error Not. Científica	P-value
Intercepto	-63434.16	1.018e+04	3.23e-09***
Carwidth	1115.80	1.586e+02	4.16e-11***
Enginesize	0.19	2.901e-02	2.94e-10***
Marca Audi	1801.26	1.227e+03	0.143865
Marca Bmw	10181.33	9.485e+02	2e-16***
Marca Buick	6777.28	1.483e+03	9.11e-06***
Marca Honda	379.05	6.857e+02	0.581129
Marca Mazda	745.85	6.593e+02	0.259481
Marca Mitsubishi	-394.80	6.924e+02	0.569293
Marca Jaguar	4486.62	1.999e+03	0.026019*
Marca Nissan	-596.06	6.208e+02	0.338272
Marca Peugeot	1437.98	9.025e+02	0.112855
Marca Plymouth	-56.82	8.723e+02	0.948139
Marca Porsche	6898.55	1.845e+03	0.000249***
Marca Saab	4132.15	9.638e+02	2.97e-05***
Marca Subaru	-219.14	7.125e+02	0.758764
Marca Dodge	124.00	7.890e+02	0.875298
Marca Volkswagen	-109.74	7.311e+02	0.880859
Marca Volvo	3690.80	8.390e+02	1.87e-05***
Marca Isuzu	1827.25	1.129e+03	0.107218
Carbody hatchback	-367.26	9.405e+02	0.696655
Carbody wagon	-271.41	9.918e+02	0.784671
Carbody convertible	4594.69	1.163e+03	0.000113***
Carbody sedan	-130.02	9.221e+02	0.888031
Cylindernumber eight	199.66	1.649e+03	0.903780
Cylindernumber four	-2264.92	8.169e+02	0.006158**
Cylindernumber six	265.87	1.026e+03	0.795756
Enginelocation rear	9443.70	2.558e+03	0.000297***

Interpretación de parámetros

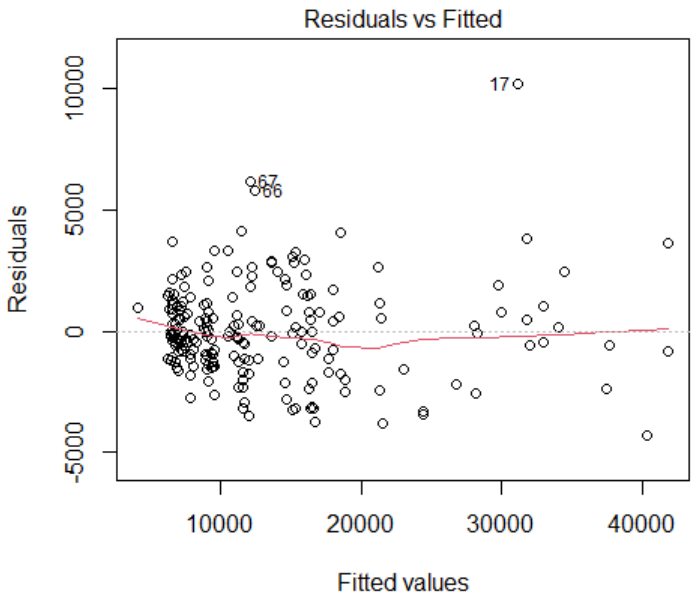
- β_0: No tiene interpretación. Un automóvil sin motor, peso, etc.

