ENAP | Maestría en Métodos Cuantitativos para la Gestión y Análisis de Datos en Organizaciones

Nivelación: Conceptos básicos de estadística y econometría

Trabajo Práctico Integrador

Maestrando: Bravo, Juan José

Año: 2024

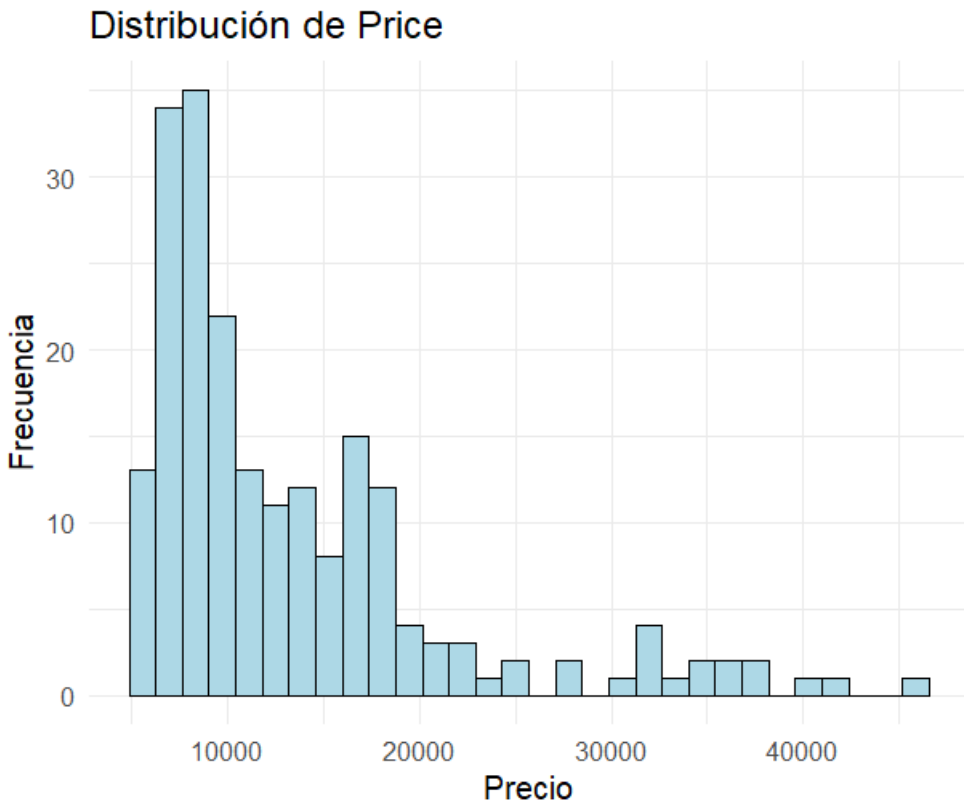
Contenido

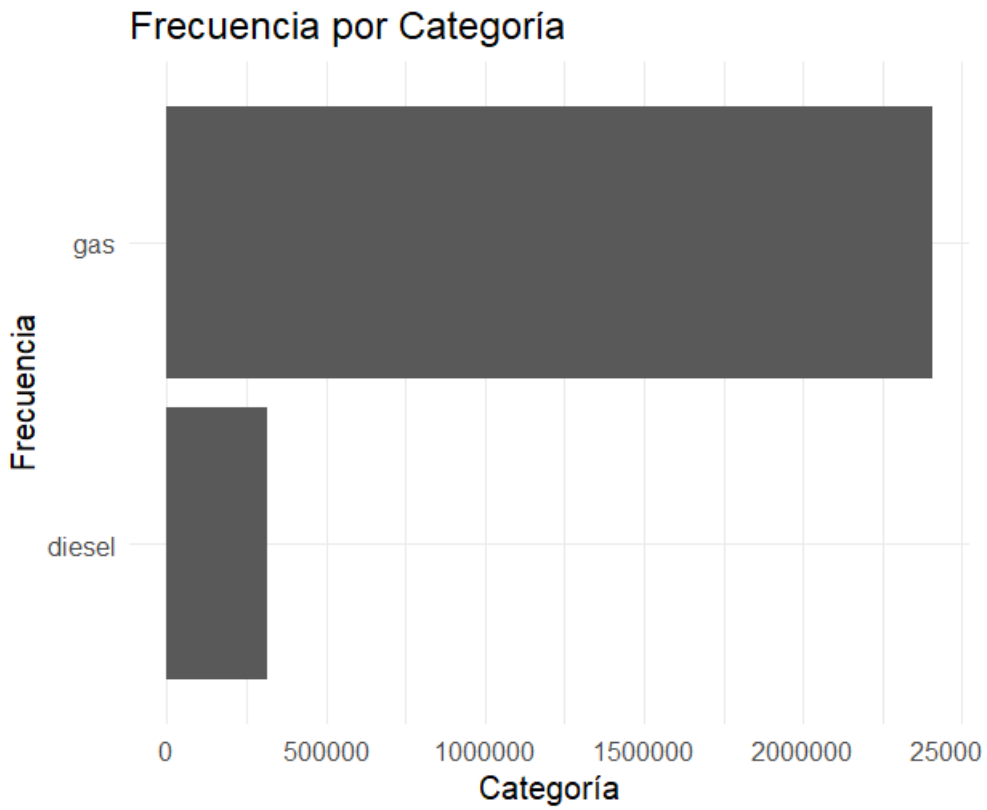
1. Análisis exploratorio de Datos

Realizamos un análisis exploratorio para conocer nuestro conjunto de datos y verificar por ejemplo como se comportan algunas categorías, de que tipo son, etc. En el conjunto de datos no se encuentran datos faltantes.

```
> dim(df)
[1] 205  26
```

```
> str(df)
'data.frame':   205 obs. of  26 variables:
 $ car_ID          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ symboling       : int  3 3 1 2 2 2 1 1 1 0 ...
 $ CarName         : chr  "alfa-romero giulia" "alfa-romero stelvio" "alfa-romero Quadrifoglio" "audi 100 ls"
...
