

Selección e Implementación de Modelos de Analítica

Emmanuel Naranjo Blanco – A00835704; Fernanda Martínez Valles – A01722279; Valeria Rojas Minor – A011298657; Emiliano Vázquez Rodríguez – A00344781; Santiago Alonso Mendoza Franco - A00836973

Desarrollo de Proyectos de Análisis de Datos – IN1002B

Tecnológico de Monterrey – Campus Monterrey

Fecha de entrega: 03/12/2023

Profesor: Dr. Alexander Garrido Ríos

Resumen

En este proyecto se busca determinar los factores que influyen en el precio de venta de los vehículos de la marca Toyota usados a través de un análisis cuantitativo. Se utilizará un conjunto de datos proporcionado por el profesorado, y de la cual se emplearán modelos de analítica, como la regresión lineal múltiple y regresión logística, para identificar patrones e interacciones relevantes en los datos. El análisis abordará elementos clave, como la introducción al contexto y propósito, la descripción de datos, la verificación de supuestos, la especificación del modelo, el ajuste del modelo, la importancia de las variables, la interpretación del coeficiente, la evaluación de colinealidad, el análisis residual y la conclusión. Se prestará especial atención a la evaluación de la bondad del ajuste, la importancia de las variables mediante valores p , los intervalos de confianza como R^2 adj, y la interpretación de los coeficientes. Además, se considerará la multicolinealidad entre variables independientes y se examinarán los residuos en busca de patrones o valores atípicos. El informe concluirá con un resumen de hallazgos, la determinación de la adecuación del modelo para predicciones, limitaciones del análisis, y recomendaciones para futuras investigaciones. Se incluirán los detalles analíticos para resolver el reto mediante el software R Studio.

Palabras clave: R studio; Regresión Lineal Múltiple; Regresión Logística; precio; Toyota.

Introducción

En el presente informe, se llevó a cabo un Análisis de Exploración de Datos (EDA) sobre un conjunto de datos relacionados con el precio de autos usados de la marca Toyota. El objetivo principal fue comprender la estructura y características de los datos antes de realizar suposiciones, identificar errores, entender patrones y determinar relaciones entre variables numéricas y categóricas. Se trabajó con un dataset que contenía 37 atributos y 1,436 instancias; se siguió la metodología CRISP-DM, haciendo énfasis en la exploración de datos mediante EDA. De este modo, se identificaron variables clave, como el kilometraje, el año de manufactura, la edad del auto y su peso, que mostraron relaciones significativas con el precio de venta. Además, se detectaron outliers, se realizó una búsqueda de datos faltantes; y finalmente, se concluyó que no era necesario realizar imputación de datos adicionales, pero se destacó la necesidad de emplear estrategias de modelación, incluyendo Regresión Logística y Lineal, para identificar los factores que tienen mayor influencia en la determinación del precio de venta de un auto Toyota.

En el presente informe argumentativo, se expondrá un modelo de Regresión Lineal con el propósito de realizar análisis de relación entre las variables predictoras y la variable de salida, que en este caso es la variable 'Precio'. En este contexto, nos centraremos en evaluar el rendimiento del modelo en un conjunto de validación mediante el uso de métricas de análisis estadístico como el R^2 ajustado para determinar la veracidad del modelo.

Además, se abordará el desafío de clasificar el precio de los autos mediante la aplicación de la técnica de Regresión Logística. Este enfoque estadístico permitirá modelar la relación entre las variables predictoras categóricas y la variable de respuesta (precio), estimando los coeficientes del modelo. El proceso comprenderá la obtención del conjunto de datos, el entrenamiento del clasificador logístico y la realización de predicciones basadas en dicho modelo para un nuevo conjunto de datos. Todo el desarrollo de los algoritmos y análisis se realizará utilizando el entorno de programación R.

Como se mencionó anteriormente, se utilizarán dos modelos estadísticos: Regresión Lineal Múltiple y Regresión Logística, con el fin de responder a nuestra pregunta de investigación: “¿Qué factores determinan el precio de venta de los vehículos marca Toyota usados en México?”.

Exploración de Datos

En esta sección, se llevará a cabo una descripción integral del conjunto de datos. Este análisis abarcará la presentación detallada de las variables incluidas en el conjunto de datos, destacando su naturaleza como variables dependientes o independientes. Se proporcionará información esencial sobre los tipos de variables, ya sean numéricas o categóricas, y se detallarán sus escalas de medición.

i. Importación de Librerías y Dataset

Inicialmente se instalan los siguientes paquetes y se procede con cargar la base de datos.

```
library("ggfortify"), library("statsr"), library("skimr"), library("forecast"),  
library("ggplot2"), library("dplyr"), library("broom"), library("ggpubr"),  
library("gvlma"), library("readxl"), library("caret"), library("tidyverse"),  
library("pillar"), library("psych"), library("readr"), library("GGally"), lib  
rary("corrplot"), library("reshape2"), library("gmodels"), library("mice")  
  
corolla_complete=read.csv("corolla.csv")  
  
View(corolla_complete)
```

ii. Visualización de los Atributos y sus Datos

Iniciaremos con la exploración de datos inicial, la cual permitirá revisar los tipos atributos, la existencia o no de missing data, los percentiles, la distribución de los datos; entre otros.

La función “dim” proporciona las dimensiones del dataframe, es decir, el número de instancias y atributos. Este es un paso básico pero crucial para entender la dimensión del conjunto de datos. De la cual se obtiene que estamos trabajando con una estructura de 1,436 instancias y 37 atributos.

```
dim(corolla_complete)
```

```
## [1] 1436 37
```

Se presenta una descripción de los 37 atributos, de los cuales en este caso tomaremos como nuestra variable dependiente a “Price” y las demás serán consideradas las variables independientes:

- Id: Identificador único para cada registro.
- Model: Modelo del automóvil.
- Price: Precio de venta del automóvil.
- Age_08_04: Edad del automóvil en meses.
- Mfg_Month: Mes de fabricación.
- Mfg_Year: Año de fabricación.
- KM: Kilometraje del automóvil.
- Fuel_Type: Tipo de combustible utilizado.
- HP: Caballos de fuerza del motor.
- Met_Color: Indicador de si el automóvil tiene pintura metalizada.
- Automatic: Indicador de si el automóvil es automático.
- cc: Cilindrada del motor.
- Doors: Número de puertas del automóvil.
- Cylinders: Número de cilindros en el motor.
- Gears: Número de marchas.
- Quarterly_Tax: Impuesto trimestral.
- Weight: Peso del automóvil.
- Mfr_Guarantee: Indicador de garantía del fabricante.

- BOVAG_Guarantee: Indicador de garantía BOVAG.
- Guarantee_Period: Período de garantía. ABS: Sistema de frenos antibloqueo. Airbag_1: Airbag frontal. Airbag_2: Airbag lateral.
- Airco: Aire acondicionado.
- Automatic_airco: Aire acondicionado automático.
- Boardcomputer: Computadora de a bordo.
- CD_Player: Reproductor de CD.
- Central_Lock: Cierre centralizado.
- Powered_Windows: Ventanas eléctricas.
- Power_Steering: Dirección asistida.
- Radio: Radio incorporado.
- Mistlamps: Luces antiniebla.
- Sport_Model: Modelo deportivo.
- Backseat_Divider: Separador de asiento trasero.
- Metallic_Rim: Rines metálicos.
- Radio_cassette: Radio casete.
- Tow_Bar: Enganche de remolque.

Ahora, la función ‘skim’ proporciona rápidamente una visión general del marco de los datos. En este caso, nos muestra que el número de valores que faltan es cero, así como las características estadísticas básicas.

```
skim(corolla_complete)
```







Data summary


















Name	corolla_complete
Number of rows	1436
Number of columns	37
<hr/>	
Column type frequency:	
character	2
numeric	35
<hr/>	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Model	0	1	15	75	0	372	0
Fuel_Type	0	1	3	6	0	3	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Id	0	1	721.56	416.48	1	361.75	721.5	1081.25	1442	
Price	0	1	10730.82	3626.96	4350	8450.00	9900.0	11950.00	32500	
Age_08_04	0	1	55.95	18.60	1	44.00	61.0	70.00	80	
Mfg_Month	0	1	5.55	3.35	1	3.00	5.0	8.00	12	
Mfg_Year	0	1	1999.63	1.54	199898	1998.00	1999.0	2001.00	2004	
KM	0	1	68533.26	37506.45	1	43000.00	63389.5	87020.75	243000	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
HP	0	1	101.50	14.98	69	90.00	110.0	110.00	192	
Met_Color	0	1	0.67	0.47	0	0.00	1.0	1.00	1	
Automatic	0	1	0.06	0.23	0	0.00	0.0	0.00	1	
cc	0	1	1576.86	424.39	1300	1400.00	1600.0	1600.00	16000	
Doors	0	1	4.03	0.95	2	3.00	4.0	5.00	5	
Cylinders	0	1	4.00	0.00	4	4.00	4.0	4.00	4	
Gears	0	1	5.03	0.19	3	5.00	5.0	5.00	6	
Quarterly_Tax	0	1	87.12	41.13	19	69.00	85.0	85.00	283	
Weight	0	1	1072.46	52.64	1000	1040.00	1070.0	1085.00	1615	
Mfr_Guarantee	0	1	0.41	0.49	0	0.00	0.0	1.00	1	
BOVAG_Guarantee	0	1	0.90	0.31	0	1.00	1.0	1.00	1	
Guarantee_Period	0	1	3.82	3.01	3	3.00	3.0	3.00	36	
ABS	0	1	0.81	0.39	0	1.00	1.0	1.00	1	
Airbag_1	0	1	0.97	0.17	0	1.00	1.0	1.00	1	
Airbag_2	0	1	0.72	0.45	0	0.00	1.0	1.00	1	
Airco	0	1	0.51	0.50	0	0.00	1.0	1.00	1	
Automatic_airco	0	1	0.06	0.23	0	0.00	0.0	0.00	1	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Boardcomputer	0	1	0.29	0.46	0	0.00	0.0	1.00	1	█
CD_Player	0	1	0.22	0.41	0	0.00	0.0	0.00	1	█
Central_Lock	0	1	0.58	0.49	0	0.00	1.0	1.00	1	█
Powered_Windows	0	1	0.56	0.50	0	0.00	1.0	1.00	1	█
Power_Steering	0	1	0.98	0.15	0	1.00	1.0	1.00	1	█
Radio	0	1	0.15	0.35	0	0.00	0.0	0.00	1	█
Mistlamps	0	1	0.26	0.44	0	0.00	0.0	1.00	1	█
Sport_Model	0	1	0.30	0.46	0	0.00	0.0	1.00	1	█
Backseat_Divider	0	1	0.77	0.42	0	1.00	1.0	1.00	1	█
Metallic_Rim	0	1	0.20	0.40	0	0.00	0.0	0.00	1	█
Radio_cassette	0	1	0.15	0.35	0	0.00	0.0	0.00	1	█
Tow_Bar	0	1	0.28	0.45	0	0.00	0.0	1.00	1	█

