International Journal of Mathematics and Statistics Invention (IJMSI)
E-ISSN: 2321 – 4767 P-ISSN: 2321 - 4759
www.ijmsi.org Volume 4 Issue 6 || August. 2016 || PP-01-10

Multiple Linear Regression Applications Automobile Pricing

Ceyhun Ozgur[1], Zachariah Hughes[2], Grace Rogers[3], Sufia Parveen[4]
Ph.D., CPIM,Professor, Valparaiso University College of Business Information & Decision Sciences Urschel Hall 223 – Valparaiso University Valparaiso, IN 46383
[2]Undergraduate Research Assistant Valparaiso University Finance – College of Business Economics – College of Arts and Sciences
[3]Undergraduate Research Assistant Valparaiso University Actuarial Science - College of Arts and Sciences Business Analytics - College of Business
[4]Graduate Research Assistant Valparaiso University M.S. in Analytics and Modeling

I. INTRODUCTION

This paper is about 470 cars selected as a representative sample of all 2005 GM cars with the make of either Chevrolet or Pontiac. The information on each car in the sample was taken from Kelley Blue Book. [5] The purpose of this paper is to develop a relatively good regression equation for predicting the price of these cars. It is known that there are many factors that influence the price of the cars, but we do not know what factors will influence the price of the cars and how these factors influence the price. If we are a buyer, we can judge whether the price of the car we are looking to buy is rational or not. If we are a seller, we can choose a rational price according to the equation and then it is good for our sales.

When the goal of developing a multiple regression model is description or prediction, the primary issue is often determining which variables to include in the model (and which to leave out).[6] All potential explanatory variables can be included in a regression model, but that often results in a cumbersome model that is difficult to understand. [2] On the other hand, a model that includes only one or two of the explanatory variables may be much less accurate than a more complex model. These models were utilized using SAS software and all of the resulting tables are given from the output of SAS [7] [8] [9] [10]. This tension between finding a simple model and finding the model that best explains the response is what makes it difficult to find a "best" model. The process of finding the most reasonable mix, which provides a relatively simple linear combination of explanatory variables, often resembles an exploratory artistic process much more than a formulaic recipe.[11] Including redundant or unnecessary variables not only creates an unwieldy model but also can lead to test statistics (and conclusions from corresponding hypothesis tests) that are less reliable. If explanatory variables are highly correlated, then their effects in the model will be estimated with more imprecision. This imprecision leads to larger standard errors and can lead to insignificant test results for individual variables that can be important in the model. Failing to include a relevant variable can result in biased estimates of the regression coefficients and invalid t-statistics, especially when the excluded variable is highly significant or when the excluded variable is correlated with other variables. Variable selection techniques are used to describe or predict a response. If our objective is to describe a relationship or predict new response variables, variable selection techniques are useful for determining which explanatory variables should be in the model. For this investigation, we will consider the response to be the suggested retail price from Kelley Blue Book.

Therefore, for this paper, the response variable is the price of these cars, which we can find from the Kelley Blue Book. Because the price is a numeric variable, it can show the level of the price. We may initially believe the following are relevant potential explanatory variables. There are 7 variables including mileage, make, type, liter, cruise control, upgraded speakers and leather seats. Specifically, mileage and liter are measured by digits, while the rest of variables are character variables, so we can use a method to change these character variables to numeric variables. In this study, there are 470 cars and they are observational subjects.

II. PRELIMINARY DATA EXPLORATION

For quantitative predictors, we will use scatterplots and correlation to assess. For categorical predictors, we will use boxplots to assess. At the same time, we will use the method of correlation to find if there are some high correlated potential predictors.

Table 2.1: Descriptive Statistics for Automobile

The Basic Data

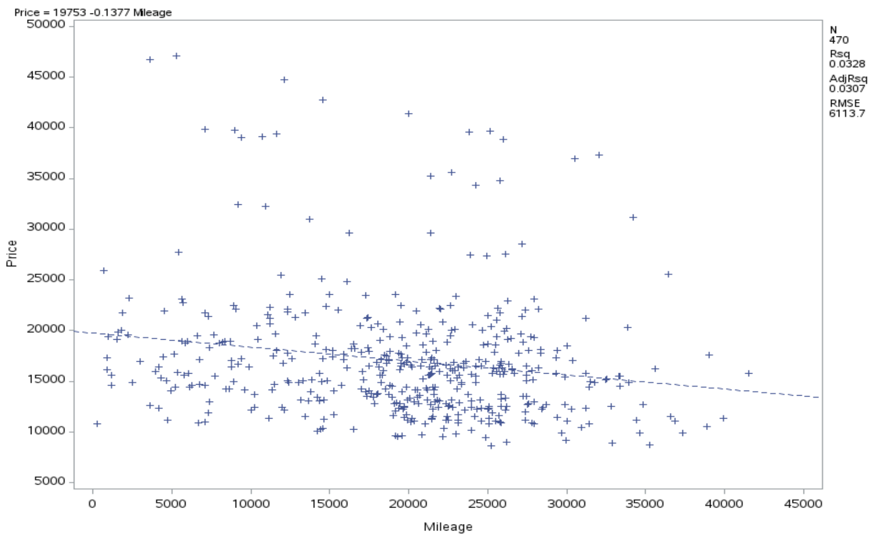
The UNIVARIATE Procedure
Variable: Price

