Análisis Regresión Lineal Múltiple

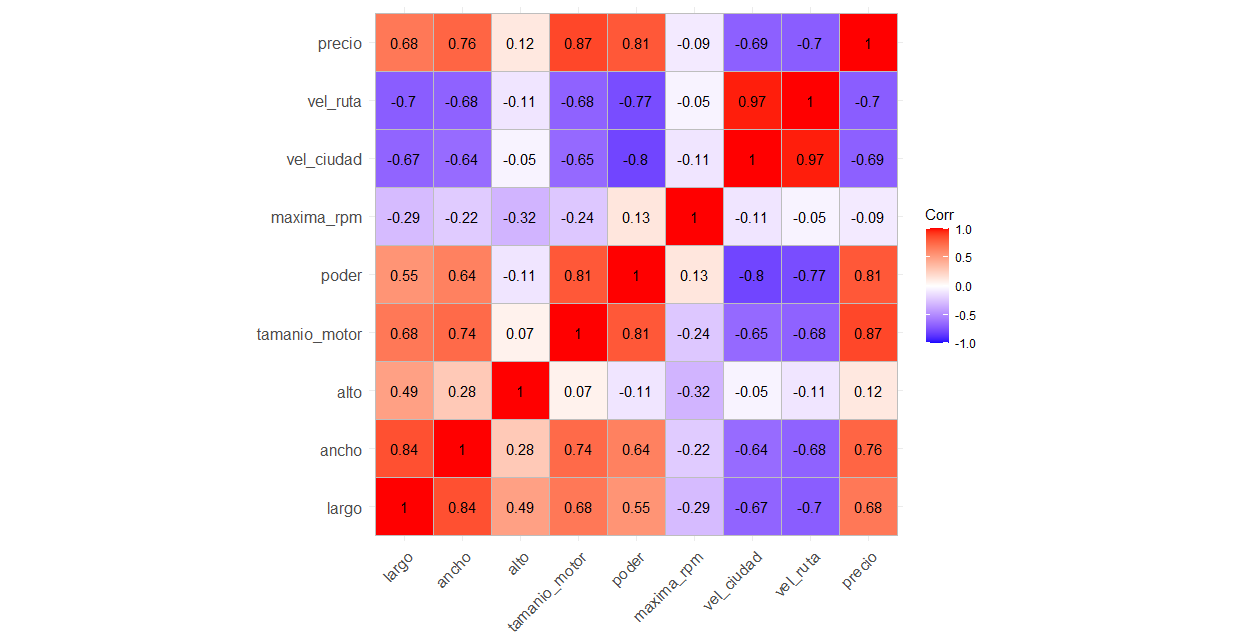
Link a los datos: <https://www.kaggle.com/hellbuoy/car-price-prediction>

El objetivo de este análisis es:

* Identificar variables que son significativas para predecir el valor de los autos.
* Que tan bien esas variables describen el precio de los autos

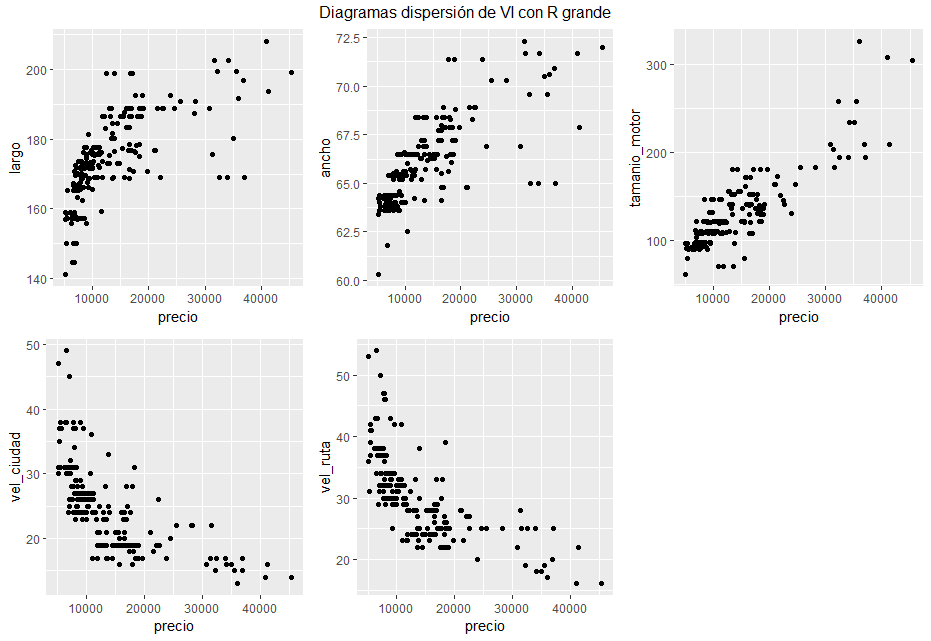
# Análisis de la relación entre variables

## Relación V.D. contra V.I.



* **largo** tiene un coeficiente de correlación **alto**, r = 0.68.
* **ancho** tiene un coeficiente de correlación **alto**, r = 0.76.
* **tamanio\_motor** tiene un coeficiente de correlación **alto**, r = 0.87.
* **poder** tiene un coeficiente de correlación **alto**, r = 0.81.
* **vel\_ciudad** tiene un coeficiente de correlación **alto**, r = -0.69.
* **vel\_ruta** tiene un coeficiente de correlación **alto**, r = -0.7
* **alto** tiene un coeficiente de correlación **bajo**, r = 0.12.
* **máxima\_rpm** tiene un coeficiente de correlación **bajo**, r = -0.09.

