BAN 5573 - Fall 2021

Members: Xuyao Liu

Junyang Yan

FINAL PEOJECT

Topic: *Analysis Car's Price*

1 Introduction

Through this analysis, car sales companies can provide their potential consumers and their most suitable models. Similarly, consumers can also use this analysis to figure out their most suitable models and help them make the right decision. Besides, we can also use this to predict the comparison of the passing car types of some new models.

Automobiles, known as "the machine that changes the world", is the third largest international trade item after tourism and oil. Automobile industry also plays an important role in the US's economic development.

Sallee and Slemrod (2011) describe notches, namely settings where the applicable regulations or tax rates change in a discontinuous fashion with respect to the values of the variable subject to the regulations or tax. Notches encourage small changes in behaviors or product adjustments, and yet result in large changes in outcomes. The US gas guzzler tax, for example, is paid for by automakers or auto importers, and the exact amount of the tax depends on the fuel economy interval that each particular car falls into. That is why we should highly consider the influence of the fuel consumed.

2 Overview of the systematic literature review method

2.1 Research Questions

We have derived our questions used in the research directly from the insights of our datasets,we want to answer these questions:

.How does the brand affect the price?

.What is the difference between search in Price VS. popularity?

.What are the challenges in this price research?

Cars dataset with features including make, model, year, engine, and other properties of the car so that we can use it to predict its price.

I just first set up a hypothesis that car brands will have a negative impact on predicting car performance. Author Sullivan M W(1998) considered the effect of brand names on demand by examining the price ratios of used twin automobiles. Twins cars usually are made in the same plant and have essentially the same physical attributes but different brand names. If the models of a twin pair are perceived as perfect substitutes, their relative price should equal unity.

Guoqiong Xu.(2011) lists some of his views about some variables that influence car exterior and interior on car price. Automobile appearance and interior decoration are important factors affecting consumers' driving comfort.Automobile economic characteristics mainly refer to fuel consumption parameters. The impact of car safety on car prices.Consumers pay attention to the characteristics of automobile safety performance is concerned about their own life, so the characteristics of automobile safety devices are very important to the impact of automobile price.brand and word of mouth on car prices also have true influence.

Lee Schipper(2010) announced without taking a position on the relative importance of standards versus fuel prices, taxes versus voluntary agreements, or other factors affecting car price and fuel use, it is still important to understand how these factors affect car economy and car price mark.

2.2 Search strategy

The main strategy is to draw conclusions from our analysis based on recent vehicle pricing intelligence and expert insights. And the analysis forms are based on using tools like Python, Knime, Tableau and literature collection.

3 Overview of studies

We have identified some studies in the literature which focus on the variables on the price decision on the car price.A quick look at the studies show that, all work performed on car price during from 2010 - 2020, On the top of that, we have also considered the publication venues for the papers selected.

4 Literature Gap

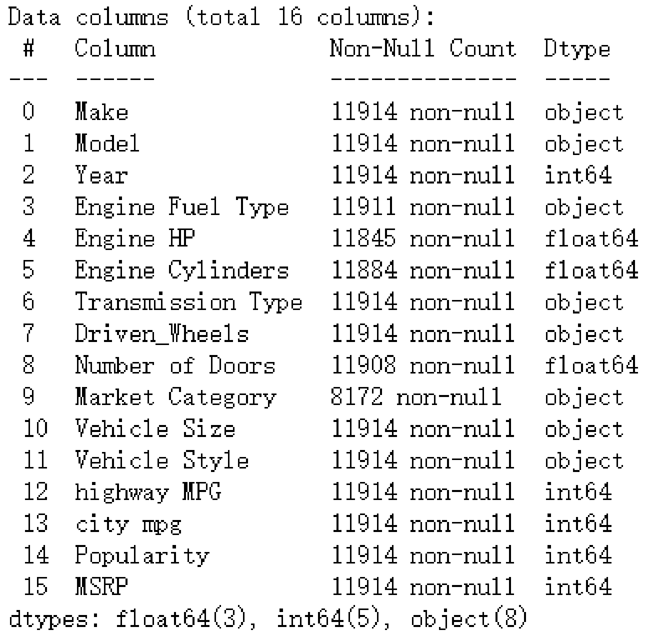
Gaps in Literature are missing pieces or insufficient information in the research literature. These are areas that have scope for further research because they are unexplored, under-explored, or outdated.