Moments			
N	470	Sum Weights	470
Mean	17060.9491	Sum Observations	8018646.1
Std Deviation	6209.84532	Variance	38562178.9
Skewness	2.15594881	Kurtosis	5.98283278
Uncorrected SS	1.54891E11	Corrected SS	1.80857E10
Coeff Variation	36.3980062	Std Error Mean	286.438805

Basic Statistical Measures			
Location		Variability	
Mean	17060.95	Std Deviation	6210
Median	15959.50	Variance	38562179
Mode	10921.90	Range	38426
		Interquartile Range	5932

Basic data about the response variable (the price of these cars). Specifically, Mean=17060.95 Median=15959.50 Standard deviation=6210.

Graph 2.1: Scatterplot between price of automobiles and milage



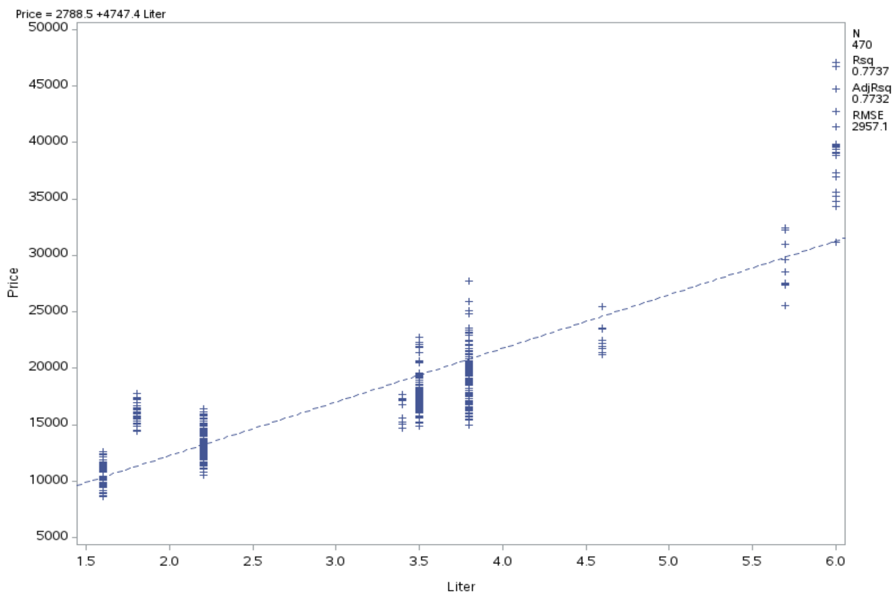


Table 2.2: Automobile Price versus its predictors

The CORR Procedure

8 Variables:	Price	Make_c	Type_c	Cruise_Control_c	Upgraded_Speakers_c	Leather_c	Mileage	Liter

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Price	470	17061	6210	8018646	8639	47065
Make_c	470	1.31915	0.46664	620.00000	1.00000	2.00000
Type_c	470	3.85106	1.41540	1810	1.00000	5.00000
Cruise_Control_c	470	1.35319	0.47847	636.00000	1.00000	2.00000
Upgraded_Speakers_c	470	1.24255	0.42908	584.00000	1.00000	2.00000
Leather_c	470	1.24468	0.43036	585.00000	1.00000	2.00000
Mileage	470	19549	8167	9187887	266.00000	41566
Liter	470	3.00638	1.15058	1413	1.60000	6.00000

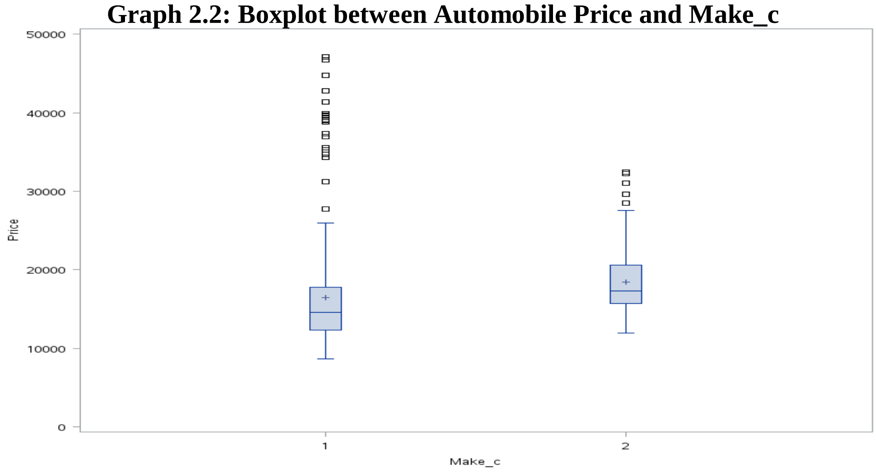
Pearson Correlation Coefficients, N = 470
Prob > \|r\| under H0: Rho=0