 $ fueltype        : chr  "gas" "gas" "gas" "gas" ...
 $ aspiration      : chr  "std" "std" "std" "std" ...
 $ doornumber      : chr  "two" "two" "two" "four" ...
 $ carbody         : chr  "convertible" "convertible" "hatchback" "sedan" ...
 $ drivewheel      : chr  "rwd" "rwd" "rwd" "fwd" ...
 $ enginelocation  : chr  "front" "front" "front" "front" ...
 $ wheelbase       : num  88.6 88.6 94.5 99.8 99.4 ...
 $ carlength       : num  169 169 171 177 177
```

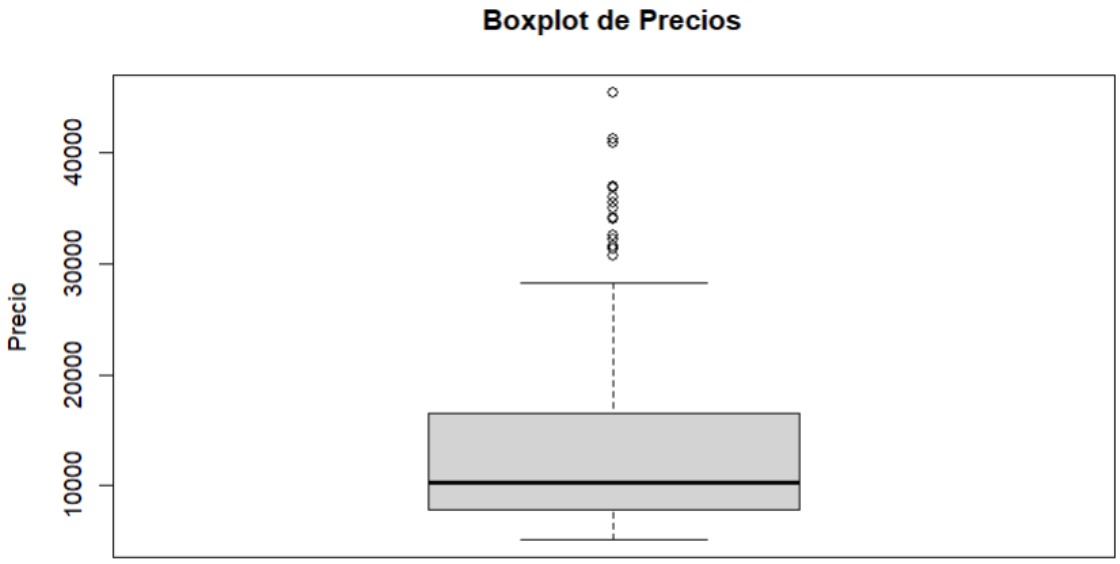





```
> asimetria_charges <- skewness(df_selected_o$price)
> asimetria_charges
[1] 1.764644
> kurtosis_charges <- kurtosis(df_selected_o$price)
> kurtosis_charges
[1] 5.948598
```

Con el valor obtenido de la asimetría podemos observar que, al ser positivo, la cola va hacia la derecha por lo que estos datos se encuentran sesgados positivamente. Esto nos da un indicio de que hay algunos autos que tienen los precios muy altos.

