Regression Analysis for Prediction of Used Car Prices:

Analyze the effect of vehicle components on the price

Name: Seojin Yoon
Student ID: A67029

Time Series Data Analysis and Forecasting
Professor Myung Suk Kim

Due date: November 06, 22

1. Problem Motivation

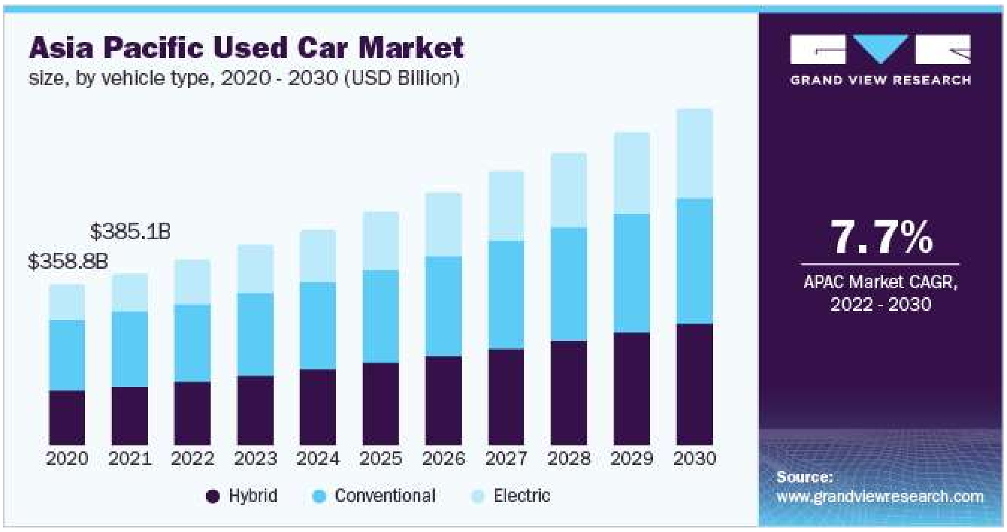


Figure 1 Used car market trend report[1]

In modern society, cars are one of the most important things. Recently, it takes a long time to acquire a new car due to a lack of semiconductor supply, and as a result, more and more people are purchasing used cars.[2] In the case of used cars, it is difficult to determine the exact price due to various factors. As used-car fraud increased,[3] the need for consumers to know the exact price. If it is analyzed through U.S. vehicle data with a large number of data and is meaningful, it can lead to analysis through domestic data. In the growing used car market, the analysis of the used car market was selected because it was an interesting and important topic as it could develop strong competitiveness and lead to trust through objective indicators.

2. Research Objectives

Used car transactions have many variables, so the price fluctuates a lot. A method that all stakeholders can be satisfied by presenting a price model with information guidance to sellers, inspectors, intermediate relationships, and buyers was studied. This report aims to create a model that explains the price through various used-car elements and to know the exact price of

[1] Used Car Market Size, Share & Trends Analysis Report By Vehicle Type (Hybrid, Conventional, Electric), By Vendor Type, By Fuel Type, By Size, By Sales Channel, By Region, And Segment Forecasts, 2022 – 2030, https://www.grandviewresearch.com/industry-analysis/used-car-market

[2] https://www.usatoday.com/story/money/cars/2022/02/13/used-cars-cost-more/6778705001/

[3] https://www.consumeraffairs.com/news/scammers-are-taking-advantage-of-record-high-used-car-prices-071822.html

a used car when the consumer enters the elements he thinks. In addition, the purpose is to derive a correlation between price and each factor.

3. Literature Review

Previous studies have shown a very low correlation between quality and price, and market prices show product reliability.[4] However, according to a recent study, there is also a result that the year, mileage, and displacement of cars affect prices, especially the year and mileage. Linear regression was used, and price, accident status, mileage, and displacement were applied as the main variables. It is a paper that conducted statistical analysis through p-value using a small number of variables.[5]

4. Data and Method Explanation

(1) The Dependent Variable (y)

The dependent variable is *Price*. Since the *Price* is finally determined by various variables, it was selected as a dependent variable. Since information on the vehicle's mileage, displacement, and the manufacturer is determined by the customer, and the *Price* is recognized accordingly, it is reasonable to set the *Price* as a dependent variable.