First I think some materials seem redundant. For example, The literature from China only provides the views and trends based on tha local area,which may not be very suitable for the US car market.

Second, the literature related to the definition of car price uses many variables but only few of them can be used in our research. On the one hand, it is based on the group's lack of analytical knowledge, on the other hand that some information redundant will affect our results.

5 Visulization

Link of data: https://www.kaggle.com/CooperUnion/cardataset.



```
Data columns (total 16 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Make               11914 non-null  object 
 1   Model              11914 non-null  object 
 2   Year               11914 non-null  int64  
 3   Engine Fuel Type   11911 non-null  object 
 4   Engine HP          11845 non-null  float64
 5   Engine Cylinders   11884 non-null  float64
 6   Transmission Type  11914 non-null  object 
 7   Driven_Wheels      11914 non-null  object 
 8   Number of Doors    11908 non-null  float64
 9   Market Category    8172 non-null   object 
 10  Vehicle Size       11914 non-null  object 
 11  Vehicle Style      11914 non-null  object 
 12  highway MPG        11914 non-null  int64  
 13  city mpg           11914 non-null  int64  
 14  Popularity         11914 non-null  int64  
 15  MSRP               11914 non-null  int64  
dtypes: float64(3), int64(5), object(8)
```

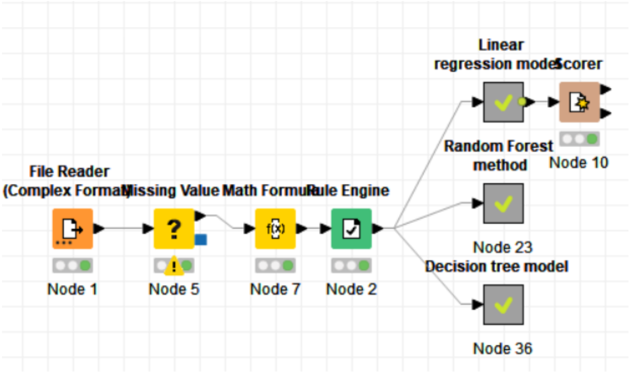
There is dataset, contain 11914 observations, and for each car they are 16 features of data:

- **1 Make:**The brand of the car
- **2 Model:**Specific product of the car
- **3 Year:**The year car published
- **4 Engine Fuel Type:**Define which energy used on car
- **5 Engine HP (horsepower):**The power of engine
- **6 Engine Cylinder:**the power unit of the engine
- **7 Transmission Type**:manual, automatic, and CVT
- **8 Driven Wheels:**Define which wheels used in running
- **9 Number of Doors:**The doors on the car
- **10 Market Category:**The market sell cars
- **11 Vehicle Size:**Size of Vehicle (compact midsize large)
- **12 Vehicle Style:**Style of vehicle
- **13 highway MPG (miles per gallon):**The gallon used on run each mile on hw
- **14 city mpg (miles per gallon):** The gallon used on run each mile in city
- **15 Popularity:** quantify as numbers,large number means greater popularity
- **16 MSRP:** Manufacturer Suggested Retail Price

There are totally 16 columns in this dataset, the highly related variables will be explained in our report.

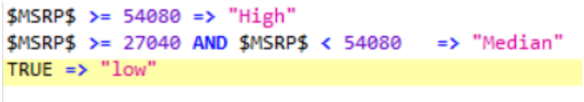
6 Methodology

6.1Knime analysis



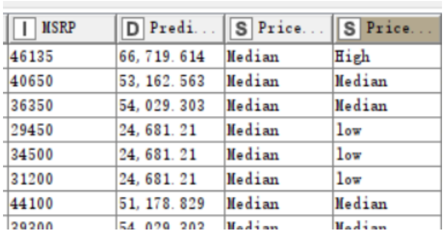
First we find the number of doors and market categories have some missing values, and the percentage of these missing values is less than 25%, so we delete the rows with missing data.

Next,I have counted the prices of all vehicles, and based on the distribution，Use math formula and rule engine set ‘Low’, ‘Median’, ‘High’ at the different trisection points (look below), to contribute to a new column.

