Control 4

KND

25 de noviembre de 2020

1 Introducción

Los datos analizados en el artículo [1] se componen de distintas variables relacionadas con el avance tecnológico de varios vehículos híbridos eléctricos. En este documento se utiliza el mismo conjunto de datos para construir, ajustar y validar un modelo de regresión lineal múltiple que explique el precio de los vehículos en términos de variables como la tasa de aceleración y el consumo de combustible entre otras. El ajuste obtenido es utilizado posteriormente para la predicción del precio de los vehículos de otro conjunto de datos.

2 Resumen del artículo

En el artículo [1] se realiza un pronóstico de tecnología por medio de análisis por envolvente de datos (TFDEA por sus siglas en inglés) cuyo objetivo es medir y comparar el avance tecnológico ocurrido entre 2004 y 2013 en distintos sectores del mercado de vehículos híbridos eléctricos.

Los autores del artículo consideran como variable de entrada el precio sugerido por el fabricante de los vehículos. Como variables de salida consideran la tasa de aceleración, el consumo de combustible y el máximo de consumo de combustible o equivalente. También se incluye una variable categórica que divide a los vehículos en siete grupos diferentes.

De acuerdo con los resultados obtenidos en el artículo, los sectores correspondientes a los vehículos de tamaño medio fueron los que tuvieron mayores avances tecnológicos durante el periodo estudiado. Las mejoras en el desempeño y la diversificación de estos últimos puede representar una amenaza para los vehículos de tamaño pequeño. Mientras que los vehículos utilitarios deportivos (SUV) se enfocan en un nicho de mercado lujoso, los vehículos especializados para el transporte de mercancías compiten contra sus equivalentes de gasolina para probar la utilidad de los vehículos híbridos.

3 Descripción de las variables del conjunto de datos y análisis exploratorio

El conjunto de datos cuenta con las siguientes variables

Variable respuesta

La variable de respuesta que se busca es el precio sugerido del vehículo (`msrp` por sus siglas en inglés), que toma valores decimales y que en los datos originales está expresada en dólares equivalentes a 2013. Los valores mínimo y máximo que se tienen son 11849.43 y 118543.6 respectiva-

mente. Para expresar esta variable en miles de dólares, lo que hacemos es dividir esa columna entre 1000.

Potenciales regresores

- tasa de aceleración (`accelrate`): Esta variable numérica indica el tiempo en segundos que le toma a cada vehículo llegar de 0 km/h a 100 km/h.
- consumo de combustible (`mpg`): También conocida como economía de combustible, esta variable numéroca indica la distancia que un vehículo puede recorrer por unidad de combustible. En la fuente original de los datos se encuentra en millas por galón. En este documento se convirtió a kilómetros por litro multiplicando por 2.8248.
- máximo de consumo de combustible o equivalente (`mpgmpge`): La Agencia de Protección Ambiental de Estados Unidos desarrolló un consumo de combustible equivalente para los vehículos híbridos eléctricos. Esta variable indica el máximo entre el consumo de combustible tradicional y el equivalente. También se convirtió a kilómetros por litro.
- modelo (`vehicle`): Es una variable categórica que indica el modelo del y que toma 109 valores diferentes.
- clase del vehículo (`carclass`): Es la categoría a la cual pertenece cada vehículo (detalle más adelante).

A pesar de las transformaciones realizadas se decidió mantener el nombre original de las variables. Además los datos cuentan con otras variables: `carid`, `year` y `carclass_id` que son variables que no aportan mucho al análisis de los datos ya que son para el manejo interno de los datos o fungen como etiquetas.

La base de datos se compone de 140 observaciones de diferentes coches y no tenemos datos faltantes ni repetidos. Existen 109 modelos de coches distintos y 7 clases.