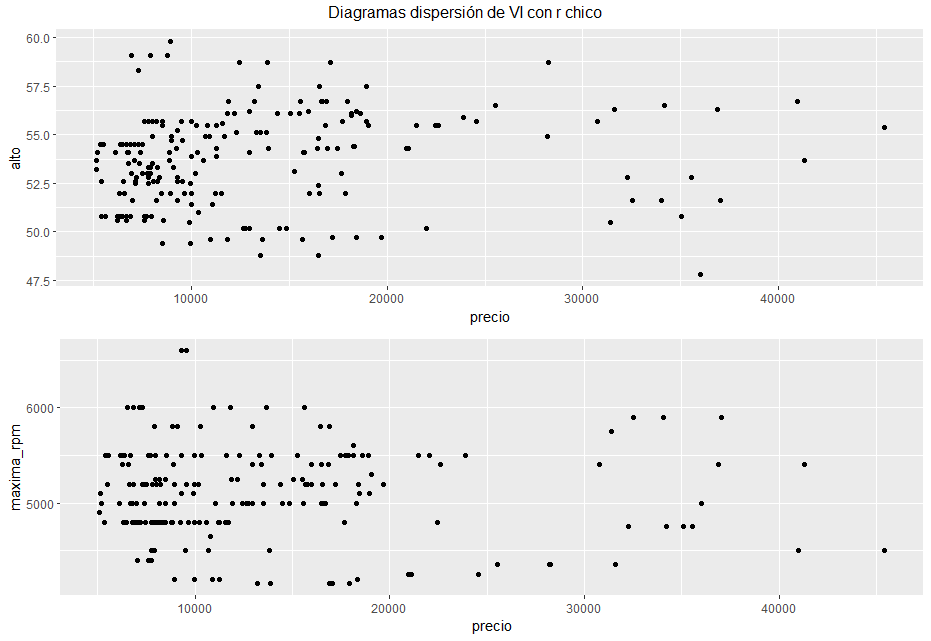
Primero vemos los diagramas de dispersión de precio contra las VI que tienen un r alto



Las VI **ancho** y **tamanio\_motor** pareciera que tienen una relación lineal con precio.

Las VI **largo**, **vel\_ciudad** y **vel\_ruta** pareciera que ajustaría mejor una regresión plinómica debido a la curvatura que se da.

Vemos los diagramas de dispersión con las VI que tienen un r bajo



Como se esperaba no hay una relación clara entre las VI **alto** y **maxima\_rpm** con **precio.**

## Relación V.I. contra V.I.

Usando la matriz de correlación usada en el punto anterior:

* **vel\_ruta** tiene un coeficiente de correlación **alto** con **largo**, **ancho**, **tamanio\_motor**, **poder** y **vel\_ciudad**.
* **vel\_ciudad** tiene un coeficiente de correlación **alto** con **largo**, **ancho**, **tamanio\_motor**, **poder** y **vel\_ruta**.
* **poder** tiene un coeficiente de correlación **alto** con **largo**, **ancho**, **tamanio\_motor**, **vel\_ciudad** y **vel\_ruta**.
* **ancho** tiene un coeficiente de correlación **alto** con **largo**, **tamanio\_motor**, **poder**, **vel\_ciudad** y **vel\_ruta**.
* **largo** tiene un coeficiente de correlación **alto** con **ancho**, **tamanio\_motor**, **poder**, **vel\_ciudad** y **vel\_ruta**.

Por esto hay que considerar de utilizar un subconjunto de variables en lugar de todas.

# Generar un modelo

Lo primero que hacemos es generar un modelo. Aunque ya conocemos posibles problemas por el análisis anterior utilizamos todas las variables, luego filtramos las VI. Los resultados son:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **precio** | | |
| *Predictors* | *Estimates* | *CI* | *p* |
| (Intercept) | -60721.05 | -89722.03 – -31720.06 | **<0.001** |
| largo | -45.62 | -140.54 – 49.31 | 0.344 |
| ancho | 632.19 | 177.20 – 1087.17 | **0.007** |
| alto | 295.00 | 36.59 – 553.40 | **0.025** |
| tamanio\_motor | 115.89 | 89.85 – 141.93 | **<0.001** |
| poder | 38.07 | 7.91 – 68.24 | **0.014** |
| maxima\_rpm | 2.25 | 0.95 – 3.56 | **0.001** |
| vel\_ciudad | -189.65 | -539.35 – 160.05 | 0.286 |
| vel\_ruta | 74.31 | -231.63 – 380.24 | 0.632 |
| tipo\_combustible [gas] | -3645.28 | -5677.26 – -1613.31 | **0.001** |
| Observations | 205 | | |
| R2 / R2 adjusted | 0.837 / 0.829 | | |

Las variables que no tienen un valor estimado significativo son:

* largo
* vel\_ciudad
* vel\_ruta

Aunque estas variables tenían un coeficiente de correlación alto con precio, también presentaban un coeficiente de correlación alto con otras VI, por lo que este pvalor alto puede deberse a un problema de multicolinealidad. Hay que analizar de descartarlas del modelo y dejar las más significativas.