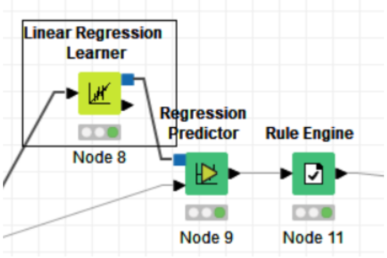


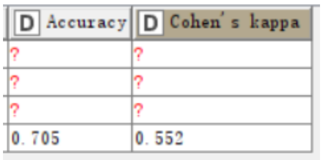
```
$MSRP$ >= 54080 => "High"
$MSRP$ >= 27040 AND $MSRP$ < 54080    => "Median"
TRUE => "low"
```

Then we use the simple regression model to find the predicted suggested retail price, using four numeric variables(Engine horsepower, Highway mpg, City mpg and popularity). send results according to the above formula, redefine the price classification.



I MSRP	D Predi...	S Price...	S Price...
46135	66, 719. 614	Median	High
40650	53, 162. 563	Median	Median
36350	54, 029. 303	Median	Median
29450	24, 681. 21	Median	low
34500	24, 681. 21	Median	low
31200	24, 681. 21	Median	low
44100	51, 178. 829	Median	Median
29300	54, 029. 303	Median	Median

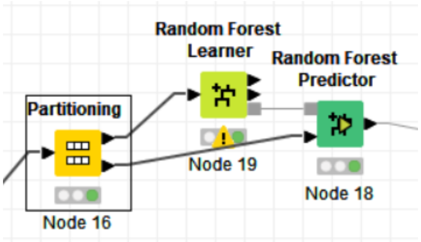




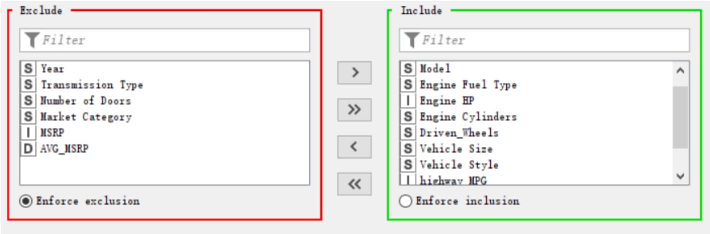
D Accuracy	D Cohen's kappa
?	?
?	?
?	?
0.705	0.552

Finally we get statistical accuracy at 0.705 and 0.552 cohen's kappa.

Then following we use a random forest and tree based method.



I split the original data into training(80%) and test(20%) data,and do filter to remove unnecessary columns.(Year, Transmission Type, Number of doors and so on)



Exclude

Filter

- S Year
- S Transmission Type
- S Number of Doors
- S Market Category
- I MSRP
- D AVG_MSRP

● Enforce exclusion

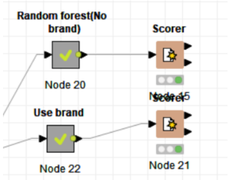
Include

Filter

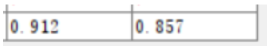
- S Model
- S Engine Fuel Type
- I Engine HP
- S Engine Cylinders
- S Driven_Wheels
- S Vehicle Size
- S Vehicle Style
- I highway MPG

○ Enforce inclusion

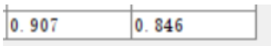
I consider Brand is a significant variable to define the car price level, so I do another model within Brand.



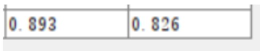
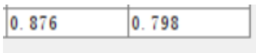
And I get different results: 0.912 0.857 (No brand) 0.907 0.846 (Use brand), so when we use no brand model we have a bit higher accuracy to predict the car price but when we consider brand as an affect variable,that eh accuracy and cohen's kappa decrease.



0.912	0.857

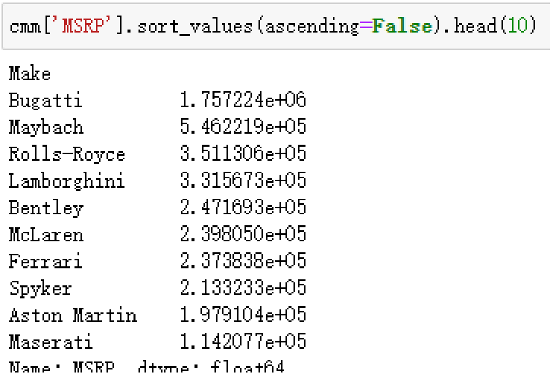


0.907	0.846

Same to use decision tree model, we get results  0.893 0.826 (No brand)  0.876 0.798 (Use brand).We get the same trend as random forest model(slightly decrease in accuracy and the cohen’s kappa).