(2) The Independent Variables (x)

The data used in the analysis include factors such as Levy, Manufacturer, Model, Prod. year, Category, Leather interior, Fuel type, Engine volume, Mileage, Cylinders, Gearbox type, Drive wheels, Doors, Wheel, Color, Airbags, Turbo. Only dependent variables and variables with meaningful results will be used, and the results will be described.

The finally selected independent variables are as follows.

- Levy
  It was judged that the *Levy* imposed according to the type of vehicle would be related to the final price.

- Prod..year
  It is an indicator of the year the car was produced. In the case of very old vehicles, prices tend to rise, but in general, the more recently produced vehicles, the higher the price.

- Mileage

[4] Ginter, James L.; Young, Murray A.; Dickson, Peter R. *A market Efficiency Study of Used Car Reliability and Prices*, Journal of Consumer Affairs. Winter87, 1987.

[5] 정재현, 김민승, 김종민. (2022). 중고차 가격 예측을 위한 영향요인 분석. 한국정보통신학회 종합학술대회 논문집, 26(1), 694-696.

It can be seen as an intuitive indicator of how much cars were used, and it is an important indicator when looking at used products.

- Engine.volume
  The larger the *Engine.volume*, the larger the vehicle and the better the performance, so it is judged to affect the price.

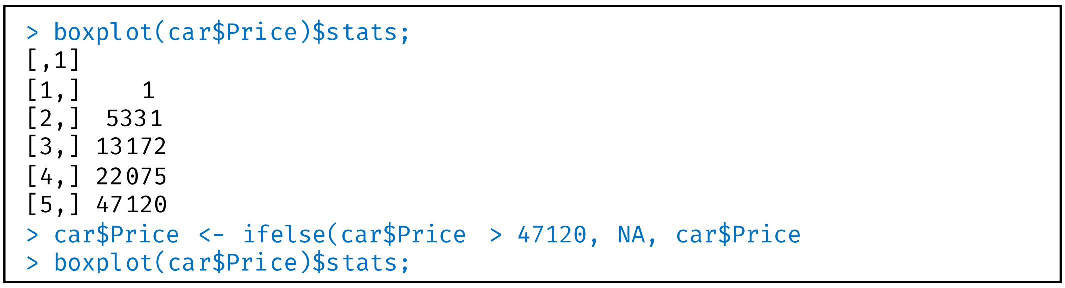
- Cylinders
  Like the *engine.volume*, the *cylinder* is known to improve the performance of the vehicle as it increases.

- Fuel.type
  It was judged that there would be a final price difference because the maintenance cost differed and the difficulty of management varied depending on the type of fuel.
  Use one-hot encoding. The final choices for this variable are Diesel, Hybrid, and LPG. Non-mainstream substances such as hydrogen were thought to be disturbing. I removed it because the p-value is very high.

- Gear.box.type
  This indicator, which is a very important factor in the way of operation, was adopted because it would affect the price.
  Use one-hot encoding. The final choices for this variable are Automatic and Tiptronic.

5. Analysis Result

Because there are various cars, it eliminated outliers from the price. The outliers were identified and removed through the boxplot. Remove any figures that deviate significantly from the graph. Replace the value changed to the missing value with the median value.



```
> boxplot(car$Price)$stats;
[,1]
[1,]     1
[2,]  5331
[3,] 13172
[4,] 22075
[5,] 47120
> car$Price <- ifelse(car$Price  > 47120, NA, car$Price
> boxplot(car$Price)$stats;
```

