MODS203 Project : Price determinants of used cars

Abdelali Boukhari
Estienne Goigoux
Sami Contesenne

I- Introduction

Buying a used car can be very difficult because there are many things to consider, and the price depends on many variables, it depends on the car's brand and model, on the features it has, on the number of kilometers the car has traveled, and many other factors that will be detailed later in this report. Not to mention the buyer also needs to have a certain degree of "trust" in the seller in order to buy a car that is more expensive than other cars with similar features, otherwise, when faced with a group of cars with the same properties most buyers will go for the one with the lowest price. So, our project's purpose is to better understand the used car's market by analyzing cars and their prices on the "Autoscout24" website.

Our analysis is mainly for the purpose of answering the following questions: Which feature (or features) have the most influence on the price of the car? Does the seller's rating also influence said price? We believe that these questions are highly significant because they provide an overview of what makes a car valuable and expensive in the eyes of the sellers, but also how accurate that evaluation is in the eyes of the buyers whose contribution is conveyed via the ratings they give to the sellers. This may also help future buyers to decide which features to take into consideration when purchasing a used car, specially from the "autoscout24" website.

All the data for our analysis comes from the "autoscout24" website, using web scraping we were able to gather the information we needed to start our analysis, we focused on popular car brands such as: BMW, Audi, Volkswagen, Mercedes, Opel… in total we extracted information about 20 different car brands, these information include : price, number of kms traveled, if the car is used or new, type of fuel, rating of seller, number of reviews, and the we proceeded to check if the cars had some features that we fixed beforehand such as "air conditioning", "light detector"… for a total of 16 features. However, to simplify our analysis we chose to restrict ourselves to one country, France.

In the analysis, we mainly used linear regression to see to what extent the price depends on each of the variables previously mentioned. We also had to use a log-transformation to reduce skewness of our original data. We also were able to find some dependencies between some of the variables, as presented in our notebook.

According to our analysis, the price of each car did indeed depend on all the variables. We found out that the more features the car has, the higher the price which is only natural, the regression also pointed out that the more the car has traveled, the less expensive it becomes. Another thing we discovered is that the better the rating a seller has, the more expensive his car is, which confirms what was mentioned about the "trust" factor in the introduction, that can result from the buyer having all the information needed, including the quality of the car, to make a decision and give a rating to a specific seller, and this particular feat represents the buyers' implicit input in the analysis. More details of the analysis are provided later in the report.

II- Background

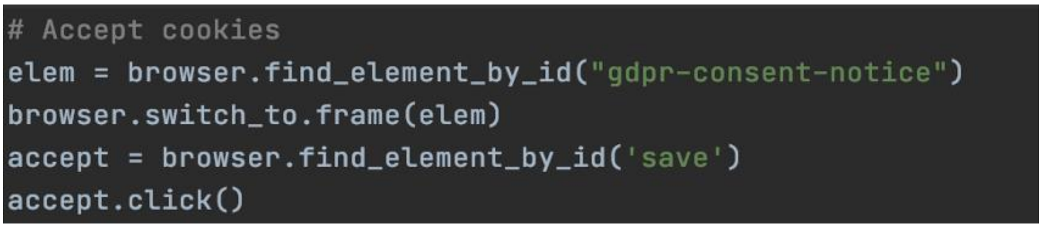
The literature we took some inspiration from during our analysis was the paper written by George Akerlof, "*The Market for 'Lemons': Quality Uncertainty and the Market Mechanism*". The market for used cars is used as the illustrative example for a market in which there are information imbalances. Cars are a product where their overall quality matters a great deal to a consumer, and it's not very easy to tell what exactly that quality is from just looking at it. The seller of the car knows exactly the quality of the car, if it's good they will tell you as such and hope you take them at their word. But if it's bad you have no real way of figuring that out and the seller has just as much incentive to convince you that it is good. What actually happens in this market for used cars after these perverse incentives to lie are used is that consumers eventually realize that while they can't actually know exactly the state of each car the general reputation of used cars has shrunken, say if consumers value a good car at $2000 and a "lemon" (a bad car) at $1000, and the market is a 50/50 split between good and bad cars, consumers would pay the expected value of $1500 for used cars. Now, say that the people who own good cars wouldn't be able to part with them at that price, only those with bad cars would be willing to do so, the bad car sellers would partake in what's called *adverse selection* and try to take advantage of this arbitrage opportunity, but consumers will catch on, and in this case no cars would sell.