	Price	Make_c	Type_c	Cruise_Control_c	Upgraded_Speakers_c	Leather_c	Mileage	Liter
Price	1.00000	0.14913	-0.07195	-0.35777	0.12475	-0.04816	-0.18108	0.87961
		0.0012	0.1193	<.0001	0.0068	0.2975	<.0001	<.0001
Make_c	0.14913	1.00000	0.26581	-0.15259	0.24084	0.18366	-0.01914	0.17490
	0.0012		<.0001	0.0009	<.0001	<.0001	0.6790	0.0001
Type_c	-0.07195	0.26581	1.00000	-0.13625	0.06312	0.16147	0.01675	-0.02167
	0.1193	<.0001		0.0031	0.1719	0.0004	0.7172	0.6393
Cruise_Control_c	-0.35777	-0.15259	-0.13625	1.00000	-0.10660	-0.12029	-0.05722	-0.43363
	<.0001	0.0009	0.0031		0.0208	0.0090	0.2156	<.0001
Upgraded_Speakers_c	0.12475	0.24084	0.06312	-0.10660	1.00000	0.34763	0.03575	0.10612
	0.0068	<.0001	0.1719	0.0208		<.0001	0.4394	0.0214
Leather_c	-0.04816	0.18366	0.16147	-0.12029	0.34763	1.00000	0.01142	-0.03244
	0.2975	<.0001	0.0004	0.0090	<.0001		0.8050	0.4829
Mileage	-0.18108	-0.01914	0.01675	-0.05722	0.03575	0.01142	1.00000	0.00840
	<.0001	0.6790	0.7172	0.2156	0.4394	0.8050		0.8559
Liter	0.87961	0.17490	-0.02167	-0.43363	0.10612	-0.03244	0.00840	1.00000
	<.0001	0.0001	0.6393	<.0001	0.0214	0.4829	0.8559	

From graph 2.1, we can find that the price and mileage have a negative relationship in general; the price decreases as the mileage increases. In detail, the correlation coefficient between price and mileage is -0.18108, with a p-Value<0.0001, which indicates that the correlation is significant. graph 2.2 we can find that the price and liter have positive relationship, which means that the price increases as the liter increases. In detail, the correlation coefficient between price and liter is 0.87961, with a p-Value<0.0001. This indicates that the correlation between price and liter is significant.

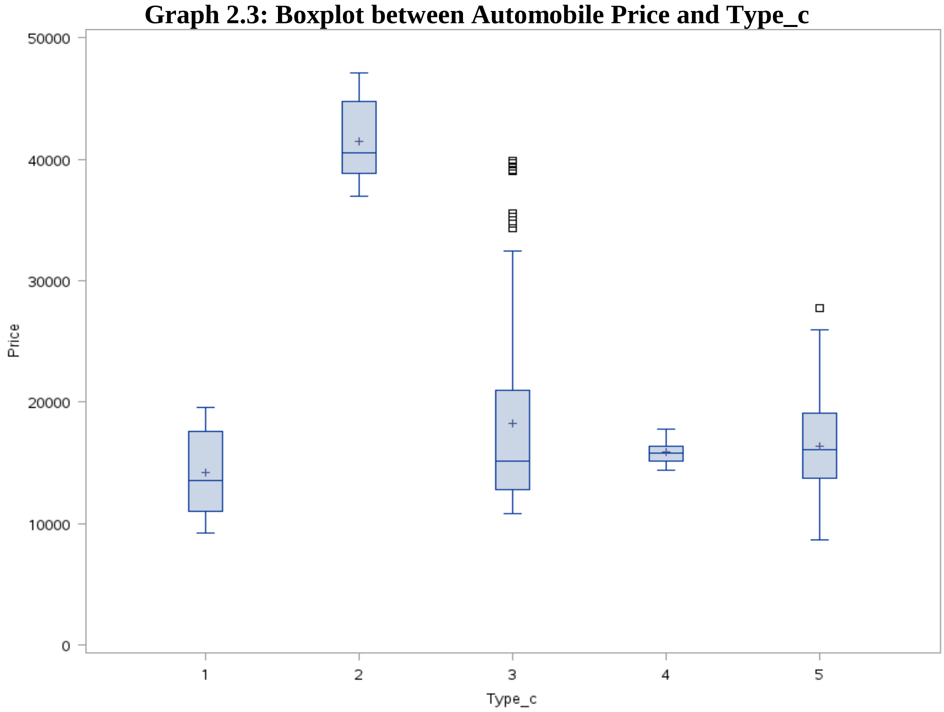
Then we should find the relationships among the potential predictors through the table. We can find that the correlation coefficient between liter and cruise control is -0.43363 and the correlation coefficient between leather and updated speaker is 0.34763. These two absolute values of correlation coefficient are both greater than 0.3 and less than 0.8, so both liter and cruise control, and leather and updated speaker have middling correlation. Other absolute values of correlation coefficients are less than 0.3, so we can conclude they have low correlation relationships.