A partir de analizar los resultados arrojados por la función ‘skim’, se puede observar que hay 2 atributos de caracteres: ‘Model’ y ‘Fuel_Type’, y 35 atributos numéricos. Sin embargo, de los 35 atributos numéricos, 21 son considerados lógicos ya que definen numéricamente 2 clases de respuesta. Además, con el uso de la función ‘sapply’, se puede observar que los atributos no cuentan con missing values.

```
sapply(corolla_complete, function(x) sum(is.na(x)))
```

##	Id	Model	Price	Age_08_04
##	0	0	0	0
##	Mfg_Month	Mfg_Year	KM	Fuel_Type
##	0	0	0	0
##	HP	Met_Color	Automatic	cc
##	0	0	0	0
##	Doors	Cylinders	Gears	Quarterly_Tax
##	0	0	0	0
##	Weight	Mfr_Guarantee	BOVAG_Guarantee	Guarantee_Period
##	0	0	0	0
##	ABS	Airbag_1	Airbag_2	Airco
##	0	0	0	0
##	Automatic_airco	Boardcomputer	CD_Player	Central_Lock
##	0	0	0	0
##	Powered_Windows	Power_Steering	Radio	Mistlamps
##	0	0	0	0
##	Sport_Model	Backseat_Divider	Metallic_Rim	Radio_cassette
##	0	0	0	0
##	Tow_Bar			
##	0			

Se puede observar de manera clara que no existen missing data entre los atributos al reflejarse un número 0 para cada una de las variables. Esto fue analizado a detalle en el reporte anterior, donde también se realizó un análisis de outliers, en el que se encontró la oportunidad de corregir un dato en el atributo 'cc', que reflejaba un registro fuera de proporción en comparación al resto. Para ello se reemplazó el valor correcto '1600' por el valor atípico que se encuentra en la instancia 81 del atributo 'cc'.

```
corolla_complete$cc[81]=1600
```

Ya que no se encontró ningún valor identificado como 'missing data (NA)', no se encontró evidencia para la realización de procedimientos relacionados a la imputación de datos.

Evaluación de Colinealidad

Para responder a la pregunta de cuáles factores influyen en "Price" pregunta, examinaremos las correlaciones entre la variable dependiente de interés con respecto a las demás variables explicativas.

Para ello, se llevará a cabo una evaluación de la multicolinealidad entre las variables independientes del conjunto de datos. Este análisis tiene como objetivo asegurarse de que las variables no estén altamente correlacionadas entre sí, lo que podría afectar la precisión de los modelos de Regresión Lineal y Logística.

Una representación de cómo los datos interactúan entre sí es mediante la matriz de correlación. La cual permitirá evaluar las relaciones lineales, analizar la multicolinealidad y seleccionar los predictores para la variable de salida.

Visualmente esto es, entre mayor el diámetro del círculo mostrado en la gráfica, mayor es la correlación entre las variables. O bien, en la tabla de correlación: 1 indica una correlación positiva perfecta, -1 indica una correlación negativa perfecta, 0 indica falta de correlación lineal.

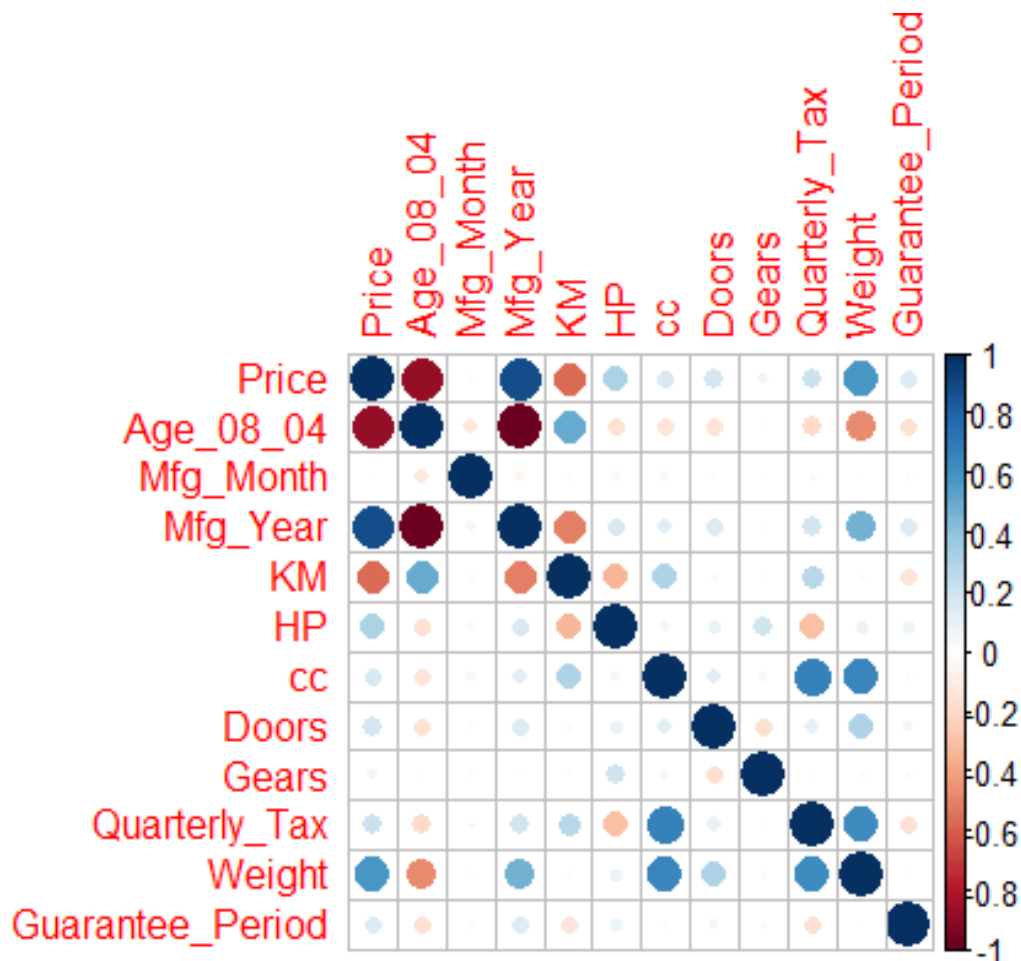
Se han seleccionado columnas específicas que corresponden a los atributos numéricos que permitirán hacer un contraste cuantitativo mediante representaciones gráficas. Por el momento no se trabajarán los datos de carácter lógico, no obstante, cabe destacar que serán analizados en

instancias posteriores del proyecto en cuestión ya que permiten representar y trabajar con información que tiene naturaleza binaria.

```
corolla.df = corolla_complete %>%
  select("Id","Model","Price","Age_08_04","Mfg_Month", "Mfg_Year","KM","HP","
cc","Doors","Cylinders", "Gears","Quarterly_Tax", "Weight","Guarantee_Period"
)

#no se toma en cuenta los atributos 1, 2 ni 11: Id, Model y Cylinders
cor.matrix=cor(na.omit(corolla.df[, -c(1,2,11)]))

#crear un gráfico visual de la matriz de correlación
corrplot(cor.matrix)
```



En el análisis EDA realizado en el informe anterior se presenta una descripción más detallada acerca de la colinealidad, a continuación, se detallarán únicamente aquellas escogidas para el modelo MLR.

Como resultados se tiene lo siguiente:

Correlación positiva:

- Mfg_Year: 0.89
- Weight: 0.58
- Gears: 0.06
- Guarantee_Period: 0.15
- Doors: 0.19
- cc: 0.17
- HP: 0.31
- Quarterly_Tax: 0.22
- Mfg_Month: -0.02

Correlación negativa:

- Age_08_04: -0.88
- KM: -0.57
- Automatic_airco: -0.44
- Automatic: -0.44
- Airco: -0.43
- Mistlamps: -0.23
- CD_Player: -0.13
- Sport_Model: -0.14
- Airbag_2: -0.09
- Airbag_1: -0.15

- Metallic_Rim: -0.01
- Radio: 0.11
- Boardcomputer: -0.05
- Central_Lock: -0.15
- Powered_Windows: -0.16
- Radio_cassette: -0.01
- Tow_Bar: -0.07

Posible Multicolinealidad Positiva:

- “Mfg_Year” y “Age_08_04” tienen una correlación de 0.89. “Weight” y “cc” tienen una correlación de 0.65.

Posible Multicolinealidad Negativa:

- “Automatic” y “Automatic_airco” tienen una correlación de -0.44.

Construcción del Modelo de Regresión Lineal Múltiple

La Regresión lineal es un método estadístico que se utiliza para modelar la relación lineal entre una variable dependiente (respuesta) y una o más variables independientes (predictoras).

Una vez analizado qué miden los distintos predictores y por qué son relevantes para predecir la variable de salida, procederemos con la propuesta de una ecuación de Regresión Lineal Múltiple para predecir el valor de “Price”. Además, se explicará la razón de cualquier transformación o interacción aplicada a las variables.

Para esto, definimos la hipótesis de Regresión Lineal Múltiple para la variable dependiente Y=Price de la siguiente manera:

$$\text{Price} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$$

donde β_0, \dots, β_p son los coeficientes, las X_p son las variables independientes a seleccionar y ϵ es el ruido o parte no explicada.

Se generaron 3 modelos de Regresión Lineal, cada uno con variables diferentes que se definían de acuerdo con la significancia de las variables del modelo anterior.

i. 'Price' en función de todas las variables predictoras

Para proceder con el modelo de Regresión Lineal se hace uso de la función “lm()” que genera el modelo para determinar la relación y significancia de todas las variables predictoras con la variable de salida “Price”.