- β_1: Cambio promedio en el precio de un automóvil por cada centímetro adicional de anchura del chasis considerando constante el comportamiento de las demás variables.
- β_2:Cambio promedio en el precio de un automóvil por cada centímetro cúbico adicional en el tamaño del motor considerando constante el comportamiento de las demás variables.
- β_3: Cambio promedio en el precio de un automóvil de marca Audi con respecto a uno que no sea de esta marca considerando constante el comportamiento de las demás variables.
- β_{20}: Cambio promedio en el precio de un automóvil de marca Isuzu con respecto a uno que no sea de esta marca considerando constante el comportamiento de las demás variables.
- β_{22}: Cambio promedio en el precio de un automóvil de tipo Convertible frente a uno que no sea de esta marca considerando constante el comportamiento de las demás variables.
- β_{23}: Cambio promedio en el precio de un automóvil de tipo sedán frente a uno que no sea de esta marca considerando constante el comportamiento de las demás variables.
- β_{25}: Cambio promedio en el precio de un automóvil con 6 cilindros en su motor frente a uno que que tenga un número distinto de cilindros considerando constante el comportamiento de las demás variables.
- β_{26}: Cambio promedio en el precio de un automóvil con 6 cilindros en su motor frente a uno que que tenga un número distinto de cilindros considerando constante el comportamiento de las demás variables.
- β_{27}: Cambio promedio en el precio de un automóvil con motor ubicado en su parte trasera frente a uno que no considerando constante el comportamiento de las demás variables.

4. Fase de Validación

Patrones no explicados

El modelo parece no presentar patrones no explicados en su gráfico de residuales vs valores ajustados, o si existe parece ser ligero. Además, al realizar la prueba RESET ("fitted"), da un *p-value* = 0.09784 y con un 95 % de significancia no se rechaza la hipótesis nula, es decir, no hay evidencia suficiente para que se necesiten transformaciones cuadráticas o cúbicas de $\hat{y}$ en el modelo.



Durante la fase de Estimación/identificación, para cumplir este supuesto, la variable **enginesize** se transformó por $enginesize^2$, ya que el gráfico de residuales vs valores ajustados presentaba un patrón no lineal y esta era la variable con coeficiente de correlación parcial más alto. A pesar de perder un poco la interpretación del tamaño del motor, se pudo validar este supuesto con el test.

Multicolinealidad
El modelo parece no presentar problemas de multicolinealidad, ya que calculándole el factor de inflación de varianza a los coeficientes, no hay ninguno que sea demasiado alto(> 10). Por lo tanto, se espera que no haya inestabilidad en $\hat{B}$.

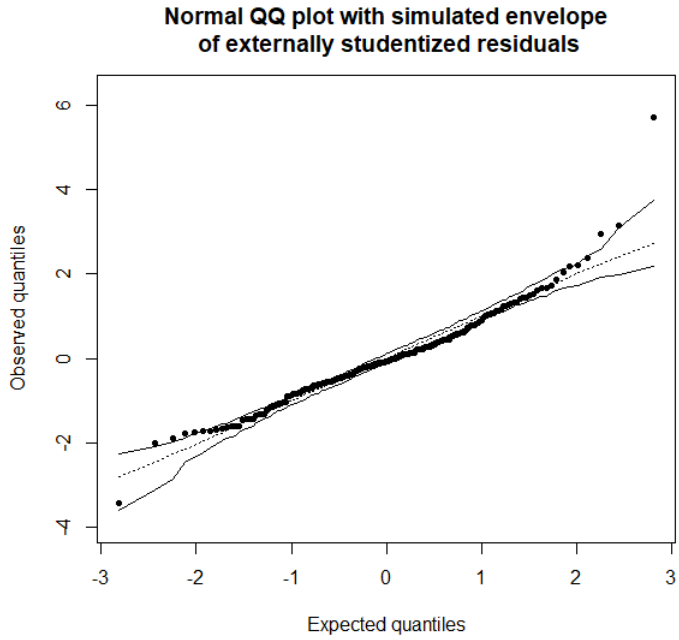
Coeficiente	**VIF**	**Coeficiente**	**VIF**
Carwidth	5.361460	Enginesize	7.945180
Marca Audi	2.311845	Marca Bmw	1.570902
Marca Buick	3.839899	Marca Honda	1.300458
Marca Mazda	1.539295	Marca Mitsubishi	1.325978
Marca Jaguar	2.682163	Marca Nissan	1.437169
Marca Peugeot	1.925655	Marca Plymouth	1.168570
Marca Porsche	3.771886	Marca Saab	1.228847
Marca Subaru	1.302580	Marca Dodge	1.216696
Marca Volkswagen	1.371594	Marca Volvo	1.664226
Marca Isuzu	1.134652	Carbody hatchback	9.2621155
Carbody wagon	4.904526	Carbody convertible	1.790327
Carbody sedan	9.858467	Cylindernumber eight	3.013830
Cylindernumber four	5.407940	Cylindernumber six	5.062452
Enginelocation rear	4.394588		