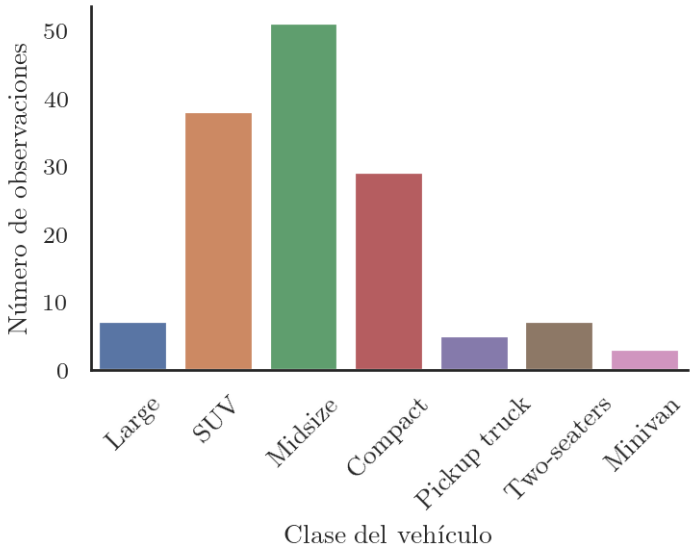

Figura 1. Número de vehículos por clase

De la Figura 1 vemos que hay más carros de tamaño medio, seguido por las SUV y los autos compactos, por lo que la mayoría de los datos de la base son coches familiares donde caben más de

3 personas. Creamos una nueva variable para las clases de coche que agrupa los coches similares en una clase. Los nuevos grupos están dados por:

- Carros grandes: Large y Pickup Truck (`Large`)
- Camionetas: SUV y Minivan (`Vans`)
- Carros pequeños: Compact y Two-seaters (`Small`)
- Carros medianos: Midsize (`Medium`)

En lo subsecuente, la variable `carclass` se refiere a esta nueva asignación de categorías.

De igual manera se creó una nueva variable para el modelo de coche que agrupa los modelos por marca, por ejemplo: *Prius*, *Highlander* y *Camry* son de la marca *Toyota*. De esta manera, la nueva variable `manufacturer` se redujo a 25 clases en lugar de las 103 de la variable `vehicle`. En la Figura 2 vemos que *Toyota*, *Honda*, *Lexus*, *Chevrolet* y *Ford* son las marcas con más observaciones.

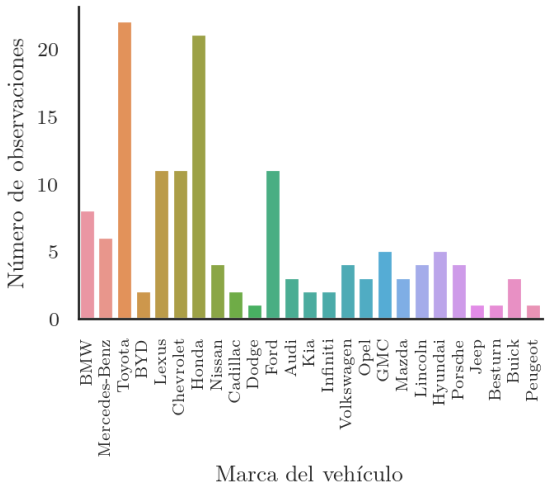

Figura 2. Número de vehículos por marca

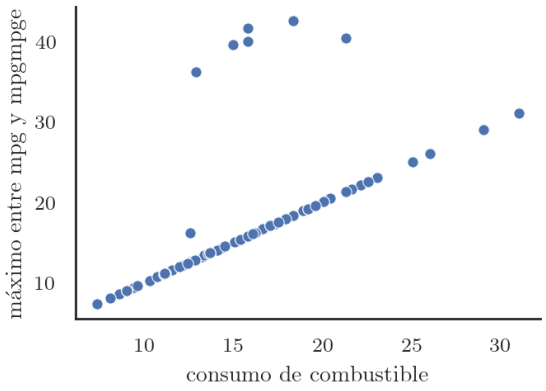

Figura 3. `mpg` vs. `mpgmpge`

Para la mayoría de las observaciones el consumo de combustible es igual al máximo entre el `mpg` y la medición alternativa del consumo de combustible. Sin embargo, hay 11 observaciones en las que la medición alternativa es más del doble de `mpg` (Figura 3).

A continuación evaluaremos la relación lineal entre los regresores y la variable respuesta.