In order to have a better understanding of the used car markets in France, we will mainly use the report titled: "**France Used Car Market Outlook to 2023 – Surge in Demand for Rental Cars Backed by Increased Online Used Car Sales".** The research offers a thorough examination of the French used car market and focuses on the overall market size in terms of used cars sales volume. When it comes to sales volume, the used car market in France is at a late stage of development in 2018. Factors like greater online dealerships, as well as the availability of insurance services, make it quick and straightforward for customers and give a pleasant experience for used car buyers. The shift from a traditional sales channel to an E-commerce channel, as well as an increase in demand for younger used cars, have all contributed to the used car industry's growth in terms of sales volume in France. It is stated in the report that the used car sales through the online channel will increase from 11% in 2018 to 15.5% in 2023. Therefore, increased sales through the digital portals will elevate the France Used car market by 2023. And according to this report, Sedan is the most preferred type of car segment in the French Used Car Market.

For our project, we extracted data from the "autoscout24" website, as it contained thousands of car adverts, and we felt that it would be quite sufficient for our study. The platform is similar to all other online platforms for online purchase, the user could type in a car brand, and then a specific model for the brand, as well as the city and other search options, but we conducted the study in the French market, so we only varied the brands and models during our web scraping. The user could then, for each car, see the type of fuel, the kms traveled, ratings… and other features. There was also a box to indicate whether the price suggested by the seller was a good deal or if it was too high, however this information was missing from most cars so we have decided against adding it to our dataset.

III- Data collection strategy

We aren't aware of any APIs that can be used to extract data from this website, so we used web scraping to gather all the data presented in the dataset turned in with the report. The access to the website wasn't blocked by some sort of human verification or "captcha" however we did find it highly difficult at first to extract data since the webpage couldn't be used unless the user accepts (or rejects) "cookies". This was a difficulty because the cookies page was in a different frame, so it wasn't possible to interact with the main webpage using selenium, however we were able to find a solution by simply switching frames as shown below :



```
# Accept cookies
elem = browser.find_element_by_id("gdpr-consent-notice")
browser.switch_to.frame(elem)
accept = browser.find_element_by_id('save')
accept.click()
```

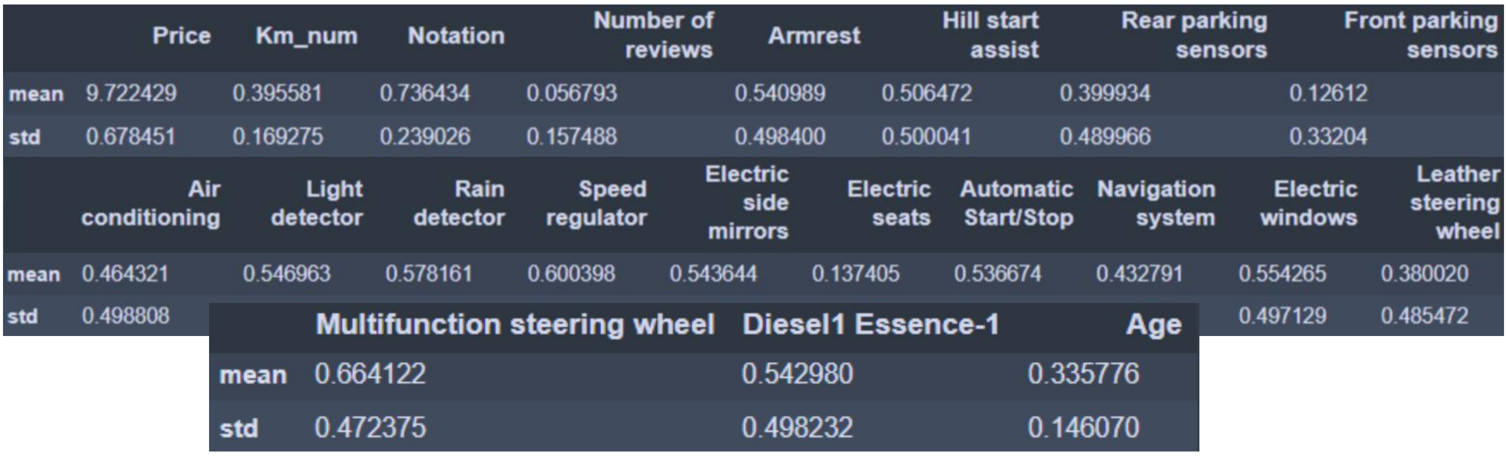
We used selenium in order to interact with the webpage and type in the search bar, click on buttons, as well as to navigate between the pages of the search results to collect links to each car's webpage. As for parsing the data from the car's webpage we used Beautiful Soup, to find elements and add them to our csv file, to finally complete our final dataset. We did use pandas and other libraries in order to create the csv file, or clean the dataset, or handle errors, but the main part of the web scraping was done using a combination of Selenium and Beautiful Soup.