En la fase de modelación se eliminó la variable horsepower debido a que estaba muy correlacionada con el tamaño del motor y traía problemas de multicolinealidad.
Homoscedasticidad
Mediante el test Breusch-Pagan juzgamos el sistema de hipótesis:

$$\begin{cases} H_0 : \text{hay homoscedasticidad (igual varianza en los residuales)} \\ H_1 : \text{no hay homoscedasticidad} \end{cases}$$

Donde obtenemos un $p\text{-}value = 0.002744$. Así que con un 95 % de significancia se rechaza la hipótesis nula y este supuesto falla. Para corregir la heteroscedasticidad durante el proceso de estimación, se intentó modelar la varianza y hacer un modelo con pesos. Sin embargo, esto generó un problema de multicolinealidad bastante alto en algunos coeficientes de variables dummy, por lo cual no se validó aquel modelo. También se intentó realizar una transformación Box-Cox, que redujo un poco el *p-value* del test Breusch-Pagan a 0.01219 pero tampoco se validó, y dado que el modelo pretende ser explicativo, esta tampoco es la mejor opción por interpretabilidad. Por lo tanto, dado que este supuesto falló, en caso de que se quiera realizar inferencia sobre los parámetros, se debería usar la matriz de varianzas y covarianzas robusta ya que $\hat{\beta}$ no tendrá eficiencia.

Normalidad
Para evaluar el supuesto de normalidad en los residuales, a continuación se muestra el siguiente QQ-Plot simulado, ya que no se pueden suponer una muestra aleatoria.



Como se puede observar, parece que no se cumple el supuesto de normalidad en los residuales. Sin embargo, n=205 es un buen número para tener una normalidad asintótica. Si falla este supuesto lo que pasa es que $\hat{\beta}$ no será un UMVUE, pero eso no se tenía debido a que no se validó la homoscedasticidad y por tanto no es eficiente, no tiene varianza mínima.

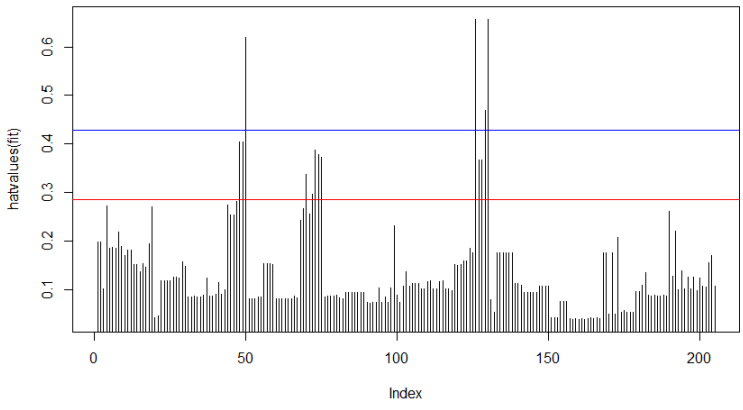
No independencia:

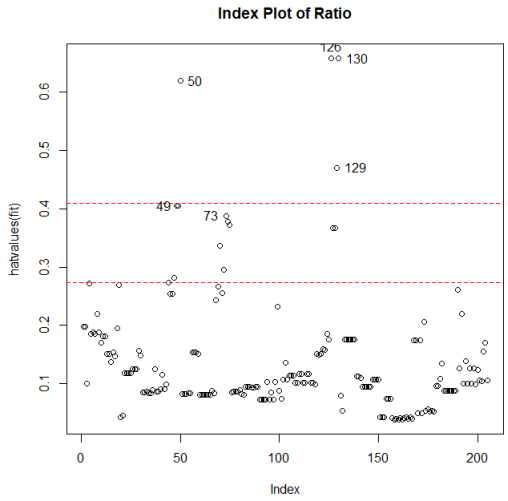
Para la evaluación de este supuesto, teniendo en cuenta que el estudio es de carácter transversal, mas no longitudinal. Este estudio no fue medido en el tiempo.