Dado que nuestro subset de datos cuenta con 12 atributos: “Age_08_04”, “Mfg_Month”, “Mfg_Year”, “KM”, “HP”, “cc”, “Doors”, “Cylinders”, “Gears”, “Quarterly_Tax”, “Weight”, “Guarantee_Period” definiremos la hipótesis de Regresión Lineal de la siguiente manera:

$$\text{Price} = B_0 + (B_1 \cdot \text{Age_08_04}) + (B_2 \cdot \text{Mfg_Month}) + (B_3 \cdot \text{Mfg_Year}) + (B_4 \cdot \text{KM}) + (B_5 \cdot \text{HP}) + (B_6 \cdot \text{cc}) + (B_7 \cdot \text{Doors}) + (B_8 \cdot \text{Cylinders}) + (B_9 \cdot \text{Gears}) + (B_{10} \cdot \text{Quarterly_Tax}) + (B_{11} \cdot \text{Weight}) + (B_{12} \cdot \text{Guarantee_Period})$$

```
modelo_lineal_1 <- lm(Price ~ Age_08_04 + Mfg_Month + Mfg_Year + KM + HP + cc  
+ Doors + Cylinders + Gears + Quarterly_Tax + Weight + Guarantee_Period, data  
= corolla.df)
```

```
# Resumen del modelo  
options(scipen = 999)  
summary(modelo_lineal_1)
```

```
##
## Call:
## lm(formula = Price ~ Age_08_04 + Mfg_Month + Mfg_Year + KM +
##      HP + cc + Doors + Cylinders + Gears + Quarterly_Tax + Weight +
##      Guarantee_Period, data = corolla.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11447.5   -705.3    -36.8    737.5   7091.2
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4931.086101  1347.318304  -3.660  0.000261 **
*
## Age_08_04    -122.863852    2.529345 -48.575 < 0.0000000000000002 **
*
## Mfg_Month    -98.694033    10.140797  -9.732 < 0.0000000000000002 **
*
## Mfg_Year      NA            NA        NA      NA
## KM            -0.017523     0.001233 -14.216 < 0.0000000000000002 **
*
## HP           38.621730     2.817309  13.709 < 0.0000000000000002 **
*
## cc           -2.426902     0.300134  -8.086  0.0000000000000013 **
*
## Doors        -34.055145    37.894366  -0.899    0.368972
## Cylinders      NA            NA        NA      NA
## Gears         528.412717   186.058364   2.840    0.004575 **
## Quarterly_Tax  9.912092    1.404924   7.055  0.00000000000026801 **
*
## Weight        19.226265     1.080922  17.787 < 0.0000000000000002 **
*
## Guarantee_Period 43.260581   11.594481   3.731    0.000198 **
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1266 on 1425 degrees of freedom
## Multiple R-squared:  0.879, Adjusted R-squared:  0.8781
## F-statistic: 1035 on 10 and 1425 DF, p-value: < 0.00000000000000022
```

A partir de este análisis donde se toman en cuenta casi la totalidad de las variables numéricas se obtiene que el modelo de Regresión Lineal resulta de la siguiente manera:

$$\begin{aligned} \text{Price} = & -4931.086101 - (122.863852 \cdot \text{Age_08_04}) - (98.694033 \cdot \text{Mfg_Month}) - (0.017523 \cdot \text{KM}) + \\ & (38.621730 \cdot \text{HP}) - (2.426902 \cdot \text{cc}) - (34.055145 \cdot \text{Doors}) + (528.412717 \cdot \text{Gears}) + (9.912092 \cdot \\ & \text{Quarterly_Tax}) + (19.226265 \cdot \text{Weight}) + (43.260581 \cdot \text{Guarantee_Period}) \end{aligned}$$

En este modelo se ignoran los atributos 'Mfg_Year' y 'Cylinders' ya que se deben a singularidades, lo que podría indicar colinealidad o algún otro problema en el modelo. De entrada, se sabe, de acuerdo a la matriz de correlación que 'Mfg_Year' y 'Age_08_04' muestran un nivel significativo de multicolinealidad, por lo tanto se ignora la variable. Por otro lado, el atributo 'Cylinders' es una variable con registros constantes, por lo tanto, no se considera una variable significativa para el modelo.

Analizando las respuestas estadísticas que arroja la función 'summary()', es evidente que los atributos en su mayoría son significantes para determinar la variable respuesta 'Price' (presentan 3 estrellas del lado derecho, que reflejan un nivel de significancia del 1%), menos la variable doors que muestra un código de significancia 0. El coeficiente de determinación (R-squared) es 0.879, lo que sugiere que el modelo explica aproximadamente el 87.9% de la variabilidad en la variable dependiente.

Para el siguiente modelo se ignorarán los atributos 'Doors', y 'Cylinders' esperando una mejoría. Además, se remplazaron los atributos 'Age_08_04' y 'Mfg_Month' con 'Mfg_Year' ya que este último presenta una relación directamente proporcional con la variable de salida y se desea evitar cualquier oportunidad de multicolinealidad entre variables.

ii. 'Price' en función de las variables predictoras 'Mfg_Year', 'KM', 'HP', 'cc', 'Gears', 'Quarterly_Tax', 'Weight', y 'Guarantee_Period'.

En un segundo acercamiento a encontrar nuestro mejor MLR, se seleccionaron las variables: 'Mfg_Year', 'KM', 'HP', 'cc', 'Gears', 'Quarterly_Tax', 'Weight', y 'Guarantee_Period'. Estos atributos fueron los que aparecieron tener un mayor nivel de significancia con la variable de salida 'Price'.

En esta ocasión se define la segunda hipótesis del modelo de Regresión Lineal de la siguiente manera, basándonos en las correlaciones observadas anteriormente, ya que pueden proporcionar información valiosa para predecir la variable 'Price' en un modelo de Regresión Lineal múltiple:

$$\text{Price} = B_0 + (B_1 \cdot \text{Mfg_Year}) + (B_2 \cdot \text{KM}) + (B_3 \cdot \text{HP}) + (B_4 \cdot \text{cc}) + (B_5 \cdot \text{Gears}) + (B_6 \cdot \text{Quarterly_Tax}) + (B_7 \cdot \text{Weight}) + (B_8 \cdot \text{Guarantee_Period})$$

```
modelo_lineal_2 <- lm(Price ~ Mfg_Year + KM + HP + cc + Gears + Quarterly_Tax
+ Weight + Guarantee_Period, data = corolla.df)

# Resumen del modelo
options(scipen = 999)
summary(modelo_lineal_2)
```

```
##
## Call:
## lm(formula = Price ~ Mfg_Year + KM + HP + cc + Gears + Quarterly_Tax +
##      Weight + Guarantee_Period, data = corolla.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11422.8   -709.4    -25.5    742.6   7200.1
##
## Coefficients:
##              Estimate      Std. Error t value      Pr(>|
t|)
## (Intercept)   -2946997.025393    60145.224511 -48.998 < 0.000000000000000
002
## Mfg_Year       1467.586177       30.259028  48.501 < 0.000000000000000
002
## KM            -0.017893        0.001226 -14.594 < 0.000000000000000
002
## HP            38.032657        2.805268  13.558 < 0.000000000000000
002
## cc           -2.356898        0.298855  -7.886 0.000000000000000
614
## Gears         560.929528       182.901994   3.067      0.002
204
## Quarterly_Tax   9.991715        1.406681   7.103 0.000000000000191
958
## Weight        18.934784        1.044591  18.127 < 0.000000000000000
002
## Guarantee_Period 43.846495       11.590675   3.783      0.000
161
##
## (Intercept)    ***
## Mfg_Year       ***
## KM             ***
## HP             ***
## cc             ***
## Gears          **
## Quarterly_Tax  ***
## Weight         ***
## Guarantee_Period ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1268 on 1427 degrees of freedom
## Multiple R-squared:  0.8784, Adjusted R-squared:  0.8778
## F-statistic: 1289 on 8 and 1427 DF, p-value: < 0.00000000000000022
```

A partir de este análisis se obtiene que el modelo de regresión lineal resulta de la siguiente manera:

$$\begin{aligned} \text{Price} = & -2946997.025393 + (1467.586177 \cdot \text{Mfg_Year}) - (0.017893 \cdot \text{KM}) + (38.032657 \cdot \text{HP}) - \\ & (2.356898 \cdot \text{cc}) + (560.929528 \cdot \text{Gears}) + (9.991715 \cdot \text{Quarterly_Tax}) + (18.934784 \cdot \text{Weight}) + \\ & (43.846495 \cdot \text{Guarantee_Period}) \end{aligned}$$

Analizando las respuestas estadísticas que arroja la función ‘summary()’, es evidente que los atributos en su mayoría son significantes para determinar la variable respuesta ‘Price’ (presentan 3 estrellas del lado derecho, que reflejan un nivel de significancia del 1%), menos la variable ‘Gears’ que muestra un código de significancia un poco menor. El coeficiente de determinación (R-squared) es 0.87, lo que sugiere que el modelo explica aproximadamente el 87.8% de la variabilidad en la variable dependiente.