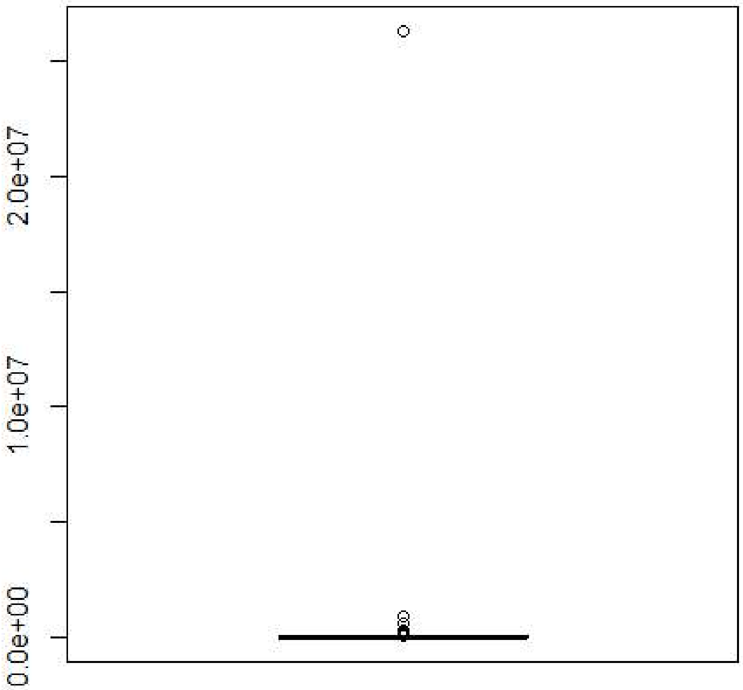


Figure 2 Dependent variable *Price* boxplot

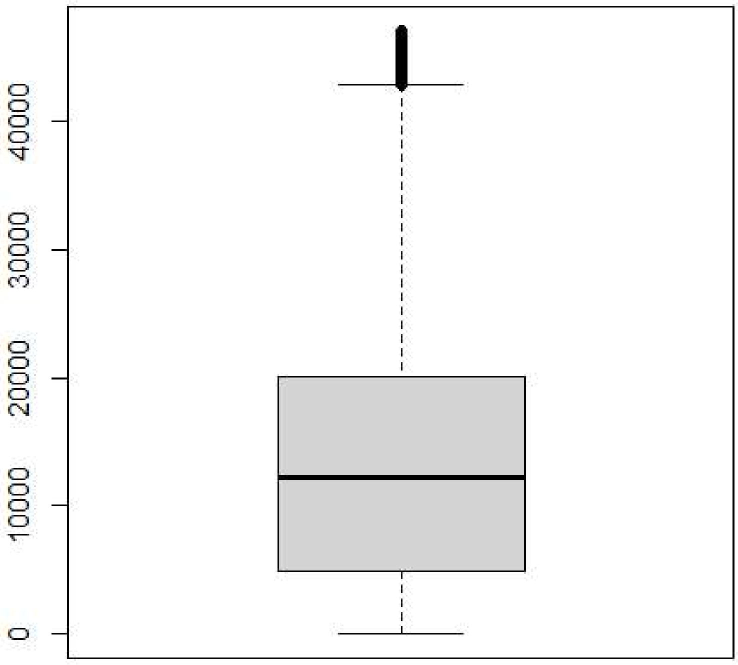


Figure 3 Dependent variable *Price* boxplot without outliers

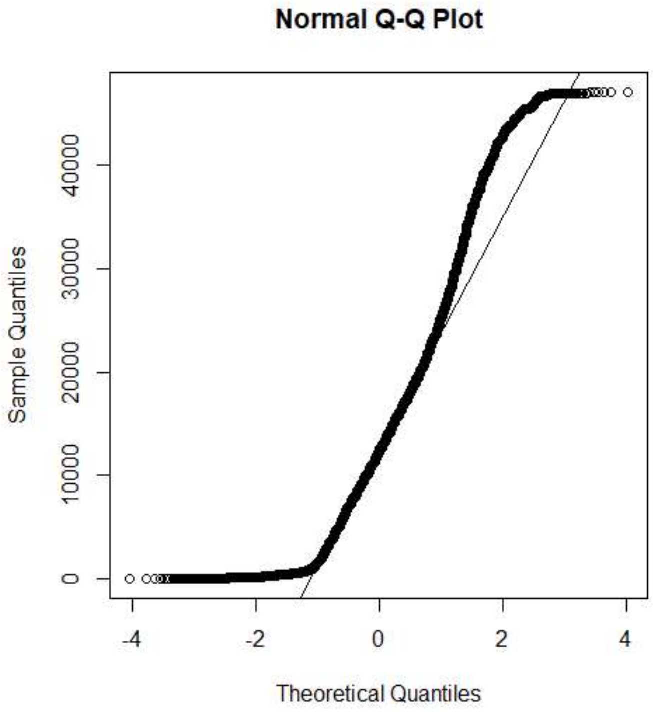
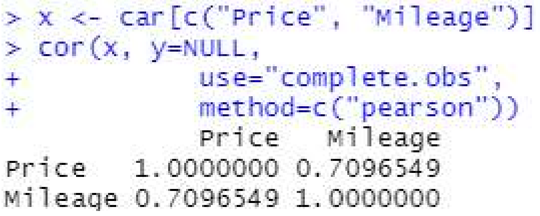


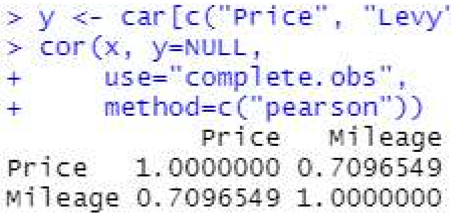
Figure 4 Dependent variable *Price* qqplot without outliers

After removing the outliers, the dependent variable has the right longtail on the qqplot, but it is found that it follows the shape of the normal distribution well.



