Car Value Analysis and Prediction

Group: Zhongce Ji, Jiamin Di, Zhengbang Pan, Minjun Li, Yuhan Duan

I. Introduction

I-1. Purpose of the report

The purpose of this report is to predict the price of 2005 used General Motors(GM) cars. The analysis method in this report will include regression method. The dataset in this report comes from KBB(Kelly Blue Book), we will use variables in this dataset to conduct a linear regression model and construct a prediction interval to achieve our purpose. Prediction car value is very important in the car markets, because the analysis of the car values can provide a comprehensive understanding for the car markets. For this project, we want to explore the determinant of car value. In our analysis, we will set Price as dependent variable and set Mileage, Make, Cylinder, Liter, Doors, Cruise, Sound, Leather as independent variables.

I-2. General background

Our dataset comes from KBB (Kelly Blue Book). KBB is a benchmark company for vehicle evaluation. All cars in this data set were less than one year old when priced and considered to be in excellent condition. A brief summary of the variable in the dataset will be recorded as following.

1. Car Price : Suggested retail price of the used 2005 GM car in excellent condition. The condition of a car can greatly affect price. This variable is quantitative variable and dependent variable.

2. Mileage : Number of miles the car has been driven. This variable is quantitative variable and independent variable.

3. Make : Manufacturer of the car. This is a category variable which has 6 different categories such as Buick, Cadillac, Chevrolet, Pontiac, SAAB and Saturn. This is also an independent variable. And we use 0 represents Buick, 1 as Cadillac, 2 as Chevrolet, 3 as Pontiac, 4 as SAAB and 5 as Saturn.

4. Model : Specific models for each car manufacturer such as Ion, Vibe, Cavalier. This is a category variable and independent variable. But we will delete this variable, because different makes have different model and there are so many categories that there is not enough comparability.

5. Trim : Specific type of car model such as SE Sedan 4D, Quad Coupe 2D. This is a category variable and independent variable. But we will delete this variable, because the trim is different in each model and there is so many different trims which does not have comparability.

6. Type : Body type such as sedan, coupe, etc. This is a category variable which has 5 categories including Sedan, Convertible, Hatchback, Coupe and Wagon. It is also an independent variable. And we use 0 represents sedan, 1 as Convertible, 2 as Hatchback, 3 as Coupe, 4 as Wagon.

7. Cylinder : Number of cylinders in the engine. This is a category variable which has 3 categories including 6, 8, 4. It is also an independent variable. And we use 0 represents 4 cylinders, 1 as 6 cylinders, 2 as 8 cylinders.

8. Liter : A more specific measure of engine size. This is a category variable which has 16 categories and independent variable. Because there is so many categories, we narrow it down to 5 categories. We group the liter which is higher than 1 and less or equal to 2 as 0, higher than 2 and less or equal to 3 as 1, higher than 3 and less or equal to 4 as 2, higher than 4 and less or equal to 5 as 3, higher than 5 and less or equal to 6 as 4.

9. Doors : Number of doors. This is a category variable and independent variable. There is 2 categories which including 2 and 4. And we use 0 represents 2 doors, and 1 as 4 doors.

10. Cruise : Indicator variable representing whether the car has cruise control. This is a category variable and independent variable. And we use 0 represents no cruise, and 1 as equipped with cruise.

11. Sound : Indicator variable representing whether the car has upgraded speakers. This is a category variable and independent variable. And we use 0 represents unupgraded speakers and 1 as upgraded speakers.

12. Leather : Indicator variable representing whether the car has leather seats. This is a category variable and independent variable. And we use 0 represents no leather equipped and 1 as leather equipped.

I-3. Discussion

Car is a part of our daily life. For most of us, we will consider practicality and comfort of a car before we buy it. However, the most important aspect is the value of a car. Car dealers usually pay you less when you trade in your used car and they will charge you more when you intend to buy a used one. So we want to predict car value

by independent variables in the dataset and tell the audience their predicted car price based on our model. As a result, we hope our prediction of price could be a good reference for potential buyers and sellers for a 2005 used car.

II. Data Analysis

II-1. Category Variable Pattern

In this part, we want to find the relationship between price and each category variable. Before people buy cars, they will consider many factors of cars. Some people may like high equipment but some people hope to buy cars with proper price including some category variables they like. We analyze each category variable patterns and try to find some rules for the customers to buy cars with moderate price.