El R-cuadrado de ambos modelos es sumamente similar, sin embargo, en este segundo modelo únicamente se tomaron en cuenta atributos que presentan una relación significativa con la variable de salida, por lo tanto, este modelo será empleado en el modelo de step forward para verificar la existencia de un mejor modelo.

iii. ‘Price’ bajo Modelo con ‘Step Forward’

En un tercer acercamiento a encontrar el mejor MLR, se empleó el método de ‘step forward’ con el fin de determinar y seleccionar los atributos con más significancia para el modelo final. El “step forward” es una técnica utilizada en la construcción de modelos de Regresión Lineal para la selección de características (features) o variables predictoras que se incluirán en el modelo final.

```
corolla.lm.null <- lm(Price ~ 1, data = corolla.df)
corolla.lm.step <- step(corolla.lm.null, scope = list(lower = corolla.lm.null
, upper = modelo_lineal_2), direction = "forward")
```

```
## Start: AIC=23540.35
## Price ~ 1
##
##              Df    Sum of Sq      RSS    AIC
## + Mfg_Year      1 14790447912  4086793552 21345
## + Weight        1  6376555420 12500686044 22951
## + KM            1  6132358520 12744882944 22978
## + HP           1  1872973334 17004268130 23392
## + Quarterly_Tax 1   907000213 17970241250 23472
## + cc           1   514350179 18362891284 23503
## + Guarantee_Period 1  405848686 18471392778 23511
## + Gears         1    75171003 18802070461 23537
## <none>                      18877241464 23540
##
## Step: AIC=21344.98
## Price ~ Mfg_Year
##
##              Df    Sum of Sq      RSS    AIC
## + Weight        1  639312864 3447480688 21103
## + HP           1  555542785 3531250767 21137
## + KM           1  383205928 3703587624 21206
## + Gears        1   59689200 4027104352 21326
## + cc           1   55496241 4031297310 21327
## + Quarterly_Tax 1   44320621 4042472931 21331
## <none>                      4086793552 21345
## + Guarantee_Period 1  4595154 4082198398 21345
##
## Step: AIC=21102.7
## Price ~ Mfg_Year + Weight
##
##              Df    Sum of Sq      RSS    AIC
## + KM           1  764943047 2682537641 20744
## + HP           1  539793383 2907687305 20860
## + cc           1  173101698 3274378990 21031
## + Quarterly_Tax 1 130490446 3316990242 21049
## + Gears        1   52432470 3395048218 21083
## + Guarantee_Period 1  20943338 3426537349 21096
## <none>                      3447480688 21103
##
## Step: AIC=20744.43
## Price ~ Mfg_Year + Weight + KM
##
##              Df    Sum of Sq      RSS    AIC
## + HP           1  237797141 2444740500 20613
## + Gears        1   59610880 2622926761 20714
## + cc           1   16208903 2666328738 20738
## + Guarantee_Period 1  10068037 2672469604 20741
## <none>                      2682537641 20744
```

```
## + Quarterly_Tax      1   2340315 2680197326 20745
##
## Step:  AIC=20613.14
## Price ~ Mfg_Year + Weight + KM + HP
##
##              Df Sum of Sq      RSS   AIC
## + cc          1  41205824 2403534676 20591
## + Quarterly_Tax  1  21156005 2423584495 20603
## + Gears         1  18716505 2426023995 20604
## + Guarantee_Period 1   6476066 2438264435 20611
## <none>                      2444740500 20613
##
## Step:  AIC=20590.73
## Price ~ Mfg_Year + Weight + KM + HP + cc
##
##              Df Sum of Sq      RSS   AIC
## + Quarterly_Tax  1  71519582 2332015094 20549
## + Gears          1  18634818 2384899858 20582
## + Guarantee_Period 1   7902738 2395631938 20588
## <none>                      2403534676 20591
##
## Step:  AIC=20549.35
## Price ~ Mfg_Year + Weight + KM + HP + cc + Quarterly_Tax
##
##              Df Sum of Sq      RSS   AIC
## + Guarantee_Period 1  22044624 2309970470 20538
## + Gears            1  14156653 2317858441 20543
## <none>                      2332015094 20549
##
## Step:  AIC=20537.71
## Price ~ Mfg_Year + Weight + KM + HP + cc + Quarterly_Tax + Guarantee_Perio
d
##
##              Df Sum of Sq      RSS   AIC
## + Gears      1  15125478 2294844992 20530
## <none>                2309970470 20538
##
## Step:  AIC=20530.28
## Price ~ Mfg_Year + Weight + KM + HP + cc + Quarterly_Tax + Guarantee_Perio
d +
##      Gears
```

```
summary(corolla.lm.step)
```

```
##
## Call:
## lm(formula = Price ~ Mfg_Year + Weight + KM + HP + cc + Quarterly_Tax +
##     Guarantee_Period + Gears, data = corolla.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11422.8   -709.4    -25.5    742.6   7200.1
##
## Coefficients:
##              Estimate      Std. Error t value      Pr(>|
t|)
## (Intercept)   -2946997.025393    60145.224511 -48.998 < 0.000000000000000
002
## Mfg_Year       1467.586177       30.259028  48.501 < 0.000000000000000
002
## Weight        18.934784        1.044591  18.127 < 0.000000000000000
002
## KM            -0.017893        0.001226 -14.594 < 0.000000000000000
002
## HP            38.032657        2.805268  13.558 < 0.000000000000000
002
## cc            -2.356898        0.298855  -7.886  0.000000000000000
614
## Quarterly_Tax  9.991715        1.406681   7.103  0.0000000000000191
958
## Guarantee_Period 43.846495     11.590675   3.783    0.000
161
## Gears         560.929528     182.901994   3.067    0.002
204
##
## (Intercept)    ***
## Mfg_Year       ***
## Weight         ***
## KM             ***
## HP             ***
## cc             ***
## Quarterly_Tax  ***
## Guarantee_Period ***
## Gears          **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1268 on 1427 degrees of freedom
## Multiple R-squared:  0.8784, Adjusted R-squared:  0.8778
## F-statistic: 1289 on 8 and 1427 DF, p-value: < 0.00000000000000022
```

Este tercer modelo que implica el uso de ‘step forward’ proporcionó un modelo idéntico al segundo modelo previamente planteado. Esto es porque el step forward comprobó que no existe posibilidad de mejoría al modelo; por lo tanto, el modelo presentado anteriormente (‘modelo_lineal_2’), es el óptimo, con mejor ajuste a los datos.

Analizando las respuestas estadísticas que arroja la función ‘summary()’, es evidente que los atributos en su mayoría son significantes para determinar la variable respuesta ‘Price’ (presentan 3 estrellas del lado derecho, que reflejan un nivel de significancia del 1%), menos la variable ‘Gears’ que muestra un código de significancia un poco menor. El coeficiente de determinación (R-squared) es 0.87, lo que sugiere que el modelo explica aproximadamente el 87.8% de la variabilidad en la variable dependiente, lo cual es un buen resultado del modelo.

iv. Seleccionar el Modelo Lineal

El R cuadrado ajustado proporciona una medida de la calidad del modelo ajustado teniendo en cuenta el número de variables en el modelo. Un valor más alto indica un mejor ajuste, razón por la cual escogemos como mejor MLR nuestra opción (ii), ya que brinda un valor de R cuadrado ajustado igual a 0.87 y además, significancia en todas las variables independientes.

```
# Obtener el R cuadrado ajustado
r_cuadrado_ajustado_1 <- summary(modelo_lineal_1)$adj.r.squared
r_cuadrado_ajustado_2 <- summary(modelo_lineal_2)$adj.r.squared
r_cuadrado_ajustado_3 <- summary(corolla.lm.step)$adj.r.squared

# Imprimir el resultado
cat("El R cuadrado ajustado para el modelo (i) es:", r_cuadrado_ajustado_1, "\n\n")

## El R cuadrado ajustado para el modelo (i) es: 0.8781432

cat("El R cuadrado ajustado para el modelo (ii) es:", r_cuadrado_ajustado_2, "\n\n")

## El R cuadrado ajustado para el modelo (ii) es: 0.8777517

cat("El R cuadrado ajustado para el modelo (iii) es:", r_cuadrado_ajustado_3, "\n\n")

## El R cuadrado ajustado para el modelo (iii) es: 0.8777517
```

El coeficiente de determinación (R-squared) es 0.87 en los modelos sugiere que el modelo explica aproximadamente el 87% de la variabilidad en la variable dependiente, lo cual es muy bueno. Aunque el coeficiente de determinación sea unos decimales más alto en el primer modelo (i), se opta por el segundo modelo (ii) ya que todas las variables independientes tienen una relación significativa con la variable de salida en el modelo.

Training y Testing del MLR Seleccionado

i. Training

Para comprobar la relación lineal entre las variables, se establecerá un escenario de entrenamiento y validación para evaluar un modelo lineal. El conjunto de entrenamiento se crea seleccionando aleatoriamente el 60% de los índices (utilizando la función “set.seed”) y el conjunto de validación se forma excluyendo las filas correspondientes al conjunto de entrenamiento, siendo el 40% restante del subset “corolla.df”.

Primeramente se asignaron 860 datos, es decir el 60% de los datos al entrenamiento (“train.index”). Posteriormente, se seleccionaron las variables que presentaron significancia en el modelo elegido (‘modelo_lineal_2’), y utilizando estos dos supuestos, se generó el dataframe de entrenamiento (“training.df”) y el dataset de validación o prueba (“validing.df”).

```
selected_variable <- c(3,6, 7, 9, 12, 15, 16, 17, 20)

names(corolla_complete[selected_variable])

## [1] "Price"           "Mfg_Year"        "KM"              "HP"
## [5] "cc"              "Gears"           "Quarterly_Tax"   "Weight"
## [9] "Guarantee_Period"
```



```
# Establecer semilla para reproducibilidad
set.seed(123)

# Obtener el número total de instancias (1426)
total_instancias <- nrow(corolla.df)

# Definir el tamaño del conjunto de entrenamiento (60%)
tamano_entrenamiento <- round(0.6 * total_instancias)

# Crear índices aleatorios para el conjunto de entrenamiento
train.index <- sample(1:total_instancias, size = tamano_entrenamiento, replace = FALSE)

# Crear conjunto de entrenamiento
training.df <- corolla_complete[train.index, selected_variable]

# Crear conjunto de validación
validing.df <- corolla_complete[-train.index, selected_variable]
```

Ahora se procederá a emplear MLR mediante la función “lm()” para el conjunto de datos “training.df” con el fin de entrenar el modelo. Es de esta forma que podemos comprobar que en efecto existe relaciones lineales entre las variables independientes y la variable dependiente.

```
corolla.lm <- lm(Price~., data = training.df)

options(scipen = 999)
summary(corolla.lm)
```

```
##
## Call:
## lm(formula = Price ~ ., data = training.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10533.8   -677.7    -25.0     720.6    7323.9
##
## Coefficients:
##              Estimate      Std. Error t value      Pr(>|
t|)
## (Intercept)   -3024978.325233    80786.095663  -37.444 < 0.00000000000000
002
## Mfg_Year       1507.722648      40.644627   37.095 < 0.00000000000000
002
## KM             -0.017650       0.001636  -10.788 < 0.00000000000000
002
## HP             39.467270       3.805911   10.370 < 0.00000000000000
002
## cc            -2.475342       0.394974   -6.267    0.0000000000
583
## Gears          490.700743      238.566177    2.057    0.04
000
## Quarterly_Tax  10.935451       1.900825    5.753    0.0000000012
216
## Weight        17.077551       1.395565   12.237 < 0.00000000000000
002
## Guarantee_Period 40.402103     13.785237    2.931    0.00
347
##
## (Intercept)    ***
## Mfg_Year       ***
## KM             ***
## HP             ***
## cc             ***
## Gears          *
## Quarterly_Tax  ***
## Weight         ***
## Guarantee_Period **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1292 on 853 degrees of freedom
## Multiple R-squared:  0.8715, Adjusted R-squared:  0.8703
## F-statistic: 723.4 on 8 and 853 DF, p-value: < 0.000000000000000022
```