Graph 2.2: Boxplot between Automobile Price and Make_c



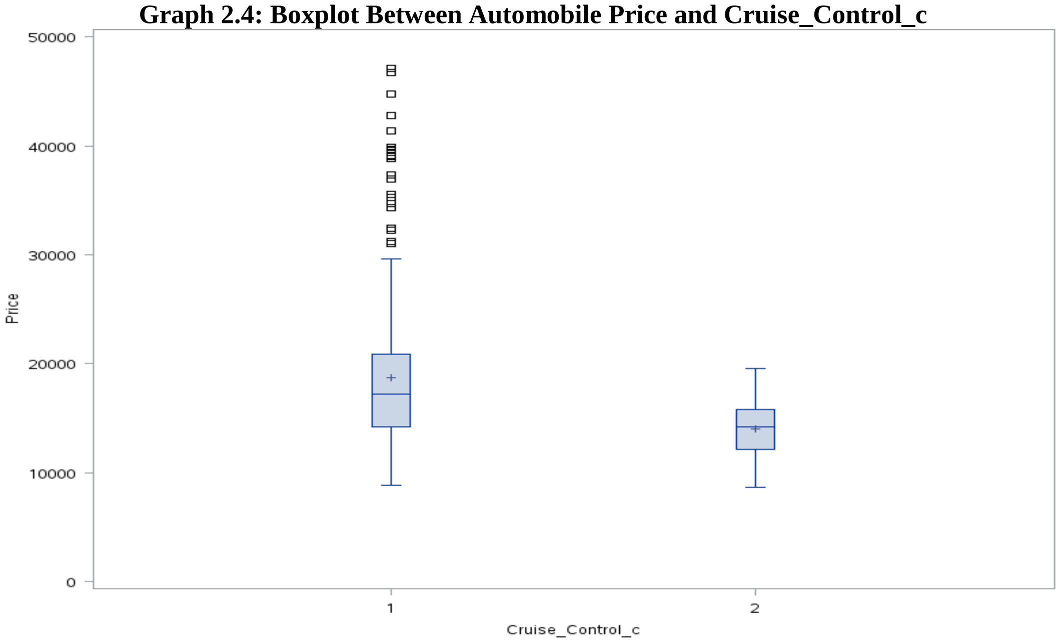
From graph 2.2, we can find that the price range is from less than 10,000 dollars to nearly 26,000 dollars for cars belonging to Chevrolet. In addition, the price range is from nearly 12,000 dollars to nearly 28,000 dollars for cars belonging to Pontiac. Both of these two have some outliers.

Graph 2.3: Boxplot between Automobile Price and Type_c



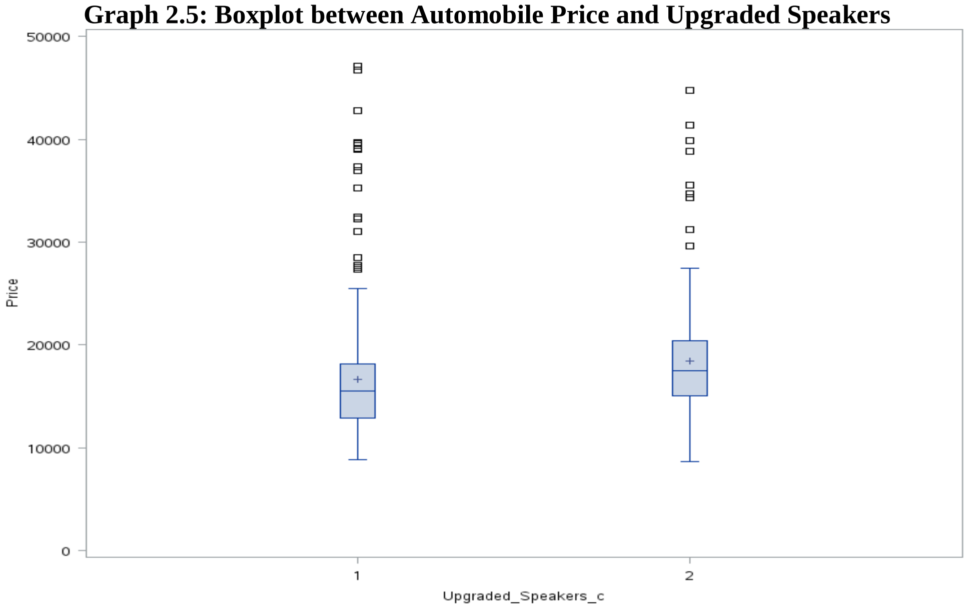
In graph 2.3, we observe that the price range is from nearly 9000 dollars to nearly 19000 dollars when the type of car is a Hatchback. The price range is from nearly 37000 dollars to greater than 40000 dollars when the type of car is a Convertible. The price range is from nearly 11000 dollars to nearly 33000 dollars when the type of car is a Coupe. The price range is from nearly 14000 dollars to nearly 18000 dollars when the type of car is a Wagon. In addition, the price range is from nearly 8000 dollars to greater than 25000 dollars when the type of car is a Sedan. The types of Coupe and Sedan have some outliers.