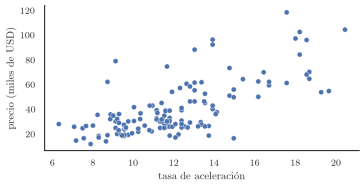

Figura 4. `accelrate` vs `precio`

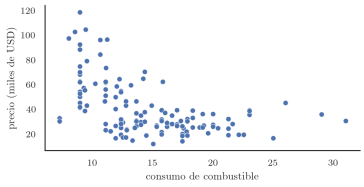

Figura 5. `mpg` vs `precio`

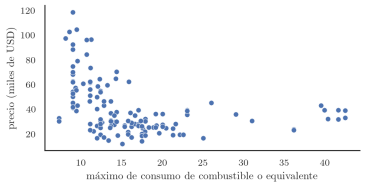

Figura 6. `mpgmpge` vs `precio`

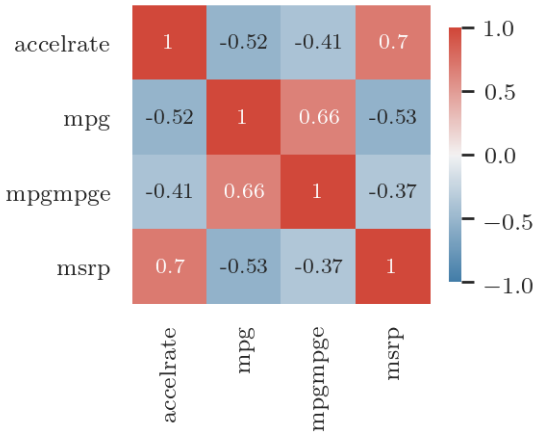

Figura 7. Correlación entre las variables numéricas

La tasa de aceleración es la única que presenta una correlación lineal clara con el precio (Figura 4). Tanto para el consumo de combustible como para el máximo de consumo de combustible o equivalente, la relación con la variable objetivo es no-lineal (Figuras 5 y 6). También podemos ver que las variables `mpg` y `mpgmpge` tienen un comportamiento similar, lo cual tiene sentido ya que solo difieren en 11 observaciones.

Esto se puede confirmar en la Figura 7 ya que `accelrate` tiene una correlación de 0.7 con la variable objetivo, mientras que para `mpg` y `mpgmpge`, además de que las correlaciones son negativas, estas no son muy grandes en magnitud.

4 Modelo de regresión lineal múltiple

4.1 Modelo inicial

En primer lugar, se ajustó un modelo de regresión lineal tomando como respuesta a la variable `msrp` y como regresores a las variables numéricas `accelrate`, `mpg` y `mpgmpge` y a los factores `carclass`, `manufacturer`.

Vemos que `mpg` no es significativo pues tiene un valor-p $0.081 > 0.05$

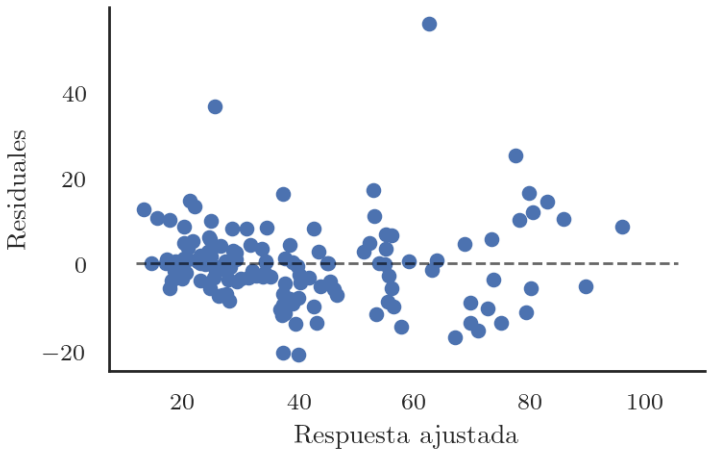

Figura 8. Residuales contra respuesta ajustada

En la Figura 8 se encuentran graficados los residuales contra las respuestas ajustadas del obtenidas a partir del modelo inicial. La gráfica sugiere una desviación notable del supuesto de varianza constante de los términos aleatorios de error.