5. Observaciones de alta palanca, atípicas e influyentes:

Observaciones de alto apalancamiento:

Usamos el criterio de buscar valores para los cuales $h_{ii} > \frac{2p}{n}$ o $h_{ii} > \frac{3p}{n}$. Par identificarlas nos apoyaremos en los siguientes gráficos:

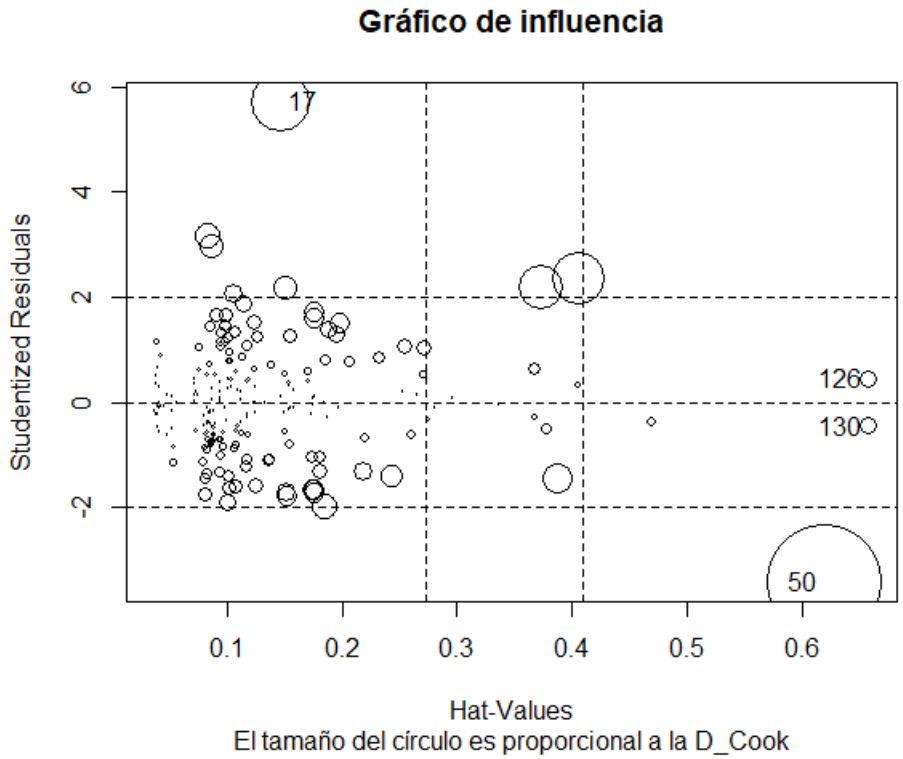




Donde identificamos que las observaciones 50,126,129,130 resultan ser de alta palanca. Para las observaciones atípicas buscamos donde se cumpla que $|r_{s,i}| > 3$ con $r_{s,i} = \frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$.De donde observamos que 17, 50 y 67 son observaciones atípicas.

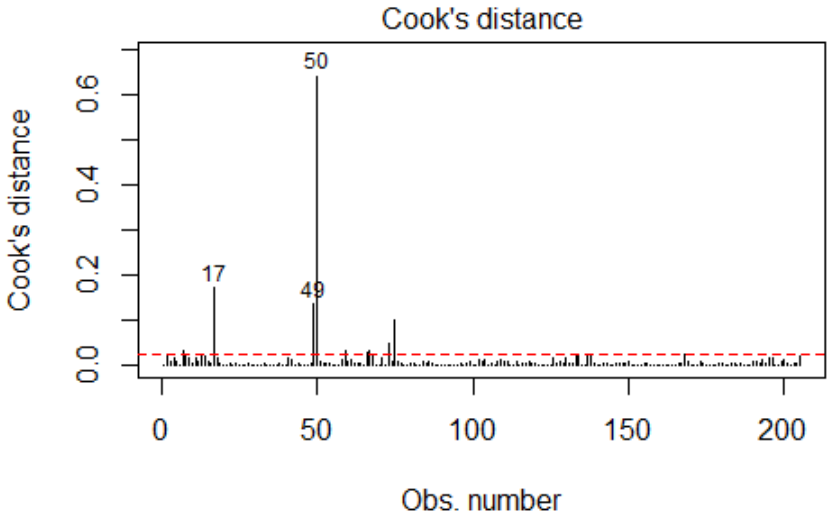
Observación	17	50	67	66	49	75
$\|r_{s,i}\|$	5.71	3.42	3.16	2.96	2.38	2.19