Graph 2.4: Boxplot Between Automobile Price and Cruise_Control_c



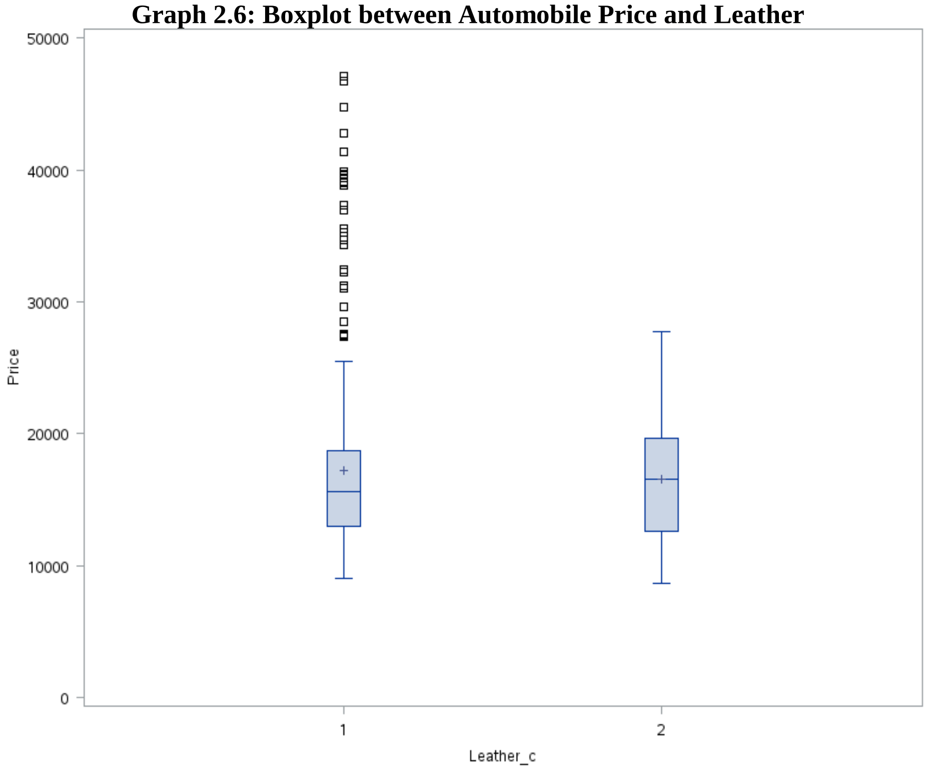
In graph 2.4 above, price ranges from 9000 dollars to nearly 29000 dollars when the the cars have cruise control. In addition, the price range is from nearly 8000 dollars to 19000 dollars when the cars do not have cruise control. The cars that have cruise control have some outliers.

Graph 2.5: Boxplot between Automobile Price and Upgraded Speakers



From graph 2.5, we can find that the price range is from nearly 9000 dollars to nearly 26000 dollars when the cars have updated speakers. In addition, the price range is from nearly 8000 dollars to 28000 dollars when the cars do not have updated speakers. Both of these two have some outliers.

Graph 2.6: Boxplot between Automobile Price and Leather



In graph 2.6, we can find that the price range is from nearly 9000 dollars to nearly 26000 dollars when the the cars have leather seats. In addition, the price range is from nearly 8000 dollars to nearly 28000 dollars when the cars do not have leather seats. The cars that have leather seats have some outliers.

As we can see from the figures, for the different categorical predictors, they all have the outliers. Now that we did some primary analysis for different predictors, we also cannot give up any predictors from the information above.

III. REGRESSION

3.1 Based on all the information and outputs above, my initial model is:

Price= $\beta_0$mileage +$\beta_1$liter +$\beta_2$cruise_control +$\beta_3$upgraded_speakers +$\beta_4$leather_seats +$\beta_5$type + $\beta_6$make+ ε .