4.2 Transformación de variables

Con el objetivo de mitigar las desviaciones de los supuestos del modelo de regresión lineal se utilizó la transformación de Box-Cox en la variable respuesta. Para obtener un valor de λ óptimo se consideraron varios valores de prueba. Con cada uno de ellos se obtuvo un ajuste y se seleccionó el valor de λ que minimizara la suma de cuadrados de los residuales de cada uno de los ajustes. En la Tabla 1 se muestra que el valor óptimo resultó ser $\lambda = 0$, lo cual sugiere que la varianza de los términos de error podría estabilizarse al aplicar una transformación logarítmica a la variable respuesta.

λ	$SC_{Res}(\lambda)$	Clasificación
-1.00	10269.7716	
-0.75	8874.2238	
-0.50	7989.9042	
-0.25	7546.4395	
0.00	**7525.9739**	**Óptimo**
0.25	7963.9670	
0.50	8959.1730	
0.75	10695.3245	
1.00	13479.5726	

Tabla 1. Valores de prueba de λ para la transformación de Box-Cox

Para la selección del modelo final se utilizó un procedimiento computacional que consiste en evaluar todas las posibles regresiones. Los parámetros de evaluación fueron el Criterio de Información de Akaike (AIC) y la C_p de Mallows y el valor p asociado con el estadístico F de la significancia de la regresión.

En la Tabla 2 se muestra que el modelo que simultáneamente minimiza el valor del AIC y el valor

model	num_regresors	AIC	Mallows Cp	f_pvalue
mpg	2	156.077	2.0	2.31×10^{-11}
mpgmpge	2	185.046	2.0	5.61×10^{-5}
carclass	4	164.835	4.0	1.23×10^{-8}
accelrate	2	109.445	2.0	1.85×10^{-21}
manufacturer	25	59.143	25.0	1.59×10^{-23}
mpg+carclass	5	152.226	5.0	5.94×10^{-11}
mpg+accelrate	3	100.821	3.0	1.47×10^{-22}
mpg+manufacturer	26	53.568	26.0	3.36×10^{-24}
carclass+accelrate	5	94.008	5.0	6.40×10^{-23}
carclass+manufacturer	28	50.344	28.0	3.20×10^{-24}
accelrate+manufacturer	26	32.315	26.0	9.74×10^{-28}
mpg+mpgmpge+manufacturer	27	48.724	27.0	9.66×10^{-25}
mpgmpge+carclass+manufacturer	29	46.563	29.0	1.41×10^{-24}
carclass+accelrate+manufacturer	29	28.718	29.0	1.84×10^{-27}
mpg+mpgmpge+carclass+manufacturer	30	38.505	30.0	1.33×10^{-25}
mpg+mpgmpge+accelrate+manufacturer	28	29.926	28.0	1.49×10^{-27}
mpgmpge+carclass+accelrate+manufacturer	30	**21.086**	30.0	2.11×10^{-28}

Tabla 2. Estadísticos relevantes para la selección del modelo final

p asociado con el estadístico F de la prueba de la significancia de la regresión es el que se muestra en la última fila de la tabla, es decir, aquel que considera a `mpgmpge`, `carclass`, `accelrate` y `manufacturer`como regresores. En todos los casos la C_p de Mallows es igual a la cantidad de regresores empleados. La cantidad de 30 regresores corresponde a (25 - 1) variables indicadoras asociadas con el factor `manufacturer`, (4 - 1) variables indicadoras asociadas con el factor `carclass`, 2 variables numéricas y 1 regresor asociado con el coeficiente β_0

4.3 Detección de valores atípicos

Para la detección de valores atípicos se calcularon los residuales estandarizados ($\hat{r}_i$) y las distancias de Cook (d_i). Con base en el criterio $|\hat{r}_i| > 3$ se detectaron dos valores atípicos: El vehículo cuyo modelo es `Crown` y el vehículo cuyo modelo es `Lexus LS600h/hL`. Esta situación se encuentra ilustrada en la Figura 9. No se detectaron valore de influencia con el criterio $d_i > 1$.