Finalmente, presentamos algunas observaciones influyentes en la siguiente tabla:



Observación	7	17	49	50	59	66	67	68	73	75
Distancia de Cook	0.03	0.169	0.133	0.64	0.029	0.028	0.03	0.023	0.046	0.1

Se pueden ver estos datos por su distancia de Cook en el siguiente gráfico.



Escogeremos para analizar especialmente las observaciones 50 y 17 pues son las que resultan parecer más influyentes para el modelo. La observación 50 es bastante influyente debido a que es atípica y tiene muy alto leverage, se puede observar que tiene el valor máximo en enginesize, lo cual la hace de alto leverage, pero además es la única con un número de cilindros de doce, lo cual es extraño, a lo que el modelo le ajusta un valor alto pero la observación real no es de un precio tan elevado, por eso además es un poco atípica. La observación 17 no es de alto leverage ya que está entre los valores normales entre las variables explicativas, pero el precio sí es demasiado elevado, siendo el segundo más alto, lo cual la hace un dato demasiado atípico. Los otros datos influyentes son debido a que tienen alto leverage pero poca atipicidad, o viceversa, pero no llegan a ser tan influyentes como la 17 y la 50.
Creemos que dado el modelo, las observaciones influyentes sí afectaron un poco la estimación de los parámetros ya que para explicar un poco mejor, la variable enginesize tuvo que elevarse al cuadrado y eso hizo que los datos atípicos en enginesize se hicieran más notorios, además de generar confusión en las variables cualitativas a la que pertenecía cada uno.

Modelo sin observaciones influyentes:

Al eliminar las observaciones influyentes y reestimar el modelo, obtenemos un nuevo modelo, el cual compararemos con el anterior en la siguiente tabla:

Parámetro	Estimación	Std. Error	t-value	P-value
Intercept	-6.832e+04	8.488e+03	-8.048	1.49e-13***
Carwidth	1.196e+03	1.325e+02	9.026	4.16e-16***
Enginesize	1.725e-01	3.102e-02	5.562	1.04e-07***
Marca Audi	1.582e+03	1.024e+03	1.544	0.1244
Marca Bmw	9.247e+03	7.753e+02	11.927	2e-16***
Marca Buick	7.076e+03	1.372e+03	5.159	6.97e-07***
Marca Honda	2.515e+02	5.436e+02	0.463	0.6442
Marca Mazda	-7.958e+02	5.578e+02	-1.427	0.1556
Marca Mitsubishi	-5.217e+02	5.496e+02	-0.949	0.3439
Marca Jaguar	6.221e+03	2.026e+03	3.070	0.0025**
Marca Nissan	-3.601e+02	4.921e+02	-0.732	0.4654
Marca Peugeot	1.273e+03	7.227e+02	1.761	0.0801 .
Marca Plymouth	-1.366e+02	6.903e+02	-0.198	0.8434
Marca Porsche	6.761e+03	1.498e+03	4.515	1.19e-05***
Marca Saab	3.936e+03	7.645e+02	5.149	7.31e-07***
Marca Subaru	-3.016e+02	5.644e+02	-0.534	0.5938
Marca Dodge	3.524e+01	6.244e+02	0.056	0.9551
Marca Volkswagen	-3.394e+02	5.815e+02	-0.584	0.5603
Marca Volvo	3.681e+03	6.701e+02	5.493	1.45e-07***
Marca Isuzu	1.818e+03	8.934e+02	2.035	0.0434*
Carbody hatchback	5.393e+02	7.845e+02	0.687	0.4928
Carbody wagon	4.265e+02	8.237e+02	0.518	0.6053
Carbody convertible	5.872e+03	9.973e+02	5.888	2.09e-08***
Carbody sedan	6.317e+02	7.808e+02	0.809	0.4197
Cylindernumber eight	-6.736e+02	1.671e+03	-0.403	0.6874
Cylindernumber four	-3.010e+03	7.304e+02	-4.121	5.93e-05***
Cylindernumber six	-9.797e+02	1.030e+03	-0.951	0.3427
Enginelocation rear	1.090e+04	2.048e+03	5.324	3.23e-07***