Los resultados son un poco diferentes al modelo lineal 2 ya que solo se ajusta al 60% de los datos de entrenamiento, lo cual puede alterar el resultado final. Sin embargo, los resultados fueron similares, por lo tanto, este modelo se usará para hacer las predicciones y obtener los residuales.

ii. Testing

Ahora se realizará la validación del modelo mediante la función “predict()”. La variable “corolla.lm.pred” almacenará las predicciones que se obtienen del modelo de Regresión Lineal; por lo tanto a función “predict()” utilizará el conjunto de datos “validing.df” para generar las predicciones basadas en el modelo “corolla.lm”.

```
# Realizar predicciones en el conjunto de validación
corolla.lm.pred <- predict(corolla.lm, validing.df)
```

Por su parte, los residuales representan la diferencia entre los valores observados y los valores predichos por el modelo. Se crea un data frame que contiene tres atributos: “Predicted” con las predicciones, “Actual” con los valores reales y “Residual” con los residuales; calculados en base al 40% de los datos totales restantes (subset “validing.df”).

```
options(scipen = 999)

some_residuais <- validing.df$Price - corolla.lm.pred

data_frame("Predicted" = corolla.lm.pred, "Actual" = validing.df$Price, "Residual" = some_residuais)
```

	Predicted	Actual	Residual
	<dbl>	<int>	<dbl>
1	16021.	13500	-2521.
2	16114.	13950	-2164.
3	16003.	14950	-1053.
4	16547.	16900	353.
5	20231.	19950	-281.
6	20594.	21500	906.
7	20398.	22500	2102.
8	15044.	15950	906.
9	16737.	16950	213.

```
## 10    15885.  17495    1610.
## # i 564 more rows
```

Previo al análisis de residuales, es importante destacar que estos se miden en unidades de la variable de salida, es decir, en este caso se miden en unidades de dinero; específicamente en dólares americanos. Al momento de analizar los residuales, es importante destacar que entre más cercanos sean los valores a 0, mejor. En este caso se obtienen valores alejados de 0; pero poniéndolos en perspectiva numérica, muestra que los resultados no están tan alejados de la situación real. En algunos casos la diferencia entre el valor actual y el predicho es de 2650 dólares, y en otros casos es de 232 dólares. Por lo cual se puede concluir que los residuales son aceptables y se pueden explicar con el hecho de que el 87% (R-cuadrado) de los datos pueden ser explicados por el modelo.

Verificación de Supuestos y Análisis de Residuos

En esta sección se evaluará si en efecto se cumplen los supuestos de la Regresión Lineal Múltiple para nuestro modelo seleccionado (ii). Estos son:

- **Linealidad:** Relaciones lineales entre las variables independientes y la variable dependiente.
- **Independencia:** Independencia de los residuales.
- **Homoscedasticidad:** La varianza de los residuos es constante en todos los niveles de las variables independientes.
- **Normalidad:** Los residuos se distribuyen normalmente.

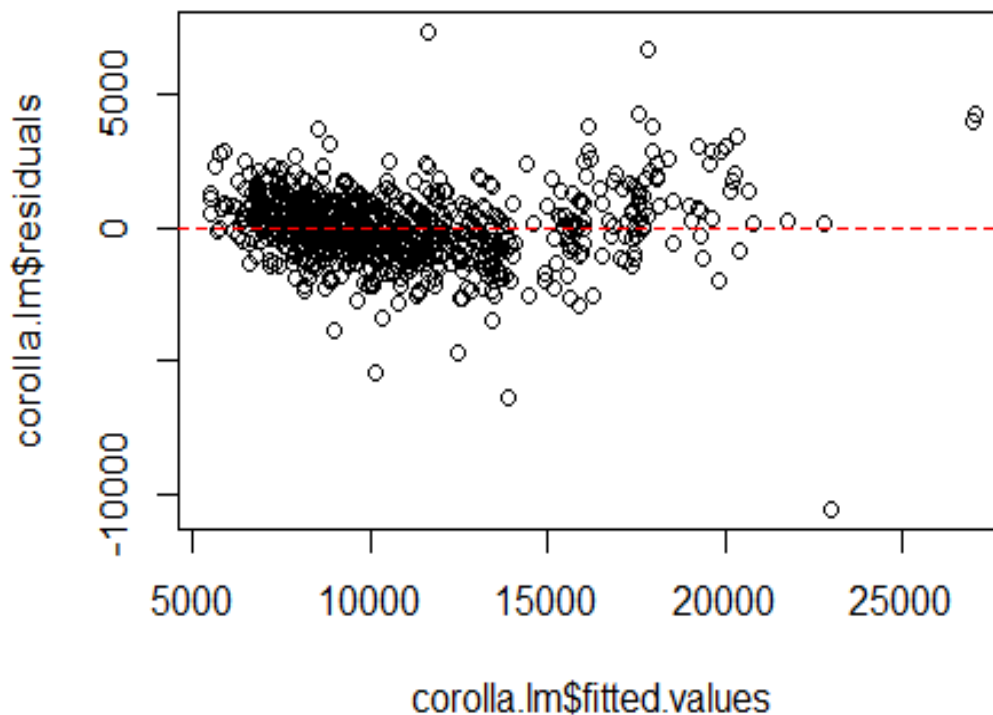
De esta manera, los siguientes gráficos ayudarán a visualizar la linealidad, independencia, homoscedasticidad y normalidad de los residuos.

i. Linealidad

La función “plot()” se utiliza para crear un gráfico de dispersión entre los valores ajustados y los residuales. Cada punto en el gráfico representa una observación en tu conjunto de datos.

A partir del análisis de linealidad se puede comprobar la relación lineal entre las variables independientes y la variable dependiente. Se observa un patrón lineal entre los datos.

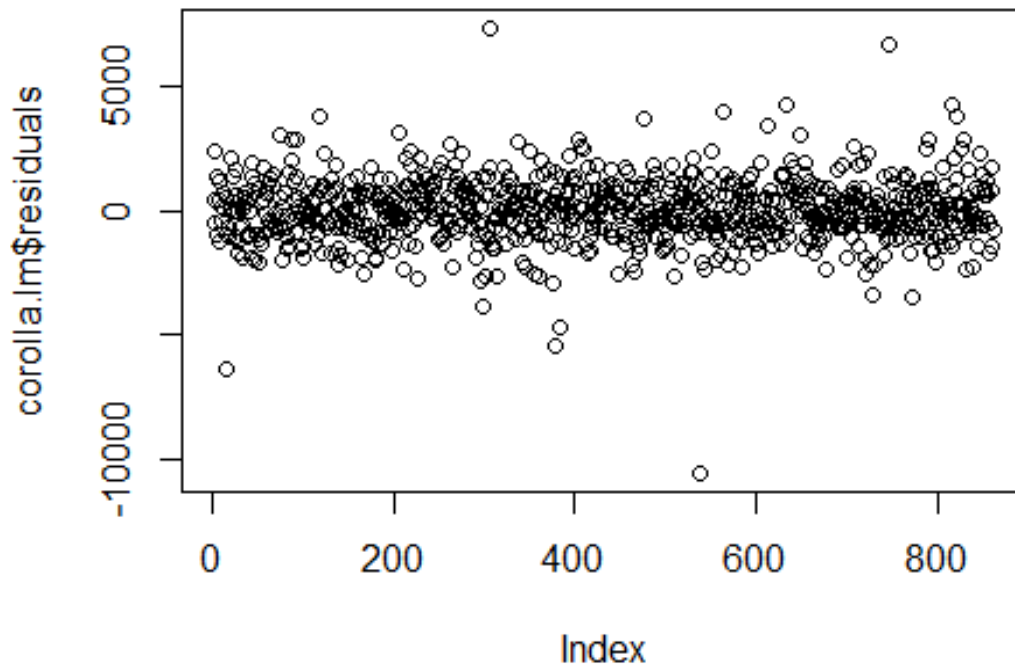
```
plot(corolla.lm$fitted.values, corolla.lm$residuals)  
abline(h = 0, col = "red", lty = 2)
```



ii. Independencia

Este código crea un gráfico de dispersión donde en el eje x se muestran las observaciones, y en el eje y se muestran los valores de los residuales correspondientes a esas observaciones. Cada punto en el gráfico representa la diferencia entre el valor observado y el valor predicho por el modelo para una observación específica. Se observa que cada valor es independiente y lineal, lo cual indica que el modelo es adecuado.