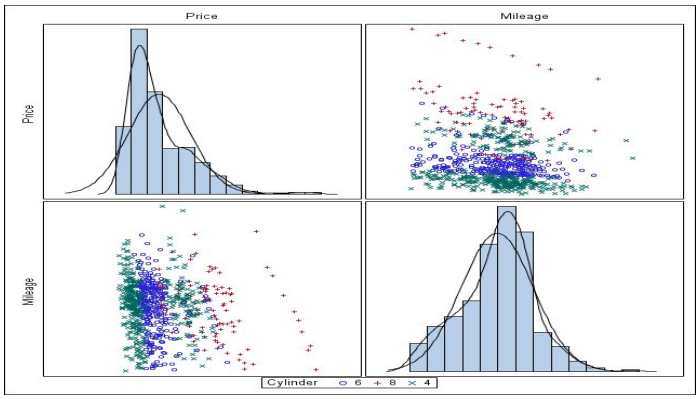

Figure 1 (Price VS Cylinder)

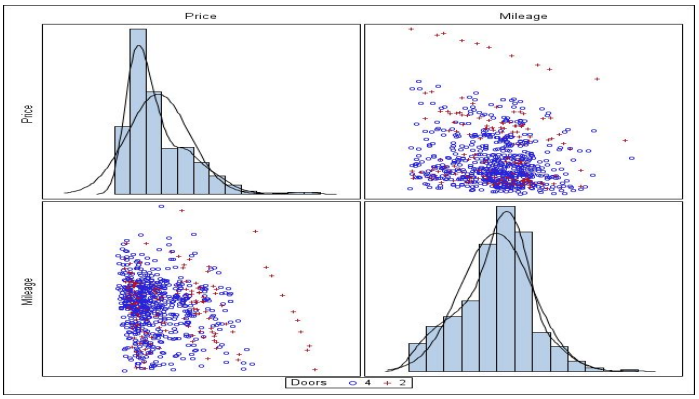

Figure 2 (Price VS Doors)

In the Figure 1,there are only three categories in the Cylinder Variable 4, 6, 8. From the plot, we can see that around the same level of mileage, cars with 8 cylinders are priced highest and with 4 cylinders are ranked lowest. There are some samples with 4 cylinders ranked in the middle, and we assume that is due to the makes of cars. Here we do not consider aspects of different make, so we will not include it in our cylinder analysis here.

In the Figure 2, there are only two different categories in the Doors Variable 4, 2. At the same level of mileage, price of 2 doors cars are wider than 4 doors cars. Without consideration of brands, 4-door cars take more market.

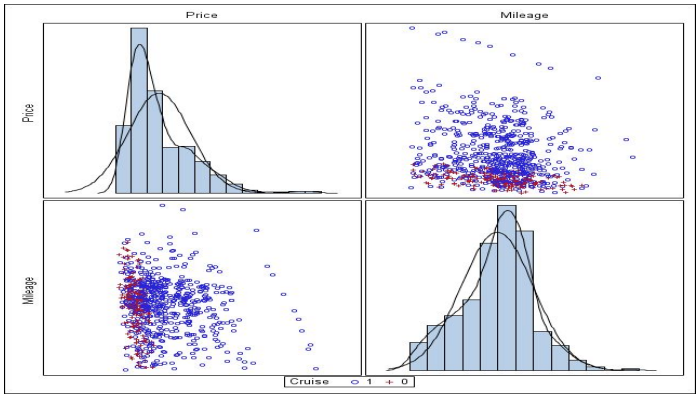

Figure 3 (Price VS Cruise)

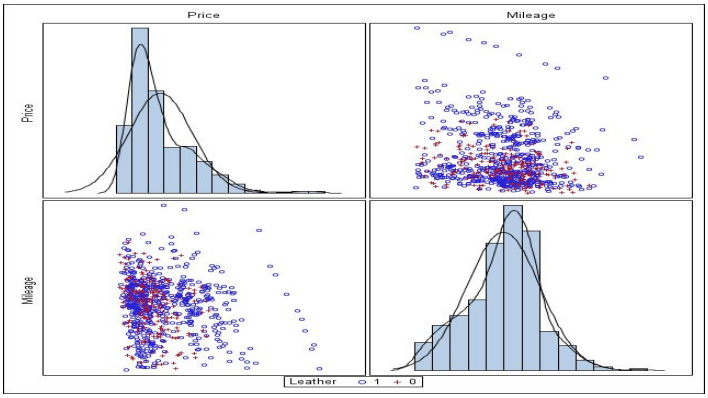

Figure 4 (Price VS Leather)

In the Figure 3, there are only two different categories in the Cruise, one is upgraded with cruise and the other one is not. At the same level of mileage, most of cars are more expensive with cruise than cars without cruise.

In the Figure 4, At the same level of mileage, cars with leather(marked in blue) have higher price range than car that without leather.