6.2 Python analysis

1,Brands have value in price decision making?



```
cmm['MSRP'].sort_values(ascending=False).head(10)
```

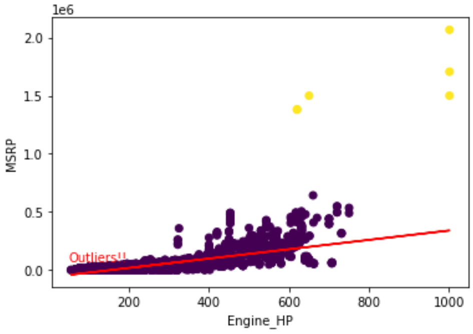
```
Make
Bugatti          1.757224e+06
Maybach          5.462219e+05
Rolls-Royce      3.511306e+05
Lamborghini      3.315673e+05
Bentley          2.471693e+05
McLaren          2.398050e+05
Ferrari          2.373838e+05
Spyker           2.133233e+05
Aston Martin     1.979104e+05
Maserati         1.142077e+05
Name: MSRP, dtype: float64
```

2. Can we explain the price of a car using the engine horsepower? How much?

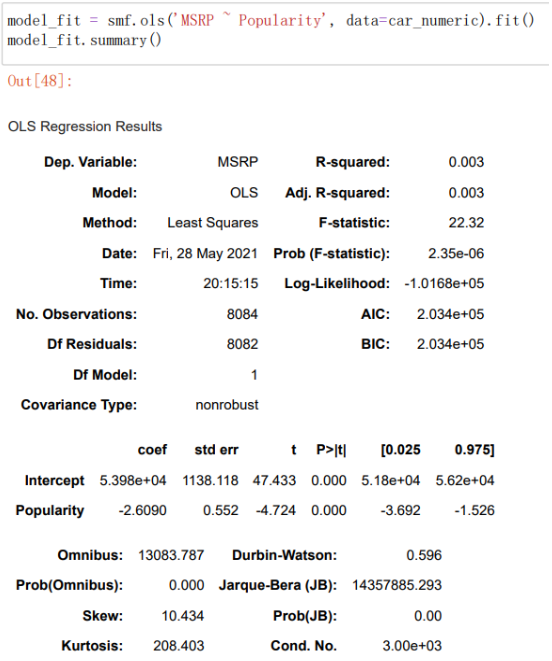
Price = $\hat{\propto}$+ $(\hat{\beta})$(popularity or engine horsepower)+ ϵ

Null Hypothesis: the coefficient is significantly different from zero. (Does the x variable add value to the prediction of y?)

$H0:\beta=0$; $Ha:\beta\neq0$



We use a scatter plot to find the relationship between Engine HP and MSRP, and we treat the MSRP>1100000 as outliers.



```
model_fit = smf.ols('MSRP ~ Popularity', data=car_numeric).fit()
model_fit.summary()
```

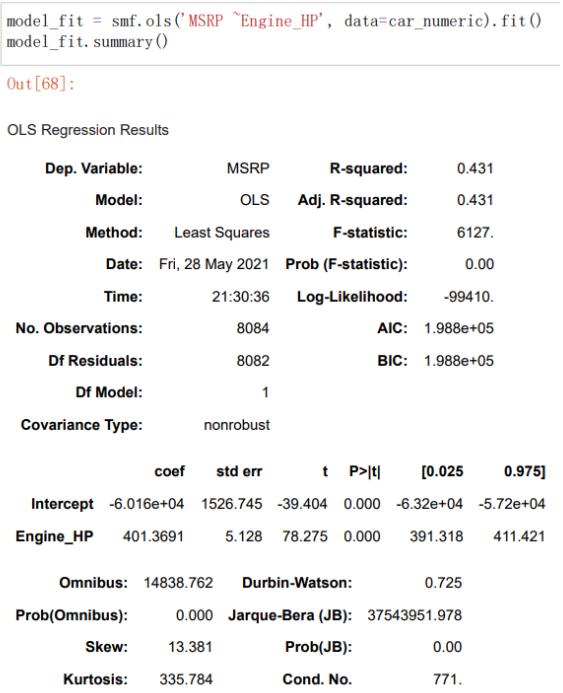
Out[48]:

OLS Regression Results

Dep. Variable:	MSRP	R-squared:	0.003
Model:	OLS	Adj. R-squared:	0.003
Method:	Least Squares	F-statistic:	22.32
Date:	Fri, 28 May 2021	Prob (F-statistic):	2.35e-06
Time:	20:15:15	Log-Likelihood:	-1.0168e+05
No. Observations:	8084	AIC:	2.034e+05
Df Residuals:	8082	BIC:	2.034e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.398e+04	1138.118	47.433	0.000	5.18e+04	5.62e+04
Popularity	-2.6090	0.552	-4.724	0.000	-3.692	-1.526

Omnibus:	13083.787	Durbin-Watson:	0.596
Prob(Omnibus):	0.000	Jarque-Bera (JB):	14357885.293
Skew:	10.434	Prob(JB):	0.00
Kurtosis:	208.403	Cond. No.	3.00e+03



```
model_fit = smf.ols('MSRP ~Engine_HP', data=car_numeric).fit()
model_fit.summary()
```

Out[68]:

OLS Regression Results

Dep. Variable:	MSRP	R-squared:	0.431
Model:	OLS	Adj. R-squared:	0.431
Method:	Least Squares	F-statistic:	6127.
Date:	Fri, 28 May 2021	Prob (F-statistic):	0.00
Time:	21:30:36	Log-Likelihood:	-99410.
No. Observations:	8084	AIC:	1.988e+05
Df Residuals:	8082	BIC:	1.988e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.016e+04	1526.745	-39.404	0.000	-6.32e+04	-5.72e+04
Engine_HP	401.3691	5.128	78.275	0.000	391.318	411.421

Omnibus:	14838.762	Durbin-Watson:	0.725
Prob(Omnibus):	0.000	Jarque-Bera (JB):	37543951.978
Skew:	13.381	Prob(JB):	0.00
Kurtosis:	335.784	Cond. No.	771.

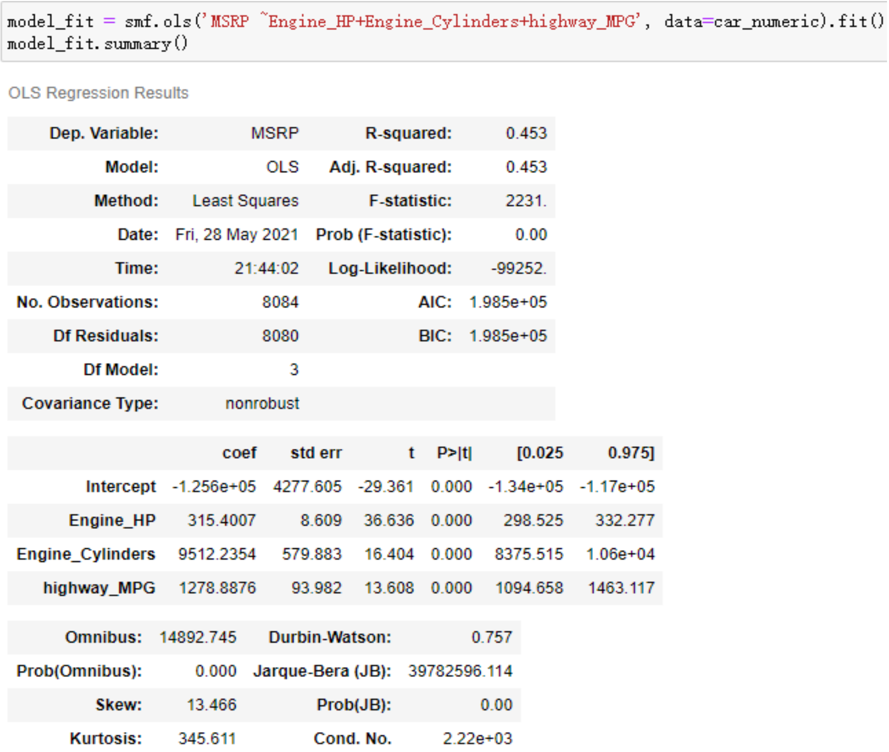
3. Can we explain the MSRP of a car using the three main variables?

MSRP or Popularity = ∝ ̂+ (𝛽1) ̂(engine horsepower)+ (𝛽2) ̂(engine cylinders)+(𝛽3) ̂(highway MPG)+є

Null hypothesis: The variation explained by the model is by chance

Null Hypothesis: the coefficient is significantly different from zero. (Does the x variable add value to the prediction of y?)