En cuanto a la kurtosis, se puede interpretar que con el número obtenido, se está ante la presencia de colas más largas hacia los costados y un pico pronunciado, esto también está indicando la presencia de valores atípicos en el precio de los autos.



Se procede a trabajar los outliers identificados en la variable Price y se vuelve a medir el modelo:

```
> asimetria_charges <- skewness(df_selected_o$price)
> asimetria_charges
[1] 1.213071
> kurtosis_charges <- kurtosis(df_selected_o$price)
> kurtosis_charges
[1] 3.644283
```

Estos valores están indicando que, aunque el valor ha disminuido en comparación con el anterior (1.764644), sigue indicando que hay una cola más larga hacia los precios más altos. El nuevo valor de curtosis es 3.644283 indica que manejando los outliers, la distribución ha disminuido y se ha vuelto más normal.

2. Estimación por MCO de un modelo de regresión lineal

```
Call:
lm(formula = price ~ fueltype + carbody + enginesize + horsepower +
    curbweight, data = df_selected)

Residuals:
    Min      1Q  Median      3Q     Max
-9012.5 -1525.7  -131.4  1668.0 13884.8

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -6988.681   2436.942  -2.868 0.004586 **
fueltypegas      -1573.831    949.058  -1.658 0.098855 .
carbodyhardtop   -1864.311   1809.130  -1.031 0.304045
carbodyhatchback -4863.867   1430.599  -3.400 0.000817 ***
carbodysedan     -3467.408   1409.477  -2.460 0.014757 *
carbodywagon     -5290.872   1563.555  -3.384 0.000863 ***
enginesize          68.916     12.907   5.339 2.57e-07 ***
horsepower          59.799     12.123   4.933 1.73e-06 ***
curbweight           4.193      1.086   3.861 0.000153 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3308 on 196 degrees of freedom
Multiple R-squared:  0.8352,    Adjusted R-squared:  0.8285
F-statistic: 124.2 on 8 and 196 DF,  p-value: < 2.2e-16
```

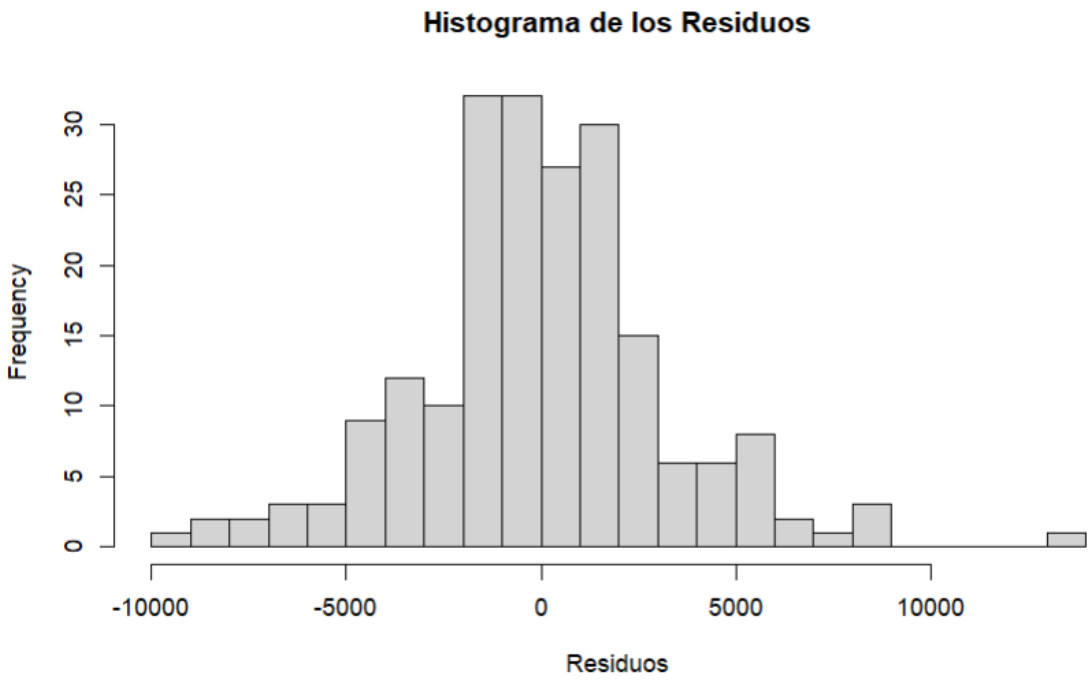
Con este modelo se pueden observar algunos puntos a tener en cuenta:

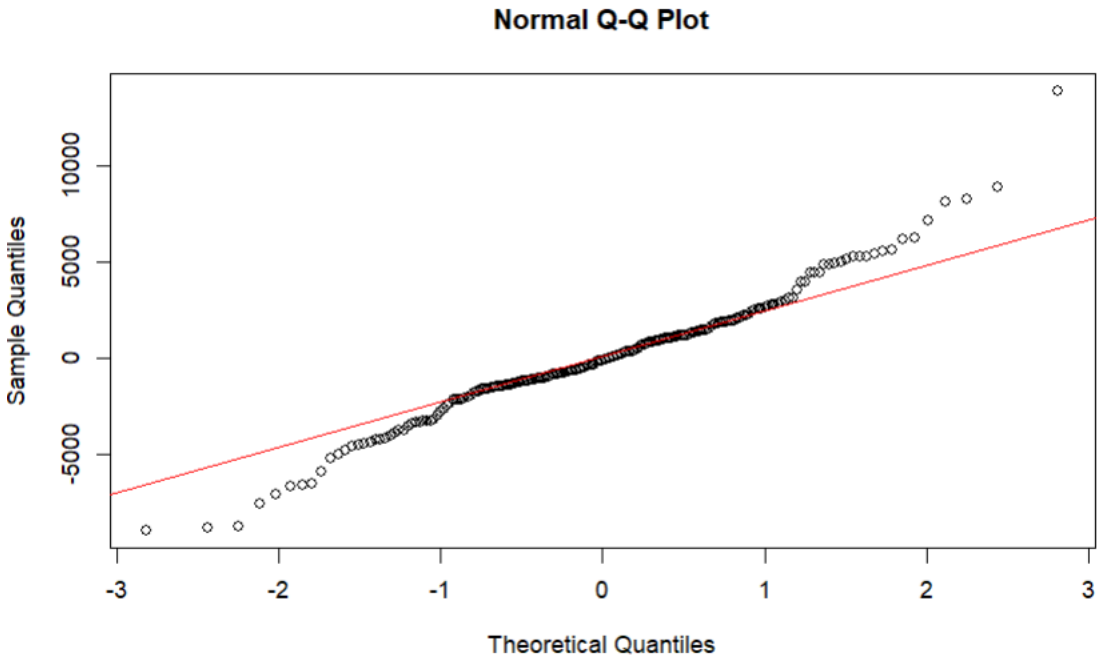
- El precio de los autos que usan gasolina es, en promedio, $1573.83 menor que los autos con otro tipo de combustible.
- **Hardtop (-1864.31)** y **hatchback (-4863.87)** muestran una relación negativa con el precio, lo que indica que estos tipos de carrocería tienden a reducir el precio en comparación con el grupo de referencia.
- **Sedan (-3467.41)** y **wagon (-5290.87)** también muestran una disminución en el precio, siendo la carrocería wagon la que más reduce el precio de los autos.
- **Enginesize (68.92, p < 0.001)**: Cada incremento de una unidad en el tamaño del motor aumenta el precio del auto en $68.92, y este efecto es altamente significativo.
- **Horsepower (59.80, p < 0.001)**: Cada incremento de una unidad en la potencia del motor incrementa el precio del auto en $59.80, también con una alta significancia.
- **Curbweight (4.19, p < 0.001)**: Un mayor peso del auto tiene un pequeño pero significativo efecto positivo en el precio. Cada incremento de una unidad en el peso del auto está asociado con un aumento de $4.19 en el precio.