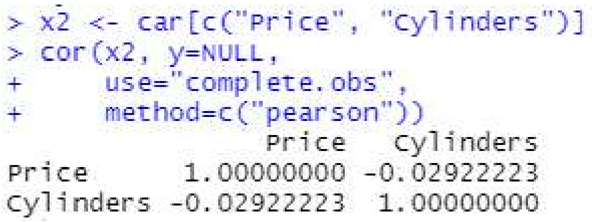
```
> x <- car[c("Price", "Mileage")]
> cor(x, y=NULL,
+            use="complete.obs",
+            method=c("pearson"))
            Price   Mileage
Price   1.0000000 0.7096549
Mileage 0.7096549 1.0000000
```

Figure 5 *Price* and *Mileage* Pearson correlation coefficient



```
> y <- car[c("Price", "Levy")]
> cor(x, y=NULL,
+     use="complete.obs",
+     method=c("pearson"))
            Price   Mileage
Price   1.0000000 0.7096549
Mileage 0.7096549 1.0000000
```

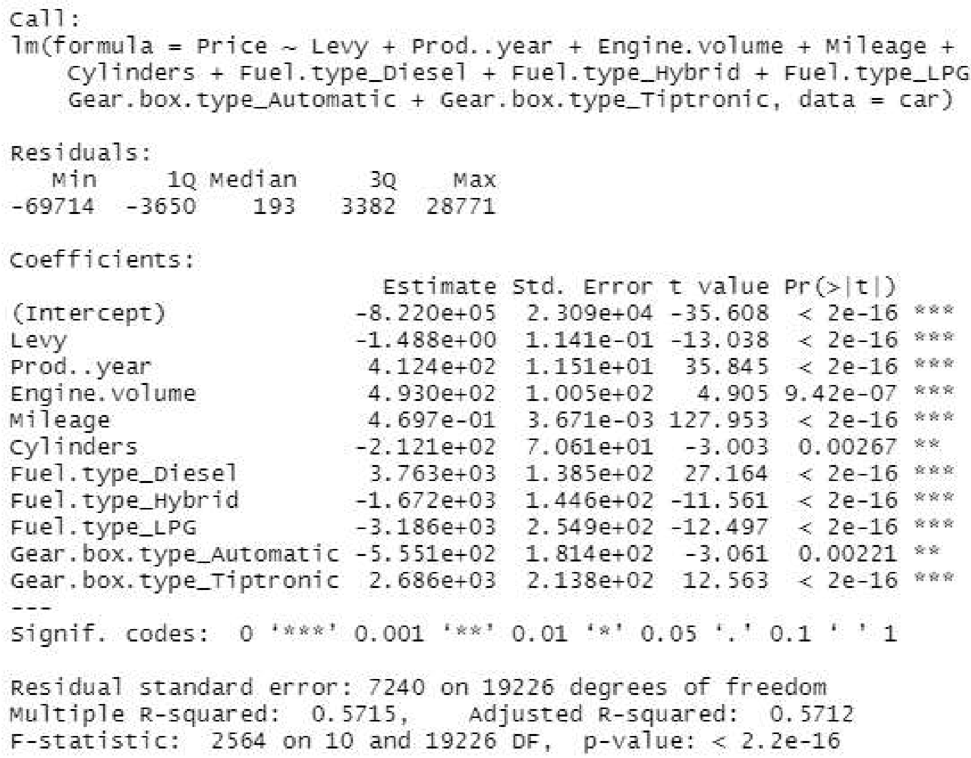
Figure 6 *Price* and *Levy* Pearson correlation coefficient



```
> x2 <- car[c("Price", "Cylinders")]
> cor(x2, y=NULL,
+     use="complete.obs",
+     method=c("pearson"))
                Price   Cylinders
Price      1.00000000 -0.02922223
Cylinders -0.02922223  1.00000000
```