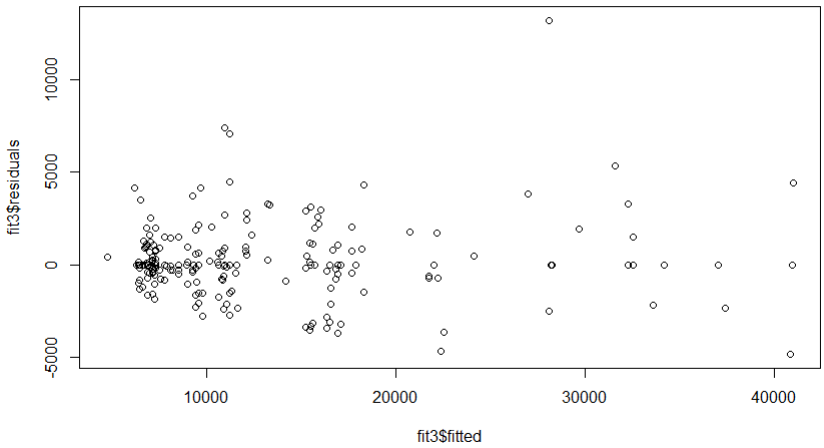
El modelo eliminando los valores influyentes si presenta cambios notorios respecto al modelo inicial, los cuales se ven en las estimaciones de variables como el intercepto. Además, una estimación pasa a ser significativas en este nuevo modelo, la cual es la estimación asociada a la marca Isuzu. De ahí, aunque las estimaciones cambian notablemente, también puede deberse a la unidad, y que realmente el cambio sea más bien pequeño. Sin embargo varias estimaciones pasaron de por ejemplo, 3000 a 4000, que es un cambio muy notorio, y más si tenemos en cuenta la cantidad de parámetros que estamos estimando.

Regression LAD:

Presentamos las estimaciones que nos da R sobre este modelo:

Paramétro	Estimación
Intercepto	-64758.49
Carwidth	1148.04
Marca Audi	2141.63
Marca Bmw	8367.53
Marca Buick	6569.49
Marca Honda	154.09
Marca Mazda	-175.43
Marca Mitsubishi	-40.95
Marca Jaguar	6695.58
Marca Nissan	70.08
Marca Peugeot	2287.42
Marca Plymouth	-115.31
Marca Porsche	7843.57
Marca Saab	4497.92
Marca Subaru	155.88
Marca Dodge	-115.31
Marca Volkswagen	-902.76
Marca Volvo	3649.57
Enginesize	0.17
Marca Isuzu	2338.80
Carbody hatchback	-375.16
Carbody wagon	72.00
Carbody convertible	4500.00
Carbody sedan	-92.45
Cylindernumber eight	900.27
Cylindernumber four	-2955.17
Cylindernumber six	-743.83
Enginelocation rear	9205.76

Sí parecen cambiar las estimaciones respecto a nuestro modelo original, sin embargo no parecen ser muy diferentes. Por otra parte, observamos que el gráfico de residuales parece tener la misma forma del estimado en el punto 5, por ello no parece haber presencia de patrones no explicados:



Para la prueba reset obtenemos el p-valor=0.09784 que resulta ser el mismo obtenido en el punto 5, así como el test de Breush- Pagan con p-valor=0.002744 y el número de condición de la matriz de diseño X es $2.58\times10^{1}2$. Concluimos que esta regresión hereda todas las propiedades de la regresión original.