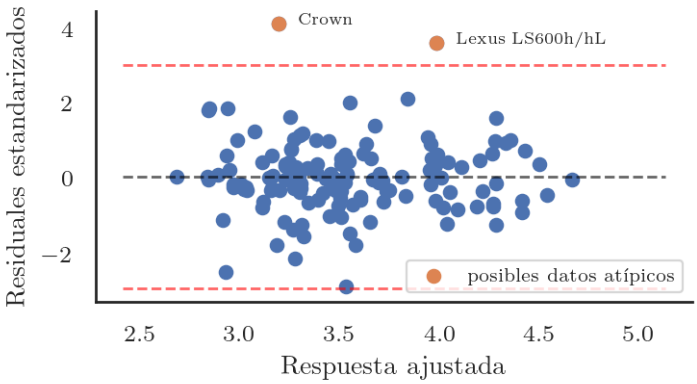

Figura 9. Residuales estandarizados contra respuesta ajustada

msrp	mpg	mpgmpge	carclass	accelrate	manufacturer	vehicle
2.4723	14.9991	14.9991	Medium	7.87	Audi	A5 BSG
2.6801	13.2985	13.2985	Medium	7.14	Besturn	Besturn B50
3.0134	19.5566	19.5566	Medium	9.90	Toyota	Prius
3.1409	12.8181	36.1372	Medium	9.24	BYD	F3DM PHEV
3.1673	12.8011	36.1372	Medium	9.52	BYD	F3DM
3.1864	21.2572	21.2572	Medium	10.20	Toyota	Prius
3.2096	12.3292	12.3292	Medium	9.09	Chevrolet	Malibu
3.2790	15.3052	15.3052	Medium	10.54	Kia	Optima K5
3.3059	14.0000	14.0000	Medium	9.51	Toyota	Prius V
3.4206	31.0015	31.0015	Medium	10.00	Toyota	Prius alpha (V)
3.4657	21.2572	40.3887	Medium	9.17	Toyota	Prius Plug-in
3.4687	21.2572	40.3887	Medium	8.82	Toyota	Prius PHV
4.1318	**15.7984**	**15.7984**	**Medium**	**8.70**	**Toyota**	**Crown**

Tabla 3. Vehículos con `accelrate` entre 6.70 y 10.70 y con `carcalss` == "Medium"

En la Tabla 3 se muestra que el vehículo `Crown` es el único de la clase `Medium` cuyo `accelrate` se encuentra entre 6.70 y 10.70 que tiene un precio en miles de dólares de 2013 `msrp` mayor a 40.

Además de que una situación similar se presenta para el caso del vehículo `Lexus LS600h/hL`, sucede que este es el vehículo más caro de todo el conjunto de datos con un `msrp` de 118.5436 miles de dólares de 2013.

Se ha decidido conservar estas observaciones debido a que no hay evidencia suficiente para concluir que la atipicidad de sus valores pueda ser atribuida a errores incurridos durante la recolección de los datos.

4.3.1 Validación del modelo

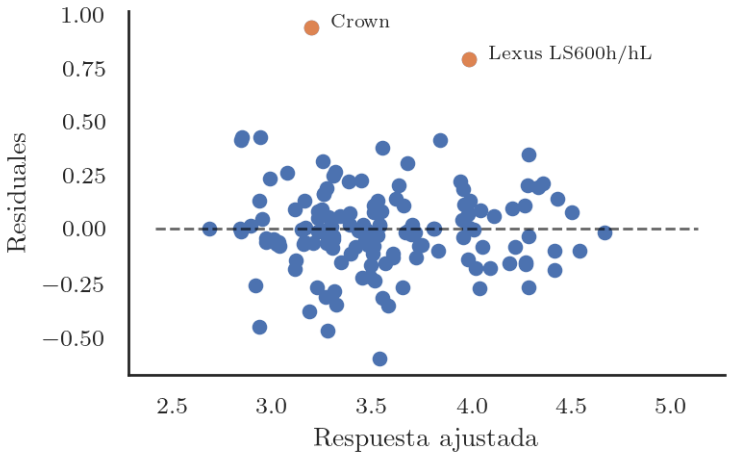

Figura 10. Residuales contra respuesta ajustada

En la Figura 10 se encuentran graficados los residuales contra las respuestas ajustadas obtenidas a partir del modelo final seleccionado. En comparación con la Figura 8, el problema de heteroscedasticidad fue ligeramente mitigado.