In our first assignment we mentioned that we weren't able to extract car ratings, so we dropped the question about the influence of car ratings on prices, but we succeeded in doing so, and our final dataset now contains this information that we have included in our analysis. In order to decide which information to extract about the cars, we ran random searches and compared information available in each car. We didn't filter our search because we wanted the analysis to be as complete as possible, our final dataset had 24 columns including the one for the names, from which we extracted the brand and model. These 24 columns included: Name, Price, Kms traveled, State, Date of the ad, Fuel, Rating, Number of reviews, the remaining columns had a True\False value to indicate whether the following specific set of features were present in the car : 'Armrest', 'Hill start assist', 'Rear parking sensors', 'Front parking sensors', 'Air conditioning', 'Automatic climate control', 'Light sensor', 'Rain sensor', 'Cruise control', 'Electric side mirrors', 'Electric seats', 'Automatic start/stop', 'Navigation system', 'Electric windows', 'Leather steering wheel', 'Multifunction steering wheel', meaning one observational unit was one car along with these 24 properties.

While we also said that we were would like to have around 100000 entries in our final dataset, we found an issue with the website where for each search result, the website doesn’t let us go over 20 pages, which gives at maximum 400 results per search even though it mentions that there are more than 1000 results available. And since we didn’t want to include brands that weren’t very well-known, we collected results from 20 different search results, which should in theory give about 8000 entries in the dataset, but after the preprocessing and data cleaning we were left with less than that. Another issue we had was that the final code for the data scraping takes a long time to run (the code was left running overnight). As for the final dataset it was extracted during the night of the 6th of January, 2022. We found having 7000 entries in our dataset was sufficient to conduct the analysis since all features of the cars were varied and seemed representative of the used cars market in France.

Concerning data preparation and cleaning, we decided to drop all lines with missing values since half of them had the notation criteria missing, which is a feature we really wanted to interpret in our analysis and since we still had a good number of observations. Then, we translated column labels to english, converted the number of kilometers and the price from a string to a float. We created a feature ‘Age’ from the date of creation of the cars and extracted the brands and models from the name of the car. After that, we began encoding categorical features : we replaced booleans True and False by 1 and 0, translated the ‘Diesel’ and ‘Essence’ types into 1 and -1, and one hot encoded the state of the car. Then, we normalized all the data on a scale from 0 to 1 so that the coefficients in the linear regression could be approximately of the same order.

IV- Data description



	Price	Km_num	Notation	Number of reviews	Armrest	Hill start assist	Rear parking sensors	Front parking sensors
mean	9.722429	0.395581	0.736434	0.056793	0.540989	0.506472	0.399934	0.12612
std	0.678451	0.169275	0.239026	0.157488	0.498400	0.500041	0.489966	0.33204

	Air conditioning	Light detector	Rain detector	Speed regulator	Electric side mirrors	Electric seats	Automatic Start/Stop	Navigation system	Electric windows	Leather steering wheel
mean	0.464321	0.546963	0.578161	0.600398	0.543644	0.137405	0.536674	0.432791	0.554265	0.380020
std	0.498808								0.497129	0.485472

	Multifunction steering wheel	Diesel1 Essence-1	Age
mean	0.664122	0.542980	0.335776
std	0.472375	0.498232	0.146070

Figure : Descriptive statistics of the dataset

We ended up with 3013 observations and 20 explicative features. One observation has 3 separate parts inside the line : the Price, which is the number on which we base our analysis, the brand and model, which will avail us to conduct precise analysis on the same type of car and the other features which are the variables affecting the price in the linear regression.

