Informe de Taller - Regresión lineal

Afanador, S.
Universidad del Valle
Probabilidad y estadística
02 Marzo 2022

Resumen

Este informe busca proponer un modelo sencillo de regresión lineal para hallar la relación entre el cilindraje medido en centímetros cúbicos (CC) de los motores de un dataset de autos con su respectiva potencia medida en caballos de fuerza(HP).

1. Descripción de conjunto de datos

Este conjunto de datos tiene las siguientes las columnas referentes al nombre del auto *name*, el año del modelo del vehículo *year*, el precio de venta del vehículo en USD *selling_price*, La cantidad de kilómetros recorridos por el vehículo *km_driven*, el tipo de vendedor *seller_type*, el tipo de transmisión de vehículo *transmission*, el orden del propietario final del vehículo *owner*, el kilometraje en Km *mileage*, el tamaño del motor en centimetros cúbicos *engine*, la potencia máxima del motor en HP *max_power*, el torque del motor Nm *torque*, y la cantidad de asientos que tiene el auto *seats*.

Este conjunto de datos cuenta con 8128 registros, fue obtenido del sitio web de Kaggle[1] de donde especificamente se usó el archivo *Car details v3.csv*.

2. Limpieza de datos

Se inicia limpiando los datos de inconsistencias y anormalidades, como primero paso se eliminan todos las filas que tengan algún dato faltante en al menos una de sus columnas. Seguidamente se elimina el texto de los valores numéricos, como es el caso de la columna *max_power* que tiene siempre las unidades *bhp* y el caso de la columna *engine* que para todos sus valores tiene las unidades *CC*.

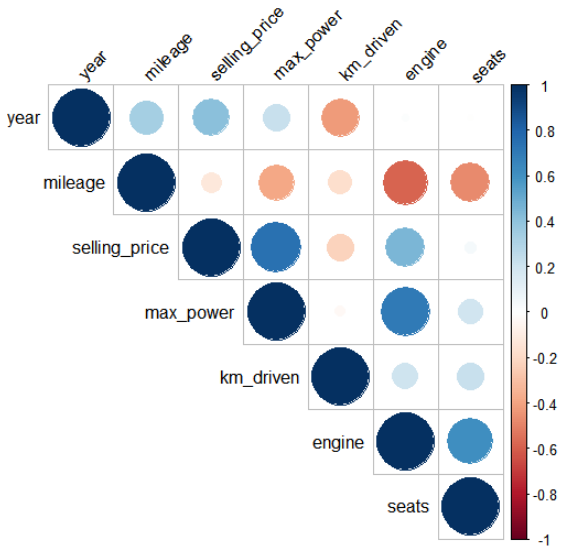
Después eliminan las columnas que no se usarán o que tienen datos de tipo *String* como es el caso de *name*, *seller_type*, *transmission*, *owner*, *torque*.

[1]Dataset: `https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho`

3. Criterio para elección de las variables

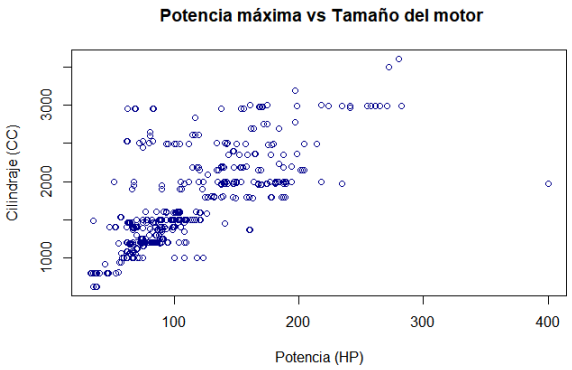
Una vez que los datos están limpios de anormalidades y se tienen las variables justas con cantidades numéricas se procede a crear una matriz de correlación para conocer cuales de las variables tienen mejor relación entre si.

Se obtienen los siguientes resultados usando la librería *corrplot* donde se puede evidenciar uno de los pares de variables que mejor correlación tienen son *engine* y la variable *max_power* con un valor de 0,743 relacionando de manera directa el tamaño del motor con la potencia máxima que puede producir.



4. Creación del modelo

Después de escoger las variables *engine* y *max_power* como las candidatas para ser analizadas se procede a visualizar gráficamente los datos de ambas variables para evidenciar el comportamiento que tienen entre sí.

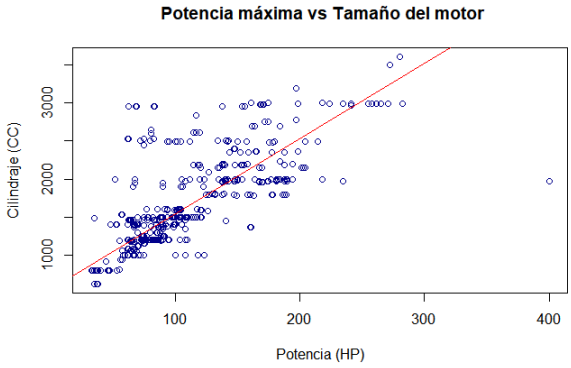


5. Parámetros del modelo de regresión

Para crear el modelo se uso la función *lm* del paquete estandar de R para hallar los valores de intercepto β_0 y pendiente de la función lineal β_1 que representa el modelo que mas se ajusta al valor esperado del Cilindraje de un motor de auto dada su potencia en caballos de fuerza.

$$y = 549{,}8 + 9{,}92x$$

Teniendo en cuenta los valores anteriores, la gráfica del modelo junto con los datos quedaría de la siguiente asi:



6. Código fuente

Acontinuación el código usado para calcular los valores del modelo simple de regresión y la creación de los gráficos incluidos en el informe.

```
#install.packages("corrplot")
library(corrplot)


res <- cor(carDetails)
round(res, 4)
corrplot(res, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
abline(549.8, 9.92,col="red")
plot(
  main= " Potencia  mxima  vs  Tamao  del  motor",
  xlab="Potencia (HP)",
  ylab="Cilindraje (CC)",
  col = "dark blue",
  carDetails$max_power,carDetails$engine)

model = lm(carDetails$engine ~ carDetails$max_power)
model
```

7. Agradecimientos

Al profesor Wilmar Sepulveda Herrera porque sin su ayuda, paciencia y conocimiento no habría sido posible este trabajo.