- **R-squared (0.8352)**: El modelo explica el 83.52% de la variabilidad en los precios de los autos, lo cual indica un buen ajuste del modelo.
- **Adjusted R-squared (0.8285)**: Ajustado por el número de variables en el modelo, el 82.85% de la variación en los precios es explicada por las variables exógenas seleccionadas. El valor de R2 ajustado es ligeramente menor que el R2 original (82.85% frente a 83.52%). Esto significa que, aunque se han incluido varias variables en el modelo, la mayoría de ellas contribuyen de manera significativa a explicar la variabilidad de price, con poca penalización por la cantidad de variables.
- **Residual Standard Error (3308)**: El error estándar de los residuos es de aproximadamente $3308, lo que indica la magnitud promedio de las desviaciones entre los valores observados y los predichos por el modelo. Esto es una estimación aceptable, pero también sugiere que algunas predicciones pueden estar lejos de los valores reales, posiblemente debido a los outliers que se vieron anteriormente.

3. Normalidad de los Residuos

A continuación, se trabaja con la normalización de los residuos y se muestra en gráficas los datos obtenidos:






```
        Shapiro-Wilk normality test

data:  residuos
W = 0.97235, p-value = 0.0004602
```

En este caso, el p-value es 0.0004602, que es menor que el umbral de 0.05. Esto significa que rechazamos la hipótesis nula (H_0) y concluimos que los residuos no siguen una distribución normal.

4. Multicolinealidad

```
> cor_matrix <- cor(df_selected_o[c("enginesize", "horsepower", "curbweight")])
> print(cor_matrix)
           enginesize horsepower curbweight
enginesize  1.0000000  0.8097687  0.8505941
horsepower  0.8097687  1.0000000  0.7507393
curbweight  0.8505941  0.7507393  1.0000000
```

La matriz obtenida identifica los siguientes puntos:

- **Enginesize y horsepower**: Correlación de **0.81.** Hay una fuerte correlación positiva entre el tamaño del motor y la potencia del motor. Esto sugiere que a medida que aumenta el tamaño del motor, también tiende a aumentar la potencia.

- **Enginesize y curbweight**: Correlación de **0.85.** Hay una correlación muy fuerte entre el tamaño del motor y el peso en vacío. Esto puede indicar que los vehículos con motores más grandes suelen ser más pesados.
- **Horsepower y curbweight**: Correlación de **0.75.** También hay una correlación positiva fuerte entre la potencia del motor y el peso en vacío. A medida que aumenta la potencia, el peso del vehículo también tiende a aumentar.

```
> vif_values <- vif(modelo_full)
> print(vif_values)
                GVIF Df GVIF^(1/(2*Df))
enginesize 5.383782  1        2.320298
horsepower 4.283329  1        2.069621
curbweight 5.959102  1        2.441127
fueltype   1.485162  1        1.218672
carbody    1.574764  4        1.058405
```

De la evaluación del modelo planteado, se desprenden las siguientes interpretaciones:

- **enginesize** (VIF = 5.38): Un VIF superior a 5 indica una posible multicolinealidad. Este valor sugiere que la varianza de los coeficientes estimados para enginesize puede estar inflada debido a la correlación con otras variables en el modelo.
- **horsepower** (VIF = 4.28): Similar al anterior, este valor también está cerca del umbral de 5, lo que indica que puede haber cierta multicolinealidad con otras variables.
- **curbweight** (VIF = 5.96): Este valor es el más alto, indicando un alta multicolinealidad. Esto significa que curbweight está muy correlacionado con otras variables en el modelo, lo que puede afectar la estabilidad de los coeficientes.
- **fueltype** (VIF = 1.49) y **carbody** (VIF = 1.57): Ambos valores están por debajo de 2, lo que sugiere que estas variables no presentan problemas de multicolinealidad.

Algunas de las recomendaciones que se pueden aplicar a este modelo son:

1. La Eliminación de las variables fuertemente correlacionadas: Es posible generar un nuevo modelo eliminando una o más de las variables con una fuerte correlación, por ejemplo, curbweight. También, se puede trabajar eliminando la variable que menos sentido tenga para el modelo.
2. El uso de PCA: Se puede reducir la dimensionalidad de los datos haciendo un análisis de componentes principales.