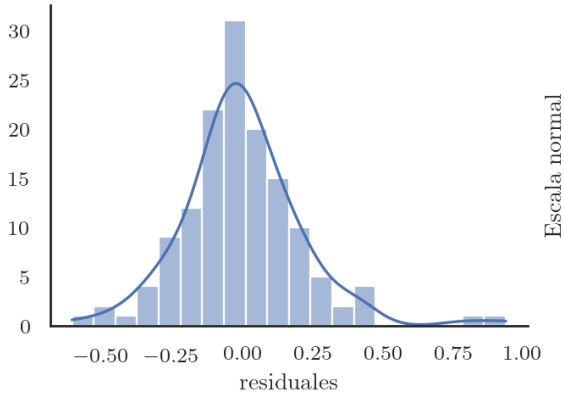

Figura 11. Histograma de los residuales

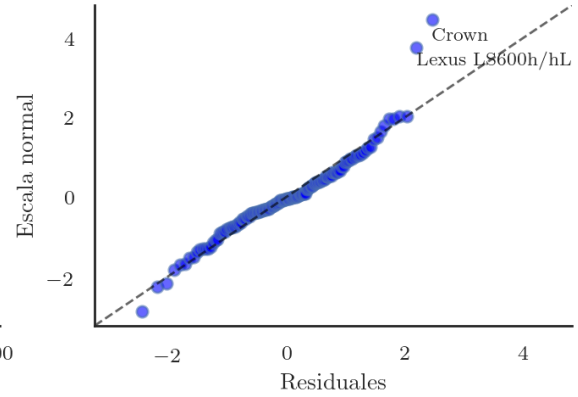

Figura 12. Gráfica cuantil-cuantil

En la Figura 11 se encuentra un histograma de los resiudales que sugiere que, salvo por algunas observaciones, el supuesto de normalidad no está siendo del todo violado. Esto se evidencia con mayor claridad en la Figura 12, que es una gráfica cuantil-cuantil de los residuales contra la escala normal. Salvo por los valores influyentes mencionados previamente, la desviación del supuesto de normalidad no es muy grave.

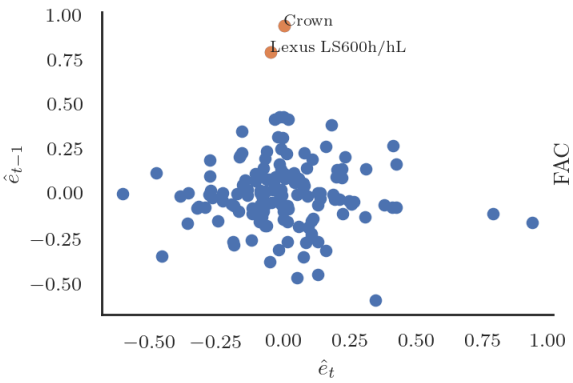

Figura 13. Gráfica $\hat{e}_t$ vs $\hat{e}_{t-1}$

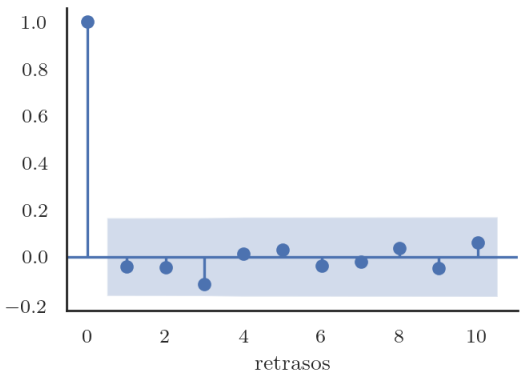

Figura 14. Función de autocorrelación

Finalmente con la Figura 13 y la Figura 14 vemos que los residuales no presentan ninguna correlación aparente entre ellos y que la correlación de orden uno es muy pequeña por lo que también se cumple el supuesto de independencia.