From the scatterplots, we can see that the linear relation is negative between price and mileage. The normal thinking is that the price will go down as the mileage goes up. That is to say, the coefficient between price and mileage is negative. Although the dots are dispersed and the linear relation is not obvious, we still include it in our initial model. Secondly, the relation between price and liter is positive, seeing from the scatterplot, so we include it in the initial model. As for the five categorical predictors, we use the side-by-side boxplots to see the distribution. Seeing the boxplots, some outliers exist. But we still include them in our initial model. For the initial model, we include all of the potential predictors in the model to suppose they can have an influence to the price.

Table 3.1: Multiple regression analysis:

Regression

The REG Procedure
Model: MODEL1
Dependent Variable: Price

Number of Observations Read	470
Number of Observations Used	470

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	14724251813	2103464545	289.11	<.0001
Error	462	3361410101	7275779		
Corrected Total	469	18085661914			

Root MSE	2697.36528	R-Square	0.8141
Dependent Mean	17061	Adj R-Sq	0.8113
Coeff Var	15.81017		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5877.29763	973.73868	6.04	<.0001
Make_c	1	-24.44993	290.19157	-0.08	0.9329
Type_c	1	-207.21374	92.99392	-2.23	0.0263
Cruise_Control_c	1	96.96023	295.70251	0.33	0.7431
Upgraded_Speakers_c	1	751.10543	317.28107	2.37	0.0183
Leather_c	1	-385.55120	316.13174	-1.22	0.2232
Mileage	1	-0.14357	0.01530	-9.39	<.0001
Liter	1	4735.22688	123.07453	38.47	<.0001

The fitted quadratic model:

Price=-24.45Make_c-207.21Type_c+96.96Cruise_Control_c+751.11Upgraded_Speakers_c-385.55Leather_c-0.14Mileage+4735.23Liter

The adjusted R^2 is 0.8113, which indicates this model fits the data well. The F-test (ANOVA) tests whether $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$. The p-Value is < 0.0001, indicating that overall the explanatory variables are significant to the response variable.

Parameter estimates and standard errors of the estimates, along with t-tests and p-Values are given too. The t-test tests whether each explanatory variable is zero. Hence, only Mileage and Liter are significant because of p-Value<0.01. Next, we put the Mileage and Liter into our model.

Table 3.2: Multiple Regression Analysis Including Mileage and Liter

Regression2

The REG Procedure
Model: MODEL1
Dependent Variable: Price

Number of Observations Read	470
Number of Observations Used	470

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	14635625024	7317812512	990.55	<.0001
Error	467	3450036890	7387659		
Corrected Total	469	18085661914			

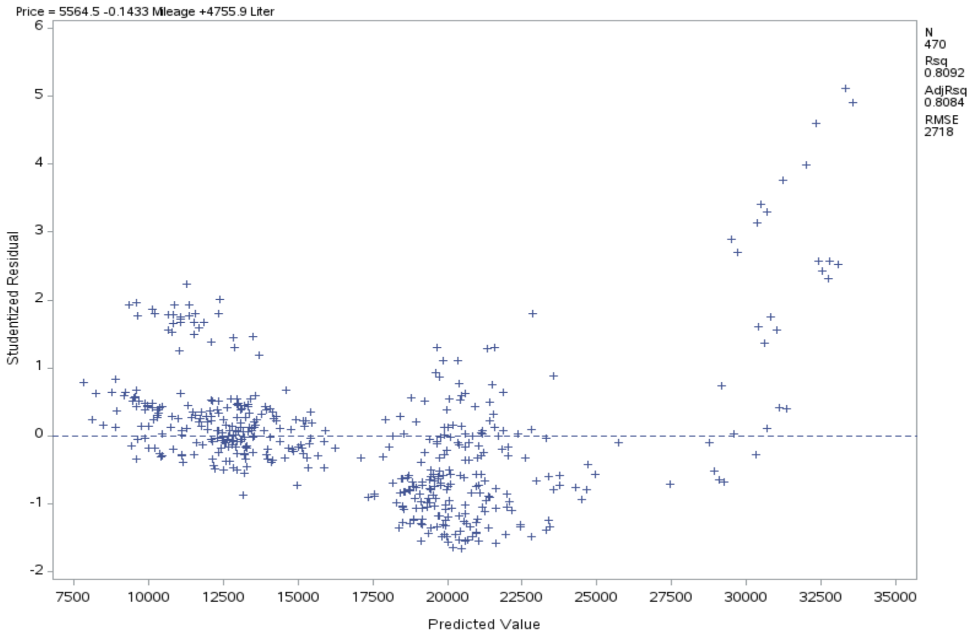
Root MSE	2718.02489	R-Square	0.8092
Dependent Mean	17061	Adj R-Sq	0.8084
Coeff Var	15.93126		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5564.46118	460.29961	12.09	<.0001
Mileage	1	-0.14332	0.01537	-9.33	<.0001
Liter	1	4755.92222	109.08489	43.60	<.0001

From table 3.2, the p-value<0.01, it means that these two variables are significant to the response variable. At the same time, the Adj R-Sq=0.8084, which means that these two variables can explain 80.84% of the change of the price. Finally, we can get the equation:

Price=4755.92Liter-0.1433Mileage+5564.46.