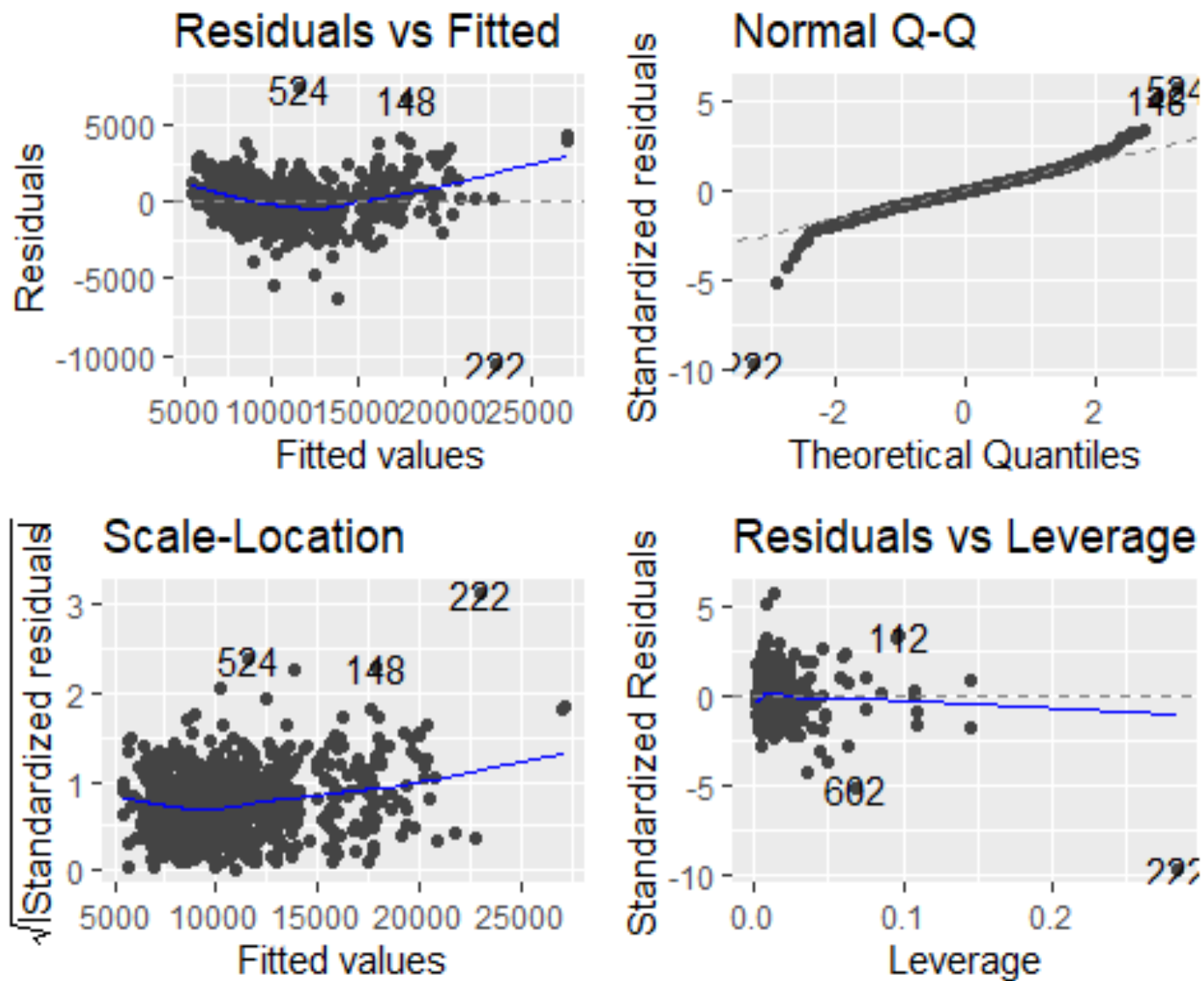
```
plot(corolla.lm$residuals, type = "p")
```



iii. Homoscedasticidad

Este gráfico comprueba que la varianza de los datos es constante, es decir, presentan homocedasticidad. La homocedasticidad es una propiedad deseable de los errores de un modelo de Regresión simple ya que permite realizar modelos más confiables. Por lo tanto, se puede decir que los datos presentan homocedasticidad, lo cual indica un modelo confiable.

```
autoplot(corolla.lm)
```

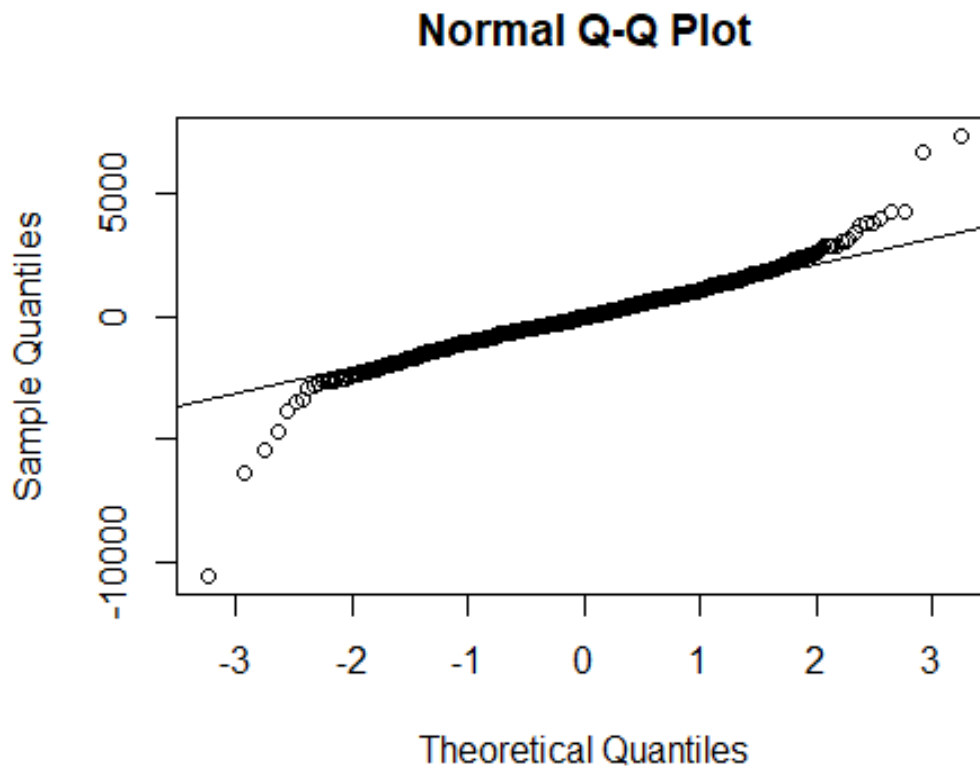


iv. Normalidad

Los residuales muestran una distribución normal al presentar una línea recta en el gráfico. La distribución normal de los residuales provoca estimaciones e intervalos de predicción más precisas, lo que indica que el MLR es aceptable y confiable.

```
qqnorm(corolla.lm$residuals)
```

```
qqline(corolla.lm$residuals)
```



Ajuste del Modelo

Ahora se analizará la veracidad del MLR. Esto implica el uso de métricas como (RMSE) o (R cuadrado) para evaluar el rendimiento del modelo. Estas métricas proporcionan una medida de la calidad global del ajuste, considerando la proporción de variabilidad “Price” explicada por el modelo y ajustándola según el número de variables independientes incluidas.

En el análisis del modelo de regresión lineal múltiple aplicado a los precios de vehículos usados de la marca Toyota, se observa un rango de precios que oscila entre \$ 4,350 y \$ 32,500. En relación con este contexto, el Root Mean Squared Error (RMSE) del modelo es de \$1239.72, que representa aproximadamente el 3.8% del rango total de precios. Este cálculo ofrece una perspectiva relativa de la magnitud de los errores en relación con la escala de los precios reales. Cabe resaltar que la evaluación precisa del rendimiento del modelo debe considerar el contexto específico del problema y las expectativas con respecto a la precisión del modelo en la predicción de los precios de los vehículos usados de la marca Toyota.

```
summary(corolla_complete$Price)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4350    8450    9900   10731   11950   32500

accuracy(corolla.lm.pred, validating.df$Price)

##              ME    RMSE     MAE        MPE     MAPE
## Test set 56.07729 1239.72 937.6244 -0.4239978 9.255647
```

Por otro lado, un R cuadrado cercano a 1 indica un buen ajuste. Además, es esencial analizar la importancia del modelo en su conjunto, evaluando la relevancia y contribución de las variables predictoras. El análisis del modelo de Regresión Lineal Múltiple revela información valiosa sobre la calidad del ajuste y la importancia de las variables independientes. El “adj.r.squared” de 0.8703 indica que aproximadamente el 87.03% de la variabilidad en “Price” puede ser explicada por las variables independientes incluidas en el modelo.

```
adjusted_r_squared <- summary(corolla.lm)$adj.r.squared

cat("El R cuadrado ajustado es:", adjusted_r_squared, "\n")

## El R cuadrado ajustado es: 0.8703404
```

En su conjunto, el modelo presenta un buen ajuste general con un alto R cuadrado ajustado y un RMSE aceptable, y las variables independientes incluidas son estadísticamente significativas en la predicción del precio del automóvil, lo que respalda su importancia en el modelo seleccionado.

Análisis de las Variables y sus Coeficientes

Se procederá a evaluar la importancia de cada variable independiente y de sus coeficientes mediante la utilización de valores p, destacando especialmente aquellas que se revelen como predictores estadísticamente significativos de la variable dependiente “Price”.

Para visualizar esto, se utiliza la función “gvlma()” para realizar una prueba global de diagnóstico para validar varias suposiciones del modelo lineal (ii), como la linealidad, la homocedasticidad y la normalidad de los residuales. La función toma como entrada el modelo de Regresión Lineal del training (corolla.lm).

```
gv_model <- gvlma(x= corolla.lm)
summary (gv_model)

##
## Call:
## lm(formula = Price ~ ., data = training.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10533.8   -677.7    -25.0     720.6    7323.9
##
```

```
## Coefficients:
##              Estimate      Std. Error t value      Pr(>|t|)
## (Intercept) -3024978.325233    80786.095663 -37.444 < 0.00000000000000
002
## Mfg_Year      1507.722648       40.644627  37.095 < 0.00000000000000
002
## KM            -0.017650         0.001636 -10.788 < 0.00000000000000
002
## HP            39.467270         3.805911  10.370 < 0.00000000000000
002
## cc           -2.475342         0.394974  -6.267      0.0000000000
583
## Gears         490.700743       238.566177   2.057      0.04
000
## Quarterly_Tax  10.935451        1.900825   5.753      0.000000012
216
## Weight        17.077551        1.395565  12.237 < 0.00000000000000
002
## Guarantee_Period 40.402103      13.785237   2.931      0.00
347
##
## (Intercept)    ***
## Mfg_Year       ***
## KM             ***
## HP            ***
## cc            ***
## Gears          *
## Quarterly_Tax  ***
## Weight         ***
## Guarantee_Period **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1292 on 853 degrees of freedom
## Multiple R-squared:  0.8715, Adjusted R-squared:  0.8703
## F-statistic: 723.4 on 8 and 853 DF,  p-value: < 0.0000000000000022
##
##
```

```
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance = 0.05
##
## Call:
## gvlma(x = corolla.lm)
##
##              Value      p-value      Decision
## Global Stat    2417.9517 0.00000000 Assumptions NOT satisfied!
## Skewness       17.7213 0.00002557 Assumptions NOT satisfied!
## Kurtosis       2274.2151 0.00000000 Assumptions NOT satisfied!
## Link Function   125.6999 0.00000000 Assumptions NOT satisfied!
## Heteroscedasticity 0.3154 0.57441531 Assumptions acceptable.
```

i. Importancia de cada variable independiente

Todas las variables presentan p-valores significativamente bajos, indicando que son estadísticamente significativas en la predicción del precio del automóvil. Esto sugiere que cada variable contribuye de manera significativa a explicar la variabilidad en el precio.

ii. Variables predictores estadísticamente significativos

Las variables más destacadas en términos de impacto estadístico significativo en el precio del automóvil son 'Mfg_Year', 'KM', 'HP', 'cc', 'Gears', 'Quarterly_Tax', 'Weight', y 'Guarantee_Period'. Estas variables tienen p-valores muy bajos, lo que sugiere una relación estadísticamente significativa con el precio del automóvil.

iii. Interpretación de los coeficientes

- *Mfg_Year*: Un aumento de una unidad en el año de fabricación resulta en un aumento estimado de \$1507.72 en el precio del automóvil, manteniendo constantes las otras variables.
- *KM*: Un aumento de una unidad en el kilometraje se asocia con una disminución estimada de \$0.0177 en el precio del automóvil, manteniendo constantes las otras variables.
- *HP*: Un aumento de una unidad en la potencia del motor ('HP') se traduce en un aumento estimado de \$39.47 en el precio del automóvil, manteniendo constantes las otras variables.