5. Heterocedasticidad

a) Analizamos la heterocedasticidad con la prueba de White:

```
> white_test <- bptest(modelo_full, ~ fitted(modelo_full) + I(fitted(modelo_full)^2))
> print(white_test)

        studentized Breusch-Pagan test

data:  modelo_full
BP = 49.309, df = 2, p-value = 1.962e-11
```

- **Estadístico BP**: 49.309
- **Grados de libertad (df)**: 2
- **Valor p**: 1.962e-11

El valor p es extremadamente bajo (mucho menor que 0.05), lo que sugiere que se puede rechazar la hipótesis nula de homocedasticidad. Esto indica que hay evidencia de heterocedasticidad.

b) Analizamos la heterocedasticidad con la prueba de Park:

```
> summary(modelo_park)

Call:
lm(formula = log(residuos^2) ~ enginesize + horsepower + curbweight +
    fueltype + carbody, data = df_selected)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5678 -1.1787  0.3138  1.2897  4.1763

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     12.7967744  1.4814520   8.638    2e-15 ***
enginesize       0.0109643  0.0078463   1.397    0.164
horsepower       0.0101277  0.0073700   1.374    0.171
curbweight       0.0003186  0.0006602   0.483    0.630
fueltypegas     -0.6942648  0.5769462  -1.203    0.230
carbodyhardtop   0.4046240  1.0997960   0.368    0.713
carbodyhatchback -0.7706778  0.8696816  -0.886    0.377
carbodysedan    -1.0824661  0.8568415  -1.263    0.208
carbodywagon    -0.5829267  0.9505079  -0.613    0.540
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.011 on 196 degrees of freedom
Multiple R-squared:  0.2327,    Adjusted R-squared:  0.2014
F-statistic: 7.432 on 8 and 196 DF,  p-value: 1.232e-08
```

Los residuales muestran valores que varían entre -8.5678 y 4.1763, lo que sugiere que hay una dispersión considerable en los residuos, lo que puede indicar variabilidad en los errores.

- **Intercepto**: El coeficiente del intercepto es significativo ($p < 0.001$), lo que indica que el modelo tiene un valor base significativo.
- **enginesize**: No es significativo ($p = 0.164$), lo que sugiere que no hay evidencia suficiente para afirmar que enginesize influye en la varianza de los residuos.
- **horsepower**: Similar a enginesize, este coeficiente no es significativo ($p = 0.171$).
- **curbweight**: Este coeficiente también es no significativo ($p = 0.630$).
- **fueltypegas**: No es significativo ($p = 0.230$), lo que sugiere que el tipo de combustible no afecta la varianza de los errores.
- **carbody**: Todos los tipos de carrocería (hardtop, hatchback, sedan, wagon) tienen coeficientes no significativos, indicando que no contribuyen significativamente a la variabilidad de los errores.

Los resultados sugieren que, aunque el modelo tiene un p-valor global significativo, las variables individuales no muestran relaciones significativas. Esto podría indicar la presencia de heterocedasticidad en el modelo original. Para solucionar esto se podrían tomar algunas medidas correctivas como considerar incluir otras variables que se observen puedan estar afectando la varianza.

Conclusiones

- **Multicolinealidad**: Encontramos que algunas de las variables que usamos en el modelo están muy relacionadas entre sí. Esto puede dificultar entender el impacto de cada variable en el precio de los autos. Sería útil simplificar el modelo eliminando o combinando algunas de estas variables.
- **Heterocedasticidad**: Realizamos pruebas para verificar si los errores en nuestras predicciones son constantes y descubrimos que no lo son. Esto significa que algunas predicciones son menos precisas que otras. Esto puede hacer que los resultados no sean del todo confiables.
- **Ajuste del Modelo**: El modelo que creamos no explica bien la variabilidad en los precios de los autos, lo que indica que faltan factores importantes que podrían estar influyendo. Esto sugiere que podríamos necesitar más información o variables adicionales.
- **Significancia de Variables**: Muchas de las variables que incluimos no mostraron un efecto claro en el modelo, lo que indica que algunas de ellas pueden no ser relevantes. Esto sugiere que deberíamos de revisar la elección de las variables a utilizar.

Pasos a continuar trabajando pueden ser:

1. Revisar las variables con las que se trabajó y ver la posibilidad de incorporar otras que estén afectando por ejemplo la varianza.

2. Se pueden probar la aplicación de otros modelos.
3. Aplicar las transformaciones necesarias para evitar complicaciones a la hora de trabajar con variables que tienen una fuerte correlación como enginesize y curbweight.