Graph 3.1: Residual Plot (1)



The residual plot shows the assumption of constant variance is not met. So we need to check for multicollinearity and check for the influential observations.

Table 3.3: Checking for Multicollinearity (1)

Checking for Multicollinearity1

The REG Procedure
Model: MODEL1
Dependent Variable: Price

Number of Observations Read	470
Number of Observations Used	470

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	14635625024	7317812512	990.55	<.0001
Error	467	3450036890	7387659		
Corrected Total	469	18085661914			

Root MSE	2718.02489	R-Square	0.8092
Dependent Mean	17061	Adj R-Sq	0.8084
Coeff Var	15.93126		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	5564.46118	460.29961	12.09	<.0001	.	0
Mileage	1	-0.14332	0.01537	-9.33	<.0001	0.99993	1.00007
Liter	1	4755.92222	109.08489	43.60	<.0001	0.99993	1.00007

Collinearity Diagnostics					
		Condition	Proportion of Variation		
Number	Eigenvalue	Index	Intercept	Mileage	Liter
1	2.81375	1.00000	0.00907	0.01722	0.01493
2	0.13729	4.52711	0.00182	0.58587	0.42195
3	0.04896	7.58098	0.98912	0.39691	0.56312

Noticing the tolerance values. Tolerance is the proportion of each variable's variance not shared with the other explanatory variables. Small tolerance values indicate collinearity. In general, we should ensure the tolerance is greater than 0.2. So these two variables do not have this problem.
Then we need to find the outliers. Studentized residuals greater than 3 or less than -3 should be deleted because they are outliers. For these 470 data, #15, #16, #17, #18, #19, #20, #21, #303, #304 and #305 should be deleted. In the second round, we delete #46, #292, #293, #294 and #365. In the third round, we delete #16, #47, #48, #49 and #50. In the fourth round, we delete #15 and #46. Finally, we have to delete 20 data. After that, we need to check for multicollinearity again.

Table 3.4: Checking for Multicollinearity (2)

Checking for Multicollinearity2

The REG Procedure
Model: MODEL1
Dependent Variable: Price

Number of Observations Read	450
Number of Observations Used	450

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7552577184	3776288592	952.58	<.0001
Error	447	1772035383	3964285		
Corrected Total	449	9324612567			

Root MSE	1991.05122	R-Square	0.8100
Dependent Mean	16331	Adj R-Sq	0.8091
Coeff Var	12.19210		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	7376.88545	353.78520	20.85	<.0001	.	0
Mileage	1	-0.12312	0.01153	-10.68	<.0001	0.99801	1.00199
Liter	1	3896.51493	91.13327	42.76	<.0001	0.99801	1.00199

Collinearity Diagnostics					
		Condition	Proportion of Variation		
Number	Eigenvalue	Index	Intercept	Mileage	Liter
1	2.82859	1.00000	0.00852	0.01698	0.01298
2	0.12531	4.75104	0.00729	0.67923	0.36132
3	0.04610	7.83343	0.98419	0.30379	0.62570

From the data in table 3.4, the tolerances of these two variables are both greater than 0.2, which indicates that these two variables have not this problem.
Then we use the stepwise method to get the equation. [1]

Table 3.5: Multiple Reression Models Using Stepwise Regression

Multiple Regression Models Using Stepwise Regressic

The REG Procedure
Model: MODEL1
Dependent Variable: Price

Number of Observations Read	450
Number of Observations Used	450

Stepwise Selection: Step 1

Variable Liter Entered: R-Square = 0.7615 and C(p) = 114.9650

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7100787483	7100787483	1430.49	<.0001
Error	448	2223825084	4963895		
Corrected Total	449	9324612567			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	5091.45449	315.17674	1295382498	260.96	<.0001
Liter	3853.15022	101.87649	7100787483	1430.49	<.0001

Table 3.6: Multiple Regresion Model Selection of Variables with the Stepwise Regression Approach

Stepwise Selection: Step 2

Variable Mileage Entered: R-Square = 0.8100 and C(p) = 3.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7552577184	3776288592	952.58	<.0001
Error	447	1772035383	3964285		
Corrected Total	449	9324612567			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	7376.88545	353.78520	1723580697	434.78	<.0001
Mileage	-0.12312	0.01153	451789701	113.96	<.0001
Liter	3896.51493	91.13327	7247089428	1828.09	<.0001