6. Uso del modelo:

Para determinar si el modelo propuesto tiene valor agregado, para estudiar los factores que afectan el precio de un automóvil. Primero se debe evaluar la siguiente prueba de hipótesis:

$$\begin{cases} H_0 : \text{La regresión no tiene valor agregado} \\ H_1 : \text{La regresión tiene valor agregado} \end{cases}$$

En este caso, al realizar la prueba F de la significancia de la regresión se evalúa el siguiente test:
τ: *Rechazar* H_0 *si* $f_c > F_{1-\alpha}(p-1)(n-p)$.
Con *p-value* el test queda de la siguiente manera:

$$p - value = p[F_c > f_c | H_0]$$

A un nivel de significancia del 5 % se obtiene un *p-value* = 2.2e-16. Con 27 y 177 grados de libertad, se rechaza la hipótesis nula, por lo que el modelo de regresión propuesto tiene valor agregado.

Ahora bien, como no se validó la homoscedasticidad, para dar respuesta a la pregunta si los coeficientes asociados a las variables cualitativas son significativos, se usarán dos modelos, uno con todas las variables del modelo, y otro con sólo las variables cuantitativas del modelo. Estos se compararan con el Wald test:

$$\begin{cases} H_0 : \text{Modelo con sólo las variables cuantitativas(en nuestro modelo)} \\ H_1 : \text{Modelo original} \end{cases}$$

Usando la matriz de varianzas y covarianzas estimada de manera robusta. El resultado de este test es un p-valor=2.2e-16 ***. Por lo tanto, los coeficientes de las variables cualitativas son significativos para explicar la variable precio.

En el modelo de regresión propuesto, no fue necesario agregar una interacción de segundo orden en el modelo. Esto se debe a que al momento de elegir el modelo que explicara el precio de un automóvil, al momento de aplicar interacciones se corría el riesgo de no cumplir los supuestos de un modelo de regresión. De acuerdo a los datos obtenidos, al momento de elegir el modelo, se tuvieron bastantes inconvenientes con los supuestos de homoscedasticidad y multicolinealidad. Por lo que se prefirió realizar una transformación a una variable (perdiendo interpretación) que realizar interacciones que no permitieran el cumplimiento de los supuestos.

Finalmente, para dar respuesta al cumplimiento de los objetivos propuestos en la primera entrega, es pertinente recordar que los objetivos propuestos (junto con el cumplimiento o no cumplimiento de los mismos) para este trabajo son los siguientes:

General:

Estudiar los factores que influyen en el precio de un modelo de automóvil en el mercado estadounidense: Para el cumplimiento de este objetivo, se planteó el modelo propuesto en la fase de Estimación/Identificación, donde, según el desarrollo de esta entrega, identificamos que las variables que pueden influir en el precio de un automóvil son: marca del automovil, tamaño del motor, ubicación del motor y el cilindraje, a lo cual ya se mostró que la regresión era significativa.

Específicos:

- **Estudiar las posibles diferencias en el precio de un automóvil de acuerdo a las variables categóricas, como el tipo de combustible, tipo de motor, seguridad, aspiración o forma del auto:** Para el cumplimiento de este objetivo, se realizó dicho análisis en la sección de Fase de Identificación, donde en el análisis descriptivo (a través de Boxplots) se muestra la relación, influencia y diferencias entre cada una de las variables cualitativas con el precio de un automóvil (Para mas información, remitirse a la sección asociada).

- **Proponer futuros planes de negocio con base en la información relevante, para producir automóviles con un precio competitivo en el mercado:** Este objetivo queda pendiente por

cumplir, teniendo en cuenta que, para efectos de esta entrega, se busca construir un modelo con fines explicativos, mas no con fines predictivos.

- **Analizar si existe relación de algún tipo entre los caballos de fuerza y el tamaño del motor, además, observar si estas variables se ven afectadas por el tipo de motor del automóvil:** Para el cumplimiento de este objetivo, debemos remitirnos a la Fase de identificación, ya que en el análisis descriptivo de dichas variables, se puede observar su respectivo coeficiente de correlación (Pearson, Kendall, Spearman o Xi). Con ello, podemos dar cumplimiento a dicho objetivo.