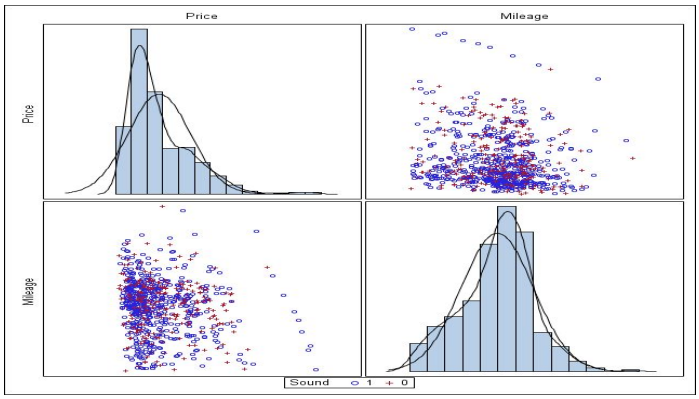

Figure 5 (Price VS Sound)

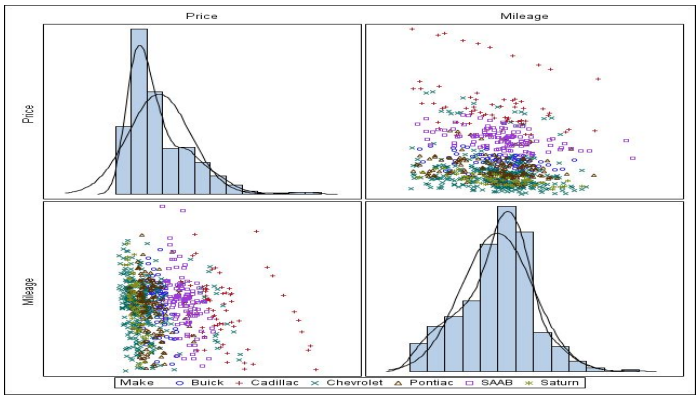

Figure 6 (Price VS Make)

In the Figure 5, from the plot, the difference between cars with or without sounds is not very clear. We can assume that sound is not a very important variable for price of cars. However, cars with sounds can price high in some particular cases related to other variable.

In the Figure 6, here we analyze influence of brands. From the plot, we can see that at same level of mileage, Cadillac price highest and Chevrolet price lowest, assuming all other variables are the same. Price from high to low: Cadillac, SAAB, Buick, Pontiac, Saturn and Chevrolet

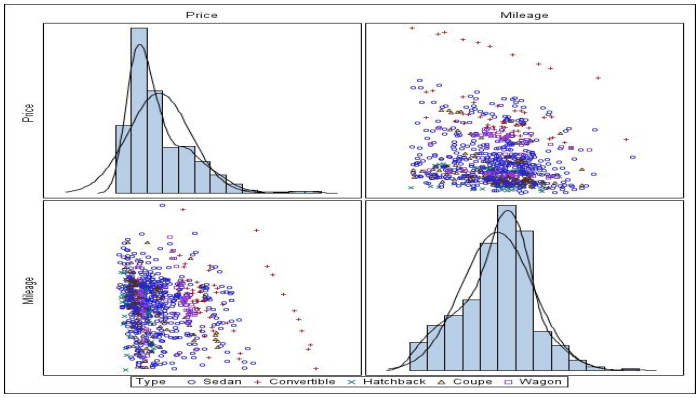

Figure 7 (Price VS Type)

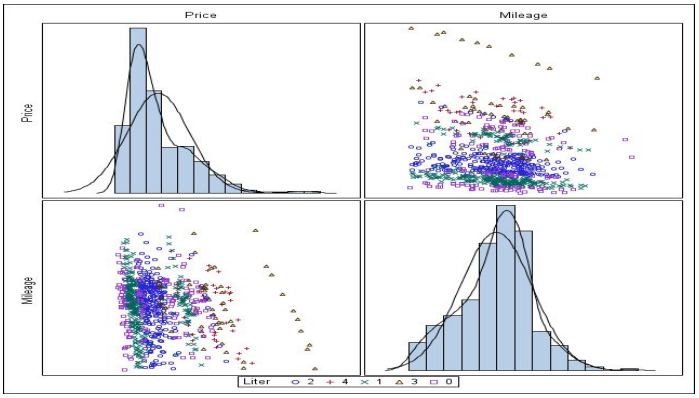

Figure 8 (Price VS Liter)

In the Figure 7, from the plot, we can find out that the range of the sedan is very wide and as we can see that the convertible seems higher than other type of car. While from the whole plot , it is hard to tell which is absolute higher than others.