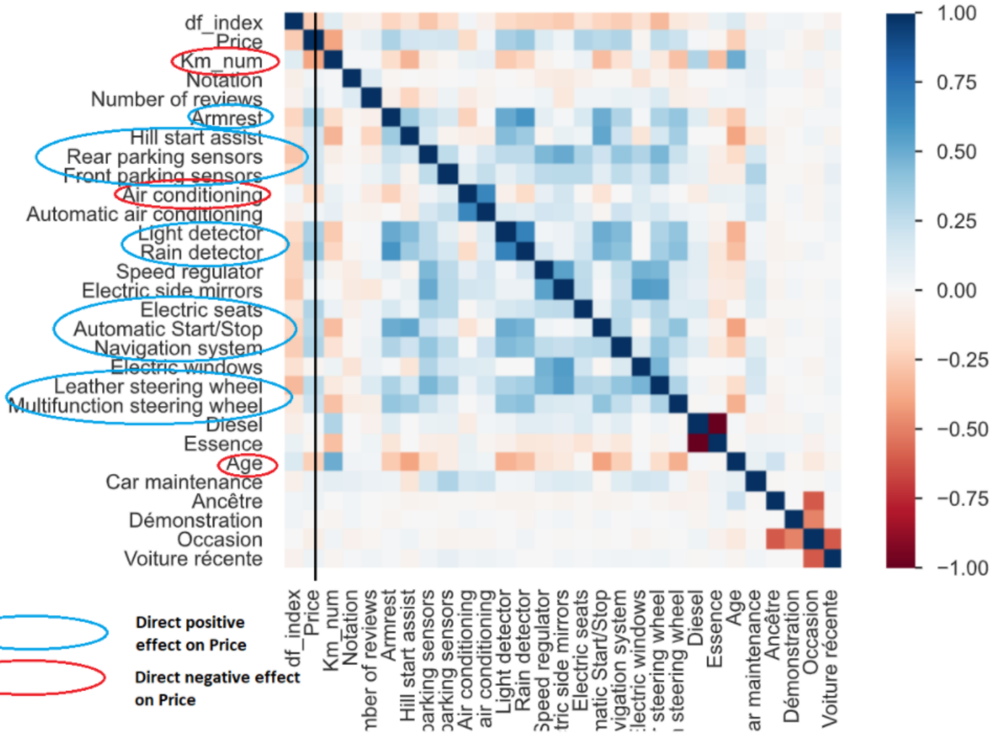
The most important features are the Price (the target), the number of km, the age and the notation and number of reviews. All the other variables are essentially options that you can have on a car and which will help us to have a good regression for the analysis.

All the variables have an average around 0.4-0.6 which is pretty normal except the notation which is pretty high (0.74). It is because a large majority of the notes given are 4.

The mean of ln(1+Price) is 9.72, so the mean of the price is approximately 16 700, which is pretty high for second-handed cars. It is due to the fact that the raw data contains outliers with cars proposed at prices exceeding 200 000 euros.

V- Analysis

Correlation map using Pearson's r on all the data :



By first printing a correlation map, we can already observe possible direct effects on price. We observe that age and kilometers seem to have a negative impact when all the options seem to have a positive effect on the price.

In the final table, the dependent variable is the price, and the control variables are numerous. Here are the most important: the number of kilometers, the rating of the seller, number of reviews, and some car options such as air conditioning, electric windows, electric seats etc.

Some of the features' distribution are skewed on the left as shown in the figure

1.