- *cc*: Un aumento de una unidad en la cilindrada ('cc') está asociado con una disminución estimada de \$2.48 en el precio del automóvil, manteniendo constantes las otras variables.
- *Gears*: Un cambio de una unidad en la cantidad de engranajes ('Gears') resulta en un aumento estimado de \$490.70 en el precio del automóvil, manteniendo constantes las otras variables.
- *Quarterly_Tax*: Un aumento de una unidad en el impuesto trimestral ('Quarterly_Tax') se asocia con un aumento estimado de \$10.94 en el precio del automóvil, manteniendo constantes las otras variables.
- *Weight*: Un aumento de una unidad en el peso ('Weight') resulta en un aumento estimado de \$17.08 en el precio del automóvil, manteniendo constantes las otras variables.
- *Guarantee_Period*: Un cambio de una unidad en el período de garantía ('Guarantee_Period') se asocia con un aumento estimado de \$40.40 en el precio del automóvil, manteniendo constantes las otras variables.

Implementación del Modelo de Regresión Logística

En esta sección, se hará el análisis de variables consideradas categóricas mediante el uso de la Regresión Logística. El objetivo principal es evaluar la posible influencia de estas variables en los precios de los automóviles. En particular, nos centraremos en atributos como el tipo de combustible, el color metalizado, la transmisión automática, garantías del fabricante, garantías BOVAG, sistema de frenos antibloqueo (ABS), presencia de airbags, aire acondicionado, sistema de audio, sistema de cierre centralizado, ventanas eléctricas, dirección asistida, sistema de radio, presencia de niebla, modelo deportivo, divisor de asiento trasero, llantas metálicas, reproductor de radio y casete, y la presencia de un gancho de remolque. Este análisis nos permitirá comprender cómo estas características influyen en la variabilidad de los precios de los vehículos, proporcionando así información valiosa para la toma de decisiones y estrategias de precios en el mercado automotriz.

A diferencia de la regresión lineal múltiple, en la Regresión Logística, la relación entre la variable dependiente Y y los parámetros β no es lineal, por lo tanto, la estimación de β no sigue el método de mínimos cuadrados. En cambio, se emplea el método de máxima verosimilitud, que busca maximizar la probabilidad de obtener los datos existentes a través de algoritmos.

Factorización de Atributos

Una vez seleccionado el subset con los atributos categóricos, se factorizarán tomando 0 como FALSE (no tienen esa característica) y 1 como TRUE (sí la tienen).

```
corolla_full <- read.csv("corolla.csv")

data.lr = corolla_full %>%

select("Price", "Fuel_Type", "Met_Color", "Automatic", "Mfr_Guarantee", "BOVAG_Guarantee", "ABS", "Airbag_1", "Airbag_2", "Airco", "Automatic_airco", "Boardcomputer", "CD_Player", "Central_Lock", "Powered_Windows", "Power_Steering", "Radio", "Mistlamps", "Sport_Model", "Backseat_Divider", "Radio_cassette", "Tow_Bar")
```

En otras palabras, en esta fase del análisis, se ejecutó una transformación esencial en las variables predictoras, llevándolas de un formato binario de 0 y 1 a sus equivalentes lógicos, es decir, de "ausente" a "false" y de "presente" a "true". Esta operación, conocida como factorización,

resultó fundamental para adecuar las características del conjunto de datos al formato más idóneo para la creación del modelo de Regresión Logística. Al convertir las variables binarias en factores lógicos, se simplifica la interpretación de los coeficientes del modelo, ya que ahora representan de manera más intuitiva el impacto de la presencia o ausencia de cada característica en la predicción de si un automóvil estará por encima o por debajo del precio promedio, tomando en cuenta ese atributo categórico específico.

```
data.lm$Fuel_Type=as.factor(data.lm$Fuel_Type)

contrasts(data.lm$Fuel_Type)

##           Diesel Petrol
## CNG           0      0
## Diesel        1      0
## Petrol        0      1

library(dplyr)

# Excluir las columnas Fuel_Type y Price
columns_to_exclude <- c("Fuel_Type", "Price")

# Convertir todas las columnas de tipo integer a factor, excluyendo las especificadas
data.lm <- mutate_at(data.lm, setdiff(names(data.lm), columns_to_exclude), function(x) {
  if (is.integer(x)) {
    x <- ifelse(x == 0, "No", "Sí")
    x <- as.factor(x)
  }
  return(x)
})
```

Division y Factorización de Price

En esta etapa del análisis, se llevó a cabo una clasificación de los precios en la base de datos para adaptar la variable de salida a los requisitos de la Regresión Logística, la cual necesita una variable de salida binomial categórica. Para este propósito, se dividió la variable de precio en dos categorías: "por debajo del promedio (BA)" y "por encima del promedio (AA)". Esta segmentación se realizó tomando como referencia la media de los precios en el conjunto de datos.

```
promedio_price <- mean(logistic.df$Price)
promedio_price

## [1] 10730.82
```

Posteriormente, se procedió a factorizar el atributo de precio para asignarle un formato lógico, representando "BA" como false y "AA" como true. Estos pasos son esenciales para preparar adecuadamente los datos y construir un modelo de regresión logística que pueda predecir la probabilidad de que un automóvil tenga un precio superior o inferior al promedio, basándose en los atributos seleccionados.

```
data.lr$Price=ifelse(test=data.lr$Price<mean(data.lr$Price),
  yes="BA", no="AA")

data.lr$Price=factor(data.lr$Price,levels=c("BA", "AA"))
```

Revisión de distribución de instancias

```
countAA=length(which(data.lr$Price=="AA"))
countBA=length(which(data.lr$Price=="BA"))

countAA

## [1] 541

countBA

## [1] 895
```

Creación Modelo de Regresión Logística Inicial

En la creación del modelo de Regresión Logística inicial, se empleó la función "glm" para generar un modelo que involucra todas las variables predictoras. Al analizar el resumen del modelo, se observa que algunas variables presentan una significancia nula o baja, lo que sugiere que podrían no ser influyentes para predecir si un automóvil tendrá un precio superior o inferior al promedio. Entre las variables que destacan por su significancia positiva se encuentran la garantía de manufactura, el sistema de frenos ABS, la presencia de aire acondicionado, tablero computarizado, lector de discos de CD, lámparas para niebla y la existencia de una barra de remolque.

Los coeficientes estimados y sus respectivos valores p indican la relación entre cada variable y la probabilidad de que el precio del automóvil esté por encima del promedio. Variables como la garantía de manufactura, el sistema ABS, la presencia de aire acondicionado, tablero computarizado, lector de discos de CD, lámparas para niebla y la barra de remolque muestran significancia estadística positiva, lo que sugiere una influencia positiva en la probabilidad de que el precio esté por encima del promedio. Estas variables se consideran como candidatas significativas para un modelo más refinado y preciso. Por otro lado, variables como **el tipo de combustible, color metálico**, y características como **el segundo airbag o el modelo deportivo**, no parecen tener una influencia estadísticamente significativa según sus valores p. Estos hallazgos proporcionan una base para la selección de variables predictoras más efectivas en la construcción de un modelo de Regresión Logística final.

```
logistic <- glm(Price ~., data = data.lr, family="binomial")
summary(logistic)
```

```
##
## Call:
## glm(formula = Price ~ ., family = "binomial", data = data.lr)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.7929     1.0111  -3.751 0.000176 ***
## Fuel_TypeDiesel  0.5644     0.8253   0.684 0.494046
## Fuel_TypePetrol  0.7242     0.7925   0.914 0.360808
## Met_ColorSí     -0.2306     0.1747  -1.320 0.186738
## AutomaticSí      0.6097     0.3235   1.885 0.059445 .
## Mfr_GuaranteeSí  0.8783     0.1702   5.161 2.45e-07 ***
## BOVAG_GuaranteeSí 0.1501     0.2977   0.504 0.614034
## ABSSí           1.6061     0.3325   4.831 1.36e-06 ***
## Airbag_1Sí       0.3130     0.6479   0.483 0.629022
## Airbag_2Sí      -0.4076     0.3050  -1.336 0.181437
## AircoSí          1.1278     0.1940   5.812 6.17e-09 ***
## Automatic_aircoSí 1.7403     0.5212   3.339 0.000840 ***
## BoardcomputerSí  2.4384     0.1948  12.515 < 2e-16 ***
## CD_PlayerSí      1.3180     0.2114   6.236 4.50e-10 ***
## Central_LockSí   -0.1469     0.3260  -0.451 0.652285
## Powered_WindowsSí 0.7481     0.3208   2.332 0.019709 *
## Power_SteeringSí -0.8666     0.6589  -1.315 0.188457
## RadioSí          1.0556     1.3535   0.780 0.435463
## MistlampsSí     -0.7084     0.2334  -3.036 0.002400 **
## Sport_ModelSí    -0.4618     0.2038  -2.266 0.023462 *
```

```
## Backseat_DividerSí -0.2132      0.2957 -0.721 0.470865
## Radio_cassetteSí   -0.1539      1.3576 -0.113 0.909773
## Tow_BarSí         -0.5053      0.1803 -2.802 0.005078 **

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1902.5  on 1435  degrees of freedom
## Residual deviance: 1076.6  on 1413  degrees of freedom
## AIC: 1122.6
##
## Number of Fisher Scoring iterations: 5
```