Asimismo, cabe resaltar que el plan tentativo de análisis propuesto en la primera entrega se cumplió de forma parcial; debido a que la última actividad no se alcanzó a cumplir. Por lo tanto, se hará mención a cada uno de los puntos enunciados en el plan propuesto y se dará una respuesta sobre el cumplimiento o no cumplimiento del mismo:

- **Realizar un análisis descriptivo sobre las variables a considerar, graficando la información relevante de los datos e identificando datos atípicos en la muestra:** Esta actividad se completó en la Fase de Identificación (Para mas información, remitirse a la sección asociada). Respecto a la identificación de datos atípicos, tanto en la fase de identificación, como en la sección de observaciones de alta palanca, atípicas e influyentes.
- **Estudiar si las variables a considerar en el estudio tienen alguna relación entre sí:** Esta actividad se completó en la Fase de Identificación al momento de identificar los coeficientes de correlación de Pearson, Kendall y Spearman; así como los coeficientes Xi y los de correlación parcial (Para mas información, remitirse a la sección asociada).
- **Analizar la media del precio de un automóvil en las diferentes categorías de autos:** Esta actividad se completó en la fase de identificación, mostrando en un Boxplot un análisis de la media de los automóviles de acuerdo a la marca asociada (Para mas información, remitirse a la sección correspondiente).
- **Estimar las distribuciones de las variables cuantitativas analizadas en el estudio:** Para efectos de este trabajo, no se realizó una estimación de la distribución de las variables cuantitativas. Esto teniendo en cuenta que se consideró pertinente realizar un análisis descriptivo mas completo en aspectos como los coeficientes de correlación, identificación de outliers y análisis con el uso de boxplots para variables cualitativas.
- **Identificar un modelo que posiblemente permita predecir el precio de un automóvil, con base en las demás variables del conjunto de datos:** Para el cumplimiento de esta actividad remitirse a las Fases de Estimación/Identificación y a la Fase de Validación.
- **Estimar los parámetros del modelo seleccionado:** Esta actividad se cumple en la fase de Estimación/Identificación (Para mas información, remitirse a la sección correspondiente).
- **Identificar los outliers, si los hay, y manejarlos, para después realizar una mejor estimación de los parámetros:** Esta actividad se cumple en la sección de Observaciones de alta palanca, atípicas e influyentes (Para mas información, remitirse a la sección correspondiente).
- **Determinar cuáles variables tienen mayor y menor incidencia para el modelo:** Para el cumplimiento de esta actividad, se identificaron los coeficientes de correlación parcial (en variables cuantitativas) y el análisis en las variables cualitativas. Este análisis se observa en la Fase de Identificación (Para mas información, remitirse a la sección correspondiente).
- **Validar el modelo, verificando si se cumplen los supuestos y observando su rendimiento, determinando si los valores del modelo se ajustan bien al verdadero precio de los automóviles:** Esta actividad se cumple en la sección de la Fase de Validación. La totalidad del cumplimiento de la actividad se encuentra en esa sección (Para mas información, remitirse a la sección correspondiente).

- **Aplicar otros modelos para el precio de los automóviles y contrastar su rendimiento:** Para el cumplimiento de esta actividad todavía no es posible dar respuesta, teniendo en cuenta que, de acuerdo al alcance de esta entrega, se escogió el modelo que pasara la totalidad (o la mayoría) de los supuestos. Por lo que no se tenía contemplado realizar una comparación con otros modelos.

7. Bibliografía

Opción 1: (https://www.inchcapemotors.com.pe/mini/preciosautosnuevos)

Opción 2: (https://www.primicias.ec/nota_comercial/autos/garage/talleres/la-demanda-de-vehiculos-usado #gsc.tab=0)

Opción 3: (https://www.moneycrashers.com/factors-affect-used-cars-resale-value/)

Opción 4: (https://carsellzone.com/blog/detail/factors-affect-car-price)

Opción 5: (https://www.cashpoint4cars.co.uk/blog/15-factors-affect-price-your-car/)