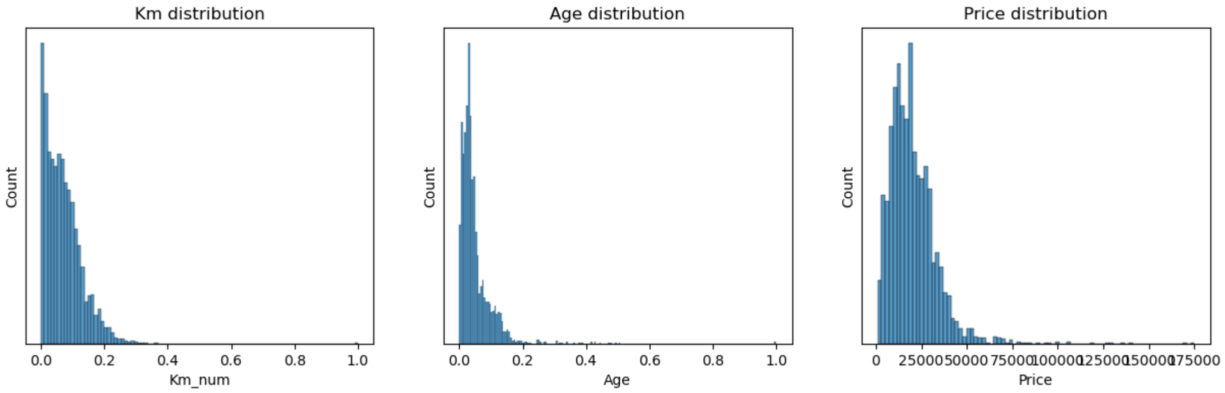


Figure 1: Some features' distribution distribution skewed on the left

This may have a negative impact on our OLS regression. We used a logarithmic transformation to remove the skewness to have