5 Predicción del precio de nuevos vehículos

Recordemos que nuestro modelo está dado por:

$$y = 2.5233 + \begin{pmatrix} -0.2331 \\ -0.2592 \\ 0.0038 \end{pmatrix}' \begin{pmatrix} Medium \\ Small \\ Vans \end{pmatrix} + \begin{pmatrix} 0.7806 \\ -0.0849 \\ -0.1878 \\ 0.2666 \\ 1.0763 \\ 0.3199 \\ 0.3493 \\ 0.0759 \\ 0.5968 \\ -0.0630 \\ 0.0789 \\ 0.3755 \\ -0.6023 \\ 0.1295 \\ 0.5186 \\ 0.3632 \\ -0.1374 \\ 0.8580 \\ 0.2471 \\ -0.1051 \\ 0.5763 \\ 0.6483 \\ 0.2047 \\ 0.3373 \end{pmatrix}' \begin{pmatrix} BMW \\ BYD \\ Besturn \\ Buick \\ Cadillac \\ Chevrolet \\ Dodge \\ Ford \\ GMC \\ Honda \\ Hyundai \\ Infiniti \\ Jeep \\ Kia \\ Lexus \\ Lincoln \\ Mazda \\ Mercedes - Benz \\ Nissan \\ Opel \\ Peugeot \\ Porsche \\ Toyota \\ Volkswagen \end{pmatrix} + 0.0102 * mpgmpge + 0.0620 * accelr \quad (1)$$

Para poder predecir los nuevos datos que tenemos, primero se tienen que hacer las mismas modificaciones que se hicieron con los datos de entrenamiento. Es decir, dividir el `msrp` entre 1000, expresar las columnas `mpg` y `mpgmpge` en términos de kilómetros por litro y clasificar las clases y las marcas de los vehículos como se hizo previamente. Una vez realizados estos cambios, procedemos a hacer la predicción y a analizar los resultados obtenidos.

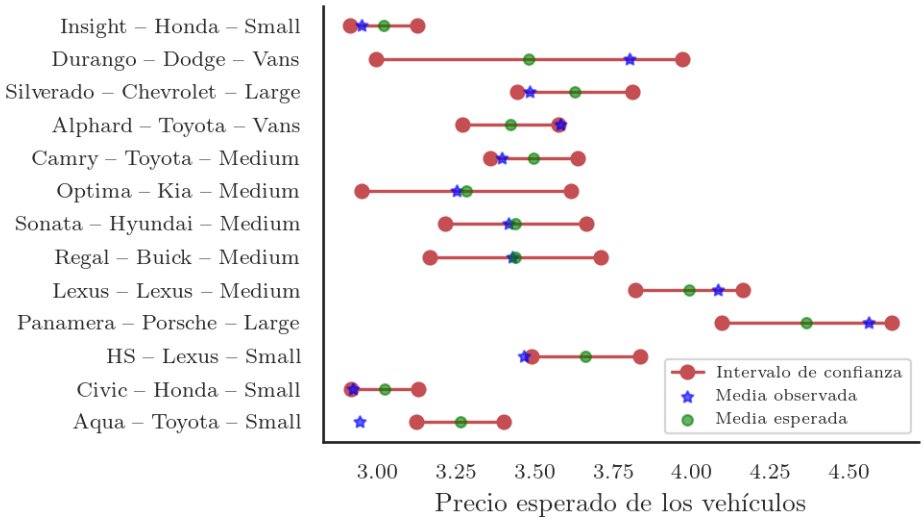

Figura 15. Intervalos de confianza del 95% para los diferentes vehículos

En la Figura 15 se pueden observar los resultados obtenidos con el modelo. En la mayoría de los casos se obtuvo una buena predicción ya que los datos observados están dentro del intervalo de confianza de las predicciones del modelo. Finamlente se calculó la $R^2_{prediccion}$ mediante la fórmula $R^2_{prediccion} = 1 - \frac{PRESS}{SS_T}$ y se obtuvo un valor de 0.9078. Lo que afirma que es una buena predicción de los datos nuevos.

6 Referencias

[1] Lim, D., Jahromi, S., Anderson, T. and Tudorie, A., 2015. *Comparing technological advancement of hybrid electric vehicles (HEV) in different market segments*. Technological Forecasting and Social Change, 97, pp.140-153.