Validación

Ahora, se selecciona aleatoriamente el 40% de las instancias de la base de datos para evaluar el rendimiento del modelo de Regresión Logística inicial. Al aplicar el modelo a este conjunto de prueba, se obtiene un rendimiento positivo, logrando acertar un 54.5% de las clasificaciones del precio. Este resultado se calcula comparando las predicciones del modelo con las categorías reales de los precios (por debajo o por encima del promedio) en el conjunto de prueba. El porcentaje de precisión (accuracy) se calcula como la proporción de clasificaciones correctas respecto al total de instancias. La precisión del 54.5% indica que el modelo tiene un rendimiento aceptable en la predicción de los precios de los automóviles en función de las variables consideradas. Sin embargo, este resultado también sugiere que hay margen para mejorar la precisión del modelo en futuras iteraciones y ajustes.

```
set.seed(1)
test.index=sample(c(1:length(data.lr$Price)),(0.4*length(data.lr$Price)))
test.df=data.lr[test.index,]
data.test=predict(logistic,test.df,type="response")
data.predict = rep("BA", dim(test.df)[1])
data.predict[data.test> .5] = "AA"
acc=sum(data.predict==data.lr)/length(data.lr$Price)
acc

## [1] 0.545961
```

Prueba Mejora de Modelo

En esta etapa del análisis, se procedió a realizar mejoras en el modelo de Regresión Logística inicial, tomando en cuenta las observaciones detalladas previamente. Se eliminaron las variables que mostraron una baja significancia y se generó un segundo modelo, el cual se evaluó utilizando la misma base de datos que el modelo original. El segundo modelo, que incorpora las variables consideradas más relevantes, exhibió una leve mejoría en su rendimiento. Al aplicar este modelo ajustado al conjunto de prueba, se observó que logra acertar un 55.2% de las clasificaciones del precio de los autos. Este aumento en la precisión sugiere que la selección cuidadosa de variables contribuyó positivamente al desempeño del modelo en la predicción de los precios de los automóviles.

```
prueba=glm(Price~Mfr_Guarantee+ABS+Airco+Automatic_airco+Boardcomputer+CD_Player+Mistlamps+Tow_Bar,data=data.lr,family="binomial")
summary(prueba)
```

```
##
## Call:
## glm(formula = Price ~ Mfr_Guarantee + ABS + Airco + Automatic_airco +
##       Boardcomputer + CD_Player + Mistlamps + Tow_Bar, family = "binomial",
##       data = data.lr)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.2867     0.2739  -11.998  < 2e-16 ***
## Mfr_GuaranteeSí    0.7180     0.1528   4.700 2.60e-06 ***
## ABSSí            1.1214     0.2654   4.225 2.39e-05 ***
## AircoSí           1.3870     0.1735   7.996 1.28e-15 ***
## Automatic_aircoSí  1.8548     0.5214   3.558 0.000374 ***
## BoardcomputerSí   2.2702     0.1758  12.912  < 2e-16 ***
## CD_PlayerSí       1.0638     0.1979   5.375 7.68e-08 ***
## MistlampsSí      -0.6715     0.1944  -3.455 0.000551 ***
## Tow_BarSí        -0.3637     0.1702  -2.136 0.032650 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1902.5  on 1435  degrees of freedom
## Residual deviance: 1140.9  on 1427  degrees of freedom
## AIC: 1158.9
##
## Number of Fisher Scoring iterations: 5

data.test2=predict(prueba,test.df,type="response")
data.predict2 = rep("BA", dim(test.df)[1])
data.predict2[data.test2> .5] = "AA"
acc2=sum(data.predict2==data.lr)/length(data.lr$Price)
acc2

## [1] 0.5522284
```

En general, la fase de aplicación de la Regresión Logística en este estudio proporcionó una perspectiva valiosa sobre los factores determinantes que influyen en la variable “Price” de los automóviles. Al ajustar y refinar el modelo inicial, se logró mejorar su capacidad predictiva, alcanzando una tasa de aciertos del 55.2% en la clasificación de los precios.

$$h(x)=g(-3.2867 + 0.7180 \times \text{Mfr_Guarantee} + 1.1214 \times \text{ABS} + 1.3870 \times \text{Airco} + 1.8548 \times \text{Automatic_airco} + 2.2702 \times \text{Boardcomputer} + 1.0638 \times \text{CD_Player} - 0.6715 \times \text{Mistlamps} - 0.3637 \times \text{Tow_Bar})$$

donde $h(x)$ representa el modelo de Regresión Logística y $g()$ es la función sigmoide, que transforma la entrada en un valor entre 0 y 1; cuyo umbral se establecerá en $g(z) = 10730.82$, lo que significa que si el resultado de la hipótesis supera este valor, la predicción será clasificada como 1 o TRUE.

Este resultado, combinado con el análisis previo de Regresión Lineal, ofrece una visión integral de los factores más relevantes que afectan el precio de los automóviles. La combinación de técnicas estadísticas, como la Regresión Lineal y Logística, permite una comprensión más completa y precisa de los factores que inciden en la determinación de los precios, lo cual es fundamental para la toma de decisiones en el contexto del presente proyecto.

Conclusiones

Para concluir este trabajo, se recapitulará de manera breve lo realizado. Siguiendo la metodología CRISP-DM, primero se realizó un breve análisis de exploración de los datos (EDA), en el cual se aseguró el entendimiento de los atributos, la inexistencia de datos faltantes, y la inexistencia de multicolinealidad mediante una matriz de correlación. Posteriormente se generaron varios modelos de regresión lineal múltiple (MLR): el primer incluyente de todas las variables numéricas, el segundo excluyendo las variables que mostraban insignificancia en el primer modelo, y el tercero siendo un step forward del segundo modelo para comprobar que no existe posibilidad de mejoría. Posteriormente se generaron subsets para entrenamiento y prueba del modelo seleccionado; y a partir de estos, se generaron predicciones y cálculo de residuales. Siguiendo, se analizaron los residuales, el ajuste del modelo, y se interpretó el resultado del modelo. Por último, se generó un modelo de regresión logística (LR) en el que se tomaron en cuenta las variables lógicas para evaluar una posible relación entre estas y la variable de salida.

A partir del modelo de regresión lineal múltiple (MLR) se obtuvieron los siguientes hallazgos:

- Se eligió el segundo modelo realizado que indica que los atributos “Mfg_Year”, “”KM”, “HP”, “cc”, “Gears”, “Quarterly_Tax”, “Weight” y “Guarantee_Period” son relevantes y significativamente relacionados a la variable de salida “Price”.
- El modelo de regresión lineal elegido arroja un R^2 de 0.87, que indica que el 87% de los datos pueden ser explicados por el modelo.
- Los residuos obtenidos en base a las predicciones y los datos actuales son aceptables.
- El Root Mean Squared Error (RMSE) del modelo es de \$1239.72, que representa aproximadamente el 3.8% del rango total de precios.

Por lo tanto, el modelo presenta un buen ajuste general de los datos con un alto R cuadrado ajustado y un RMSE aceptable, y las variables independientes incluidas son estadísticamente significativas en la predicción del precio del automóvil, lo que respalda su importancia en el modelo seleccionado.

A partir del modelo de regresión logística se obtuvieron los siguientes hallazgos:

- Las variables que muestran mayor relación con la variable de salida son: “Mfr_Guarantee”, “ABS”, “Airco”, “Automatic_airco”, “Boardcomputer”, “CD_Player”, “Mistlamps” y “Tow_Bar”.
- El modelo con mayor precisión mostró un 55.2% asertividad en la predicción de la clasificación del precio de autos.

Por lo tanto, el modelo de regresión logística se puede considerar medianamente aceptable para obtener conclusiones sobre relaciones entre variables.

Dando respuesta a la pregunta de investigación: “¿Qué factores determinan el precio de venta de los vehículos marca Toyota usados en México?”, utilizando ambos métodos de modelación (MLR y LR), se puede concluir que los atributos o factores que determinan el precio de venta de los vehículos Toyota usados en México son: “Mfg_Year”, “”KM”, “HP”, “cc”, “Gears”, “Quarterly_Tax”, “Weight”, “Guarantee_Period”, “Mfr_Guarantee”, “ABS”, “Airco”, “Automatic_airco”, “Boardcomputer”, “CD_Player”, “Mistlamps” y “Tow_Bar”.

Entre las limitaciones de un modelo de regresión lineal múltiple (MLR) se encuentra la posibilidad de multicolinealidad ya que se puede generar un sesgo en el modelo; la sensibilidad a outliers; underfitting u overfitting de los datos puede afectar la validación; y el modelo genera suposiciones importantes ya que asumen que la relación lineal entre las variables independientes y la variable dependiente; además, asume la normalidad y homocedasticidad de los errores. Entre las limitaciones de un modelo de regresión logística se encuentra la limitada veracidad del modelo; y la incapacidad para manejar relaciones no lineales, lo que significa que no puede modelar relaciones no lineales entre las variables predictoras y la variable dependiente sin la inclusión de términos polinómicos. Como sugerencia a posibles mejoras se encuentra la posibilidad de obtener más datos, ya que entre más datos, más específico es el modelo; o incluso implementar otro modelo de regresión como el Análisis de Varianza (ANOVA), el Análisis de Covarianza (ANCOVA), o Regresión Polinómica que se encuentran fuera del alcance de este curso.

Bibliografía

- [1] Khandelwal, R. (2021). *A Step-by-Step Guide to Multiple Linear Regression in R*. Medium. <https://arshren.medium.com/a-step-by-step-guide-to-multiple-linear-regression-in-r-a85d270f70f7>
- [2] Gustavo. (2022). *Understanding Logistic Regression step by step*. Medium. <https://towardsdatascience.com/understanding-logistic-regression-step-by-step-704a78be7e0a>
- [3] Singh, S. (2022). *Exploratory Data Analysis (EDA) in R: A Step-by-Step Guide with Commands*. Medium. <https://medium.com/towards-data-engineering/exploratory-data-analysis-eda-in-r-a-step-by-step-guide-with-commands-b9acbe1d557d>
- [4] Dey, J. (2022). *Step-by-step Basic Data Cleaning in R*. Medium. <https://medium.com/@joyeetadey/step-by-step-basic-data-cleaning-in-r-3441c9cf6096>

Anexos

Acceso a la data set completo empleado para realizar este proyecto: [corolla.csv](#)