Figure 7 *Price* and *Cylinders* Pearson correlation coefficient

The mileage showed a strong positive correlation, and Levy and Cylinders showed a weak negative correlation. Regression analysis was conducted through the finally selected variables.



```
Call:
lm(formula = Price ~ Levy + Prod..year + Engine.volume + Mileage + 
    Cylinders + Fuel.type_Diesel + Fuel.type_Hybrid + Fuel.type_LPG + 
    Gear.box.type_Automatic + Gear.box.type_Tiptronic, data = car)

Residuals:
   Min     1Q Median     3Q    Max 
-69714  -3650    193   3382  28771 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             -8.220e+05  2.309e+04 -35.608  < 2e-16 ***
Levy                    -1.488e+00  1.141e-01 -13.038  < 2e-16 ***
Prod..year               4.124e+02  1.151e+01  35.845  < 2e-16 ***
Engine.volume            4.930e+02  1.005e+02   4.905 9.42e-07 ***
Mileage                  4.697e-01  3.671e-03 127.953  < 2e-16 ***
Cylinders               -2.121e+02  7.061e+01  -3.003  0.00267 ** 
Fuel.type_Diesel         3.763e+03  1.385e+02  27.164  < 2e-16 ***
Fuel.type_Hybrid        -1.672e+03  1.446e+02 -11.561  < 2e-16 ***
Fuel.type_LPG           -3.186e+03  2.549e+02 -12.497  < 2e-16 ***
Gear.box.type_Automatic -5.551e+02  1.814e+02  -3.061  0.00221 ** 
Gear.box.type_Tiptronic  2.686e+03  2.138e+02  12.563  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7240 on 19226 degrees of freedom
Multiple R-squared:  0.5715,	Adjusted R-squared:  0.5712 
F-statistic:  2564 on 10 and 19226 DF,  p-value: < 2.2e-16
```

Figure 8 Regression analysis after preprocessing

When the significance level is selected as 1%, the null hypothesis is rejected. In addition, the Adjusted R-squared is 0.5712, which is considered to explain *Price* to some extent when the model is made through these variables.

With all this carried test, the final model can be written as (descended coefficients):

Used Car Price

$= -8022200 - 1.488 * (Levy) + 412.5 * (Prod\ Year) + 492.8 *$

(Engine volume) + 0.4697 ∗ (Mileage) – 212.21 ∗ (Cylinders) + 3750 ∗ (Fuel type: Diesel) - 1685 ∗ (Fuel type: Hybrid) - 3199 ∗ (Fuel type: LPG) − 555.1 ∗ (Gear box type: Automatic) + 2686 ∗ (Gear box type: Tiptronic)

6. Discussions

Since it is data on the US market, it was recognized and analyzed that the factors that consumers view important compared to the Korean market were different. However, it was surprising that mileage showed a positive correlation. Of course, I think the difference in price according to the type of car and the difference according to the manufacturer had an effect to some extent, but the moment the outlier was changed to the median value, it had a strong positive correlation, resulting in a different result than expected. Since it contains many categorical data, it has had a lot of difficulties in processing the data. Better results can be obtained if other derivatives are newly created and processed or detailed.

7. Reference List

- Used Car Market Size, Share & Trends Analysis Report By Vehicle Type (Hybrid, Conventional, Electric), By Vendor Type, By Fuel Type, By Size, By Sales Channel, By Region, And Segment Forecasts, 2022 – 2030, https://www.grandviewresearch.com/industry-analysis/used-car-market

- “Used cars cost 40.5% more than last year as gas prices rise. New car prices also climbing”. USA TODAY, Feb 13, 2022, https://www.usatoday.com/story/money/cars/2022/02/13/used-cars-cost-more/6778705001/

- “Scammers are taking advantage of record-high used car prices”, CounsumerAffairs, July 18, 2022, https://www.consumeraffairs.com/news/scammers-are-taking-advantage-of-record-high-used-car-prices-071822.html

- Ginter, James L.; Young, Murray A.; Dickson, Peter R. *A market Efficiency Study of Used Car Reliability and Prices*, Journal of Consumer Affairs. Winter87, 1987,

- 정재현, 김민승, 김종민. (2022). 중고차 가격 예측을 위한 영향요인 분석. 한국정보통신학회 종합학술대회 논문집, 26(1), 694-696