Bounds on condition number: 1.002, 4.008

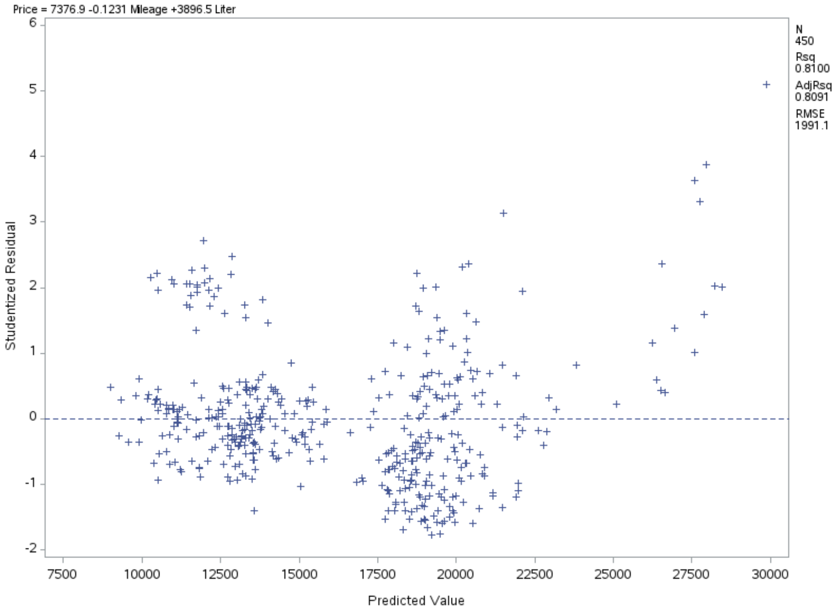
All variables left in the model are significant at the 0.1500 level.

All variables have been entered into the model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Liter		1	0.7615	0.7615	114.965	1430.49	<.0001
2	Mileage		2	0.0485	0.8100	3.0000	113.96	<.0001

In tables 3.4 and 3.5, the intercept, liter and mileage are significant because of p-value<0.001.
From table 3.6, we can get the equation:
Price=3896.51Liter-0.123Mileage+7376.89.
At last, we should check the variance.

Graph 3.2: Residual Plot (2)



Based on graph 3.2, it can meet the requirment of constant variance in general. So the best model is: **Price=3896.51Liter-0.123Mileage+7376.89.**

IV. CONCLUSION

This paperis about what factors would have influenced on the price of the cars. Initially we did the preliminary analysis and chose 7 variables to predict the price. Then we used the method of multiple linear regression to analyze how these factors affect the price of the cars. [3] After the analysis, we chose four variables(Type_c, Upgraded_Speakers_c,Mileage,Liter) to include in our model and there were too many outliers, and then we chose three variables(Mileage,Liter, Type_c) to include in our model.[4] We found that the amount of outliers is nearly same with two variables while we kept the Liter and Mileage in our model. Because the easiest model is the best, we chose two variables to predict. Finally , we got the equation that is

Price=3896.51Liter-0.123Mileage+7376.89

In which the coefficient of liter is positive and the coefficient of mileage is negative. At the same time, we can see that the abosolute value of liter coefficient is very big and the abosolute value of mileage coefficient is very small, which means that liter may have bigger influency on the price than mileage.

Once we have this question, I think mileage and liter are the most useful indexes to estimate the price of vehicle. At the same time, these two indexes have the most important influence on the price. When we are going to buy a car, we can put the information of mileage and liter into this regression model, and then we can get an estimate price to compare with the real price. Then we can decide whether to buy or not.

However, when we decide to buy a car, we should consider other factors that influence the price, including the season element, whether it has an accident or the times of repair, etc. All of these elements will have a large or small influence on the price. If we want to get a more precise model, many other factors should be considerate. It may take much time to collect the data and find the best regression model.

REFERENCES

[1]. A procedure for stepwise regression analysis (Articles & Statistical Papers, December 1992, Volume 33, Issue 1, pp 21-29
[2]. How to use SAS® to fit Multiple Logistic Regression Models Anpalaki J. Ragavan, Department of Mathematics and Statistics, University of Nevada, Reno, NV 89557(369-2008)
[3]. Introduction to Building a Linear Regression Model Leslie A. Christensen. The Goodyear Tire & Rubber Company, Akron Ohio.
[4]. Introduction to Linear Regression Analysis (By Douglas C. Montgomery, Elizabeth A. Peck, G.)
[5]. Kelley Blue Book. http://www.kbb.com
[6]. Multiple Regression: How Much Is Your Car Worth? George E. P. BoReliaSoft's Experiment Design and Analysis Reference
[7]. Regression with SAS (http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter1/sasreg1.htm)
[8]. SAS System for Regression (Third Edition, By Rudolf J. Freud, Ph.D., Ramon C. Littell, Ph.D.)
[9]. Statistics Using SAS Enterprise Guide (By James B. Davis, Ph.D.)
[10]. The Little SAS Book for Enterprise Guide 4.2 (By Susan J. Slaughter, Lora D. Delwiche)
[11]. Using Multivariate Statistics (FIFT H EDITION, Barbara G. Tabachnick California State University, Northridge Linda S. Fidell California State University, Northridge)
[12]. Special thanks to Dong Zhang for his help for this manuscript.