a more normally-shaped bell curve. In figure 2, the transformation has been applied and we can see that the features' distribution is not anymore skewed.

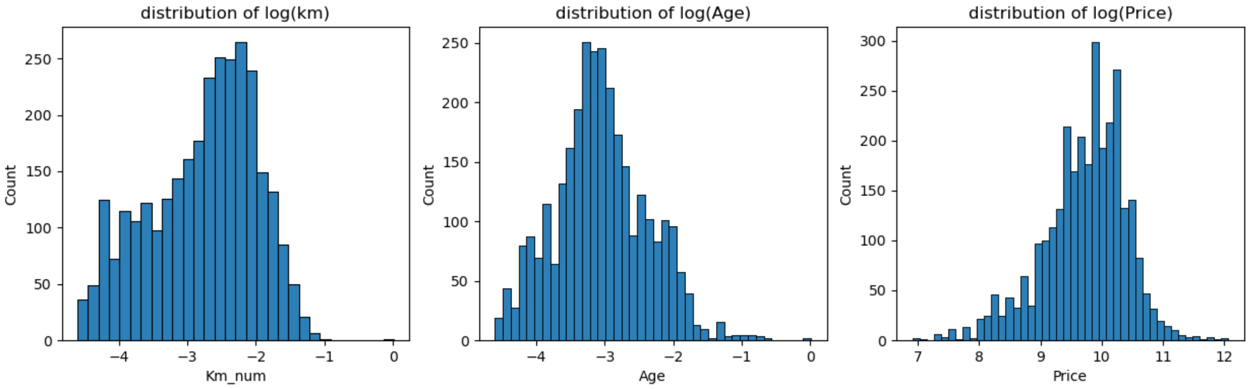
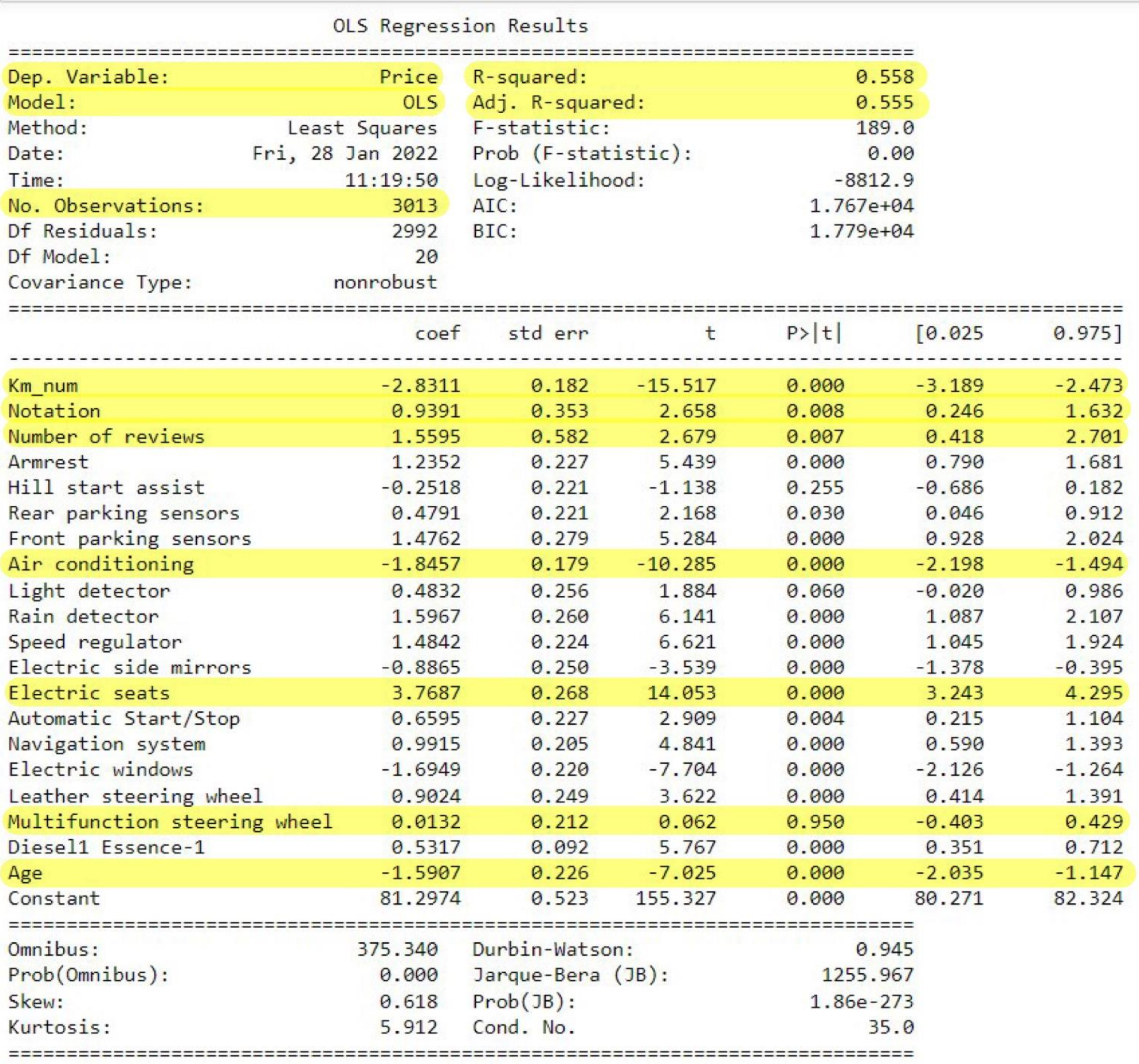