El resto de las variables presentan valores de B significativos, sin embargo hay que analizarlas en profundidad porque algunas no tienen un coeficiente correlación muy grande con el precio.

La interpretación del B de tipo\_combustible para el nivel de gas es: en promedio el precio de los autos con gas es $3645.28 más baratos que los autos con diesel.

El modelo presenta un coeficiente de determinación alto. Su interpretación es que este modelo logra explicar el 83,7% de la variabilidad del precio. A su vez el modelo se considera significativo (estadístico F)

F-statistic: 111 on 9 and 195 DF, p-value: < 2.2e-16

# Selección de los mejores predictores

Para la selección de variables se aplica el método de stepwise. El resultado es

|  |  |  |  |
| --- | --- | --- | --- |
|  | **precio** | | |
| *Predictors* | *Estimates* | *CI* | *p* |
| (Intercept) | -67058.55 | -89950.95 – -44166.16 | **<0.001** |
| ancho | 591.77 | 233.88 – 949.66 | **0.001** |
| alto | 244.30 | 30.67 – 457.92 | **0.025** |
| tamanio\_motor | 110.70 | 85.96 – 135.44 | **<0.001** |
| poder | 50.07 | 24.59 – 75.54 | **<0.001** |
| maxima\_rpm | 2.28 | 0.98 – 3.57 | **0.001** |
| tipo\_combustible [gas] | -3023.57 | -4877.22 – -1169.91 | **0.002** |
| Observations | 205 | | |
| R2 / R2 adjusted | 0.835 / 0.830 | | |

Las estimaciones de todas las variables resultan ser significativas. Pero seguimos teniendo que:

* **máxima\_rpm** aparece pero tiene un coeficiente de correlación con **precio** bajo, r = -0.09.
* **alto** tiene un coeficiente de correlación con **precio** bajo, r = 0.12.
* **poder** tiene un coeficiente de correlación alto con **ancho** y **tamanio\_motor**. Es más preocupante con **tamanio\_motor** ya que presenta un r = 0.81.
* **tamanio\_motor** presenta un alto coeficiente de correlación con **ancho** y **poder**.
* **ancho** tiene un coeficiente de correlación alto con **tamanio\_motor** y **poder**.

El coeficiente de determinación resulta ser parecido al del anterior modelo. También tiene un coeficiente de determinación ajustado parecido.

Calculamos los VIF para poder refinar este modelo

VIF

ancho 2.842209

alto 1.313894

tamanio\_motor 5.118954

poder 4.891878

maxima\_rpm 1.844859

tipo\_combustible 1.465888

Primero sacamos **tamanio\_motor** porque tiene un vif mayor a 5 y vemos como impacta

|  |  |  |  |
| --- | --- | --- | --- |
|  | **precio** | | |
| *Predictors* | *Estimates* | *CI* | *p* |
| (Intercept) | -77880.39 | -104681.10 – -51079.67 | **<0.001** |
| ancho | 1078.07 | 676.63 – 1479.52 | **<0.001** |
| alto | 221.09 | -30.33 – 472.52 | 0.084 |
| poder | 131.79 | 110.89 – 152.69 | **<0.001** |
| maxima\_rpm | -0.55 | -1.89 – 0.78 | 0.415 |
| tipo\_combustible [gas] | -2952.56 | -5134.81 – -770.30 | **0.008** |
| Observations | 205 | | |
| R2 / R2 adjusted | 0.770 / 0.764 | | |

Con este cambio en el modelo vemos que **alto** y **maxima\_rpm** pasan a tener estimaciones no significativas. Esto además concuerda con lo que vimos antes de que ambas variables tienen un coeficiente de correlación bajo con **precio**.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **precio** | | |
| *Predictors* | *Estimates* | *CI* | *p* |
| (Intercept) | -79100.06 | -102695.48 – -55504.65 | **<0.001** |
| ancho | 1252.72 | 886.81 – 1618.63 | **<0.001** |
| poder | 123.95 | 104.38 – 143.51 | **<0.001** |
| tipo\_combustible [gas] | -3426.48 | -5480.38 – -1372.58 | **0.001** |
| Observations | 205 | | |
| R2 / R2 adjusted | 0.765 / 0.761 | | |

Ahora todas las variables resultan ser significativas, vemos como son el vif

VIF

ancho 2.121211

poder 2.060552

tipo\_combustible 1.284902

# Validación de supuestos