$H0$:$\beta1$,$\beta2$,$\beta3$=0; Ha:At least one β≠0



```
model_fit = smf.ols('MSRP ~Engine_HP+Engine_Cylinders+highway_MPG', data=car_numeric).fit()
model_fit.summary()
```

```
OLS Regression Results

       Dep. Variable:             MSRP       R-squared:        0.453
              Model:               OLS   Adj. R-squared:       0.453
             Method:     Least Squares       F-statistic:       2231.
               Date:  Fri, 28 May 2021 Prob (F-statistic):      0.00
               Time:          21:44:02   Log-Likelihood:      -99252.
   No. Observations:             8084              AIC:    1.985e+05
       Df Residuals:             8080              BIC:    1.985e+05
           Df Model:                3
    Covariance Type:         nonrobust

                       coef    std err         t    P>|t|     [0.025     0.975]
       Intercept  -1.256e+05   4277.605   -29.361   0.000   -1.34e+05  -1.17e+05
       Engine_HP    315.4007      8.609    36.636   0.000     298.525    332.277
Engine_Cylinders   9512.2354    579.883    16.404   0.000    8375.515   1.06e+04
     highway_MPG   1278.8876     93.982    13.608   0.000    1094.658   1463.117

        Omnibus:  14892.745      Durbin-Watson:         0.757
  Prob(Omnibus):      0.000   Jarque-Bera (JB):  39782596.114
           Skew:     13.466           Prob(JB):          0.00
       Kurtosis:    345.611           Cond. No.      2.22e+03
```

Result and Conclusion

1,There are eight brands that have higher prices than others: Bugatti, Maybach, Rolls-Royce, Lamborghini, Bentley, McLaren, Ferrari, Spyker.

We use the mean of each brand('Make' column),To show the features of the cars.From the graph and line table we can point top8 brand with highest Price)

2,The MSRP has a high correlation to Engine horsepower.

Prob value is less than 0.05, So horsepower is significant. R-square value is a standard to show how strong is a linear regression between two variables. r is sqrt R-square value, r between -1 and +1, 0 means no correlation. Correlations of -1 or +1 imply an exact linear relationship. Positive correlations r that as x increases, so does y. Negative correlations are opposite. R-square is 0.431, r = 0.657. It is showing a positive linear regression,0.657 means it’s moderate correlation linear regression.

3,There is no clear evidence that MSRP is associated with popularity.

R-square is 0.003, r = 0.055. It is showing a positive linear regression,0.07 means it's nearly none of correlation linear regression.

4,Retail price of a car is much affected by its Parts and performance, car brands have a negative impact on customers' judgments about the performance of a car

Using a simple regression model to predict the price rank is less accurate than using other machine learning models(Such as tree base and random forest),almost 10% difference(70.5% compared to 80%),but all these models can have a high accuracy to predict price rank.

Next, I find that when we use 'Make'(As car brand) into our machine learning model, we both two model have small decrease in accuracy and cohens' kappa, means that when we consider other performances variable to define a car's price and rank, it's more reliable than to choose brand value in our consideration, When we choose a car between the same price level, a popular brand may means that the performance of a car is bit lower than others.

Limitation

The sample size can be larger, which can reduce the error, and there are much outliers in the value, which is easy to lead to errors in the conclusion(Such as we present before,MSRP have high correlation to Engine power,but in the regression model it shows not, because of 4-5 extreme outliers!)

References

Alberini, Anna; Bareit, Markus; Filippini, Massimo.(2017).What Is the Effect of Fuel Efficiency Information on Car Prices? Evidence from Switzerland.Energy Journal, v. 37, iss. 3, pp. 315-42.

Guoqiong Xu.(2011). An Empirical Study on Automobile New Product Pricing [D].Chongqing: Master's Thesis of Chongqing Normal University.

Schipper L, Hand P, Gillingham K. The Road from Copenhagen: Fuel Prices and Other Factors Affecting Car Use and CO2 Emissions in Industrialized Countries[C]//12th World Conference on Transport Research, July. 2010: 11-15.

Sullivan M W. How brand names affect the demand for twin automobiles[J]. Journal of marketing research, 1998, 35(2): 154-165.

Engers M, Hartmann M, Stern S. Annual miles drive used car prices[J]. Journal of Applied Econometrics, 2009, 24(1): 1-33.

Betts S C, Taran Z. The'brand halo'effect: Brand reliability influence on used car prices[C]//Allied Academies International Conference. Academy of Marketing Studies. Proceedings. Jordan Whitney Enterprises, Inc, 2002, 7(1): 19