Figure 2: Some features' distribution after the logarithmic transformation (the skewness is removed)

For some of the categorical features having multiple distinct values we used dummy variables to use their information. For the feature *type of fuel* as we had mainly two distinct values: diesel and essence, we chose to put numerical values 1 for the first class, -1 for the other.

We then applied an OLS regression on the table.

<u>Figure 3: Summary of OLS regression results (important information highlighted in yellow)</u>



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.558
Model:                            OLS   Adj. R-squared:                  0.555
Method:                 Least Squares   F-statistic:                     189.0
Date:                Fri, 28 Jan 2022   Prob (F-statistic):               0.00
Time:                        11:19:50   Log-Likelihood:                -8812.9
No. Observations:                3013   AIC:                         1.767e+04
Df Residuals:                    2992   BIC:                         1.779e+04
Df Model:                          20
Covariance Type:            nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Km_num                        -2.8311      0.182    -15.517      0.000      -3.189      -2.473
Notation                       0.9391      0.353      2.658      0.008       0.246       1.632
Number of reviews              1.5595      0.582      2.679      0.007       0.418       2.701
Armrest                        1.2352      0.227      5.439      0.000       0.790       1.681
Hill start assist             -0.2518      0.221     -1.138      0.255      -0.686       0.182
Rear parking sensors           0.4791      0.221      2.168      0.030       0.046       0.912
Front parking sensors          1.4762      0.279      5.284      0.000       0.928       2.024
Air conditioning              -1.8457      0.179    -10.285      0.000      -2.198      -1.494
Light detector                 0.4832      0.256      1.884      0.060      -0.020       0.986
Rain detector                  1.5967      0.260      6.141      0.000       1.087       2.107
Speed regulator                1.4842      0.224      6.621      0.000       1.045       1.924
Electric side mirrors         -0.8865      0.250     -3.539      0.000      -1.378      -0.395
Electric seats                 3.7687      0.268     14.053      0.000       3.243       4.295
Automatic Start/Stop           0.6595      0.227      2.909      0.004       0.215       1.104
Navigation system              0.9915      0.205      4.841      0.000       0.590       1.393
Electric windows              -1.6949      0.220     -7.704      0.000      -2.126      -1.264
Leather steering wheel         0.9024      0.249      3.622      0.000       0.414       1.391
Multifunction steering wheel   0.0132      0.212      0.062      0.950      -0.403       0.429
Diesel1 Essence-1              0.5317      0.092      5.767      0.000       0.351       0.712
Age                           -1.5907      0.226     -7.025      0.000      -2.035      -1.147
Constant                      81.2974      0.523    155.327      0.000      80.271      82.324
==============================================================================
Omnibus:                      375.340   Durbin-Watson:                   0.945
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1255.967
Skew:                           0.618   Prob(JB):                    1.86e-273
Kurtosis:                       5.912   Cond. No.                         35.0
==============================================================================
```

<u>Coefficient of determination:</u>

The OLS regression has been done on 3013 cars. We get a coefficient of determination of 0.56, which means that 56% of the variation in the price is explained by the control variables.

<u>p-value:</u>

The *multifunction steering wheel* has a p-value of 0.95. It means that the hypothesis that the coefficient of this feature is zero is accepted.

Thus the regression suggests that this feature has no influence on the variation of the price and can be dropped.

Interpretation of features' coefficient:

The coefficients of the features *number of kilometers* and *age* are both negative. Therefore the higher these features are the lower the price is. It was an expected outcome: an old and very used car is cheaper. Another observation is that these coefficients are relatively high compared to most of the others, meaning that these features have an important impact on the variation of the price.

It was also an expected outcome, the sellers directly fix the price of their car using these two features. The buyers also often look at these values first.

One surprising coefficient is the one linked to the *air conditioning* feature. This coefficient is negative meaning that the price tends to be lower if this option exists. This can be explained by the fact that the seller does not mention the air conditioning on high tech cars as it is obvious that this option is available on such cars.

Electric seats is the feature with the highest coefficient meaning that it has a relatively large impact on the price variation. This can be because this feature may be related to the fact that if this option exists then the car is new and of high quality.

The standard error is relatively low for these features compared to their coefficient making the interpretation above reliable.

The *rating* and *number of reviews* have positive coefficients meaning that the price gets higher if these features are high. This answers the initial question, showing that these features also have an impact on the variation of the price. This shows that in a market in which there are information imbalances, trust is important and has a direct influence on the buyer decision, allowing the seller to fix a higher price for his car. However the standard error is high and the decisions made during the preprocessing make this outcome less reliable.

Let's run our model in a test set to prevent eventual overfitting.

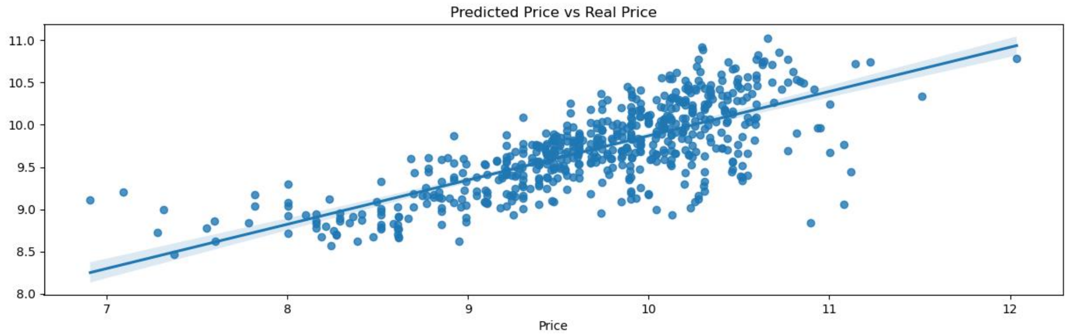


Figure 4: OLS fitting on the data from the test set

The data is well fitted in the test set showing that our model does not overfit during the training phase.

VI- Conclusion

The linear regression showed that the rating and the number of reviews are also impactful on the price fixed by the seller. This answers our initial question of this case study. Asymmetric information in the market of used cars is very important, that explains the importance of trust in this market. The ratings and number of reviews on autoscout24 gives the buyer a feeling of security which possibly leads him to buy a car a little bit more expensive than another one with the same options, if the seller of the first car can be trusted.