In the Figure 8, from the plot, we can easily find that the price of Liter which is between 4 and 5 are higher than others, and the price of Liter which is between 5 and 6 are also higher than others but is smaller than some of Liter which is between 4 and 5. What's more, in the plot we also can find that the price range of Liter which is between 1 and 2 seems wider than others.

II-2. Quantitative Variable Analysis

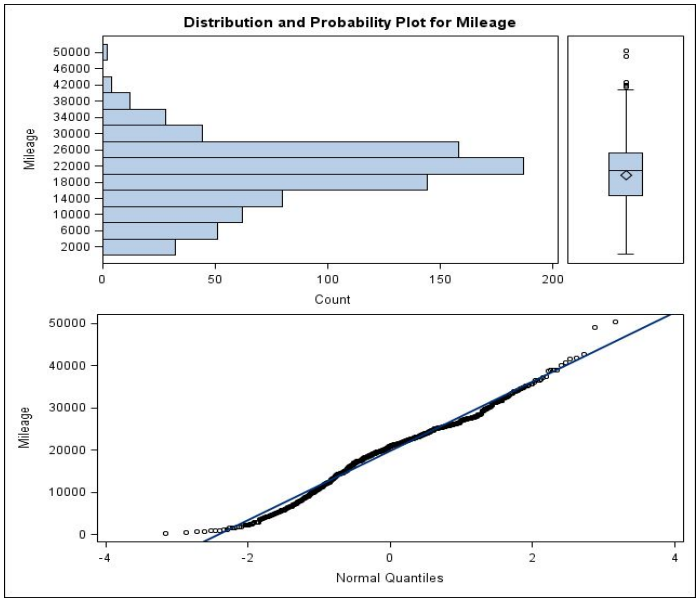

Figure 9 (Box-plot for Mileage)

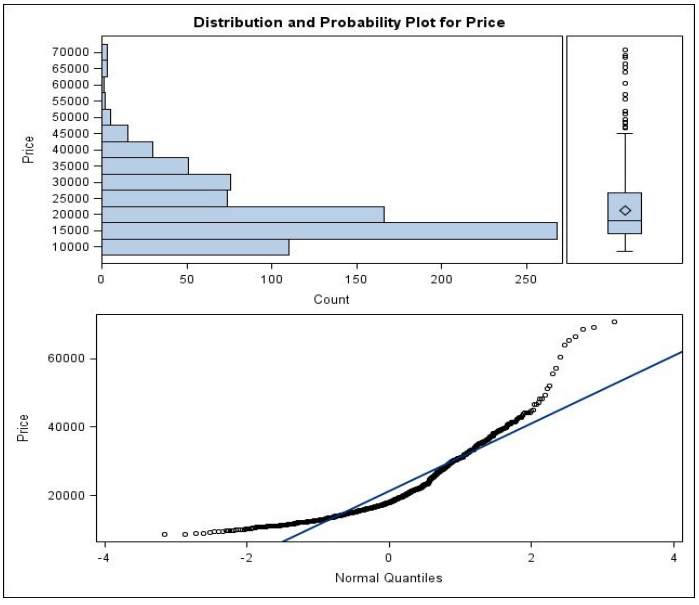

Figure 10 (Box-plot for Price)

In Figure 9 and 10, we draw the Box-plot for two quantitative variable: Mileage and Price. In Figure 9 which is Box-plot for Mileage, we find that the Mileage plot is skewed a little bit to the smaller side. Also there are some outliers which are larger than 42000. Then in Figure 10 which is Box-plot for Price, we find the plot is skewed much more to the smaller side than Mileage Box-plot. And there are many outliers in the large side. From our observation, most outliers in Price belong to the one make: Cadillac.

II-3. Correlation Analysis

Pearson Correlation Coefficients, N = 804
Prob > |r| under H0: Rho=0

	Price	Mileage
Price	1.00000	-0.14305
Price		<.0001
Mileage	-0.14305	1.00000
Mileage	<.0001	

Table 1(Correlation Coefficients table for Price vs Milage)

We do the correlation analysis for each quantitative variables, in the Pearson Correlation Coefficients table, the data shows the correlation between each variables including independent and dependent variables. If the correlation coefficients are close to 0 then we can conclude that two variable may have no correlation with each other. If the coefficients is close to 1 or -1, then we can conclude that they are highly correlated. In the table 1 which is correlation coefficients table for Price v.s. Mileage, we can find that the correlation coefficients between Price and Mileage is -0.14305. Hence we can conclude that the Price variable and Mileage variable are less correlated to each other.

III. Regression Analysis

III-1. Simple linear Regression Analysis

We first want to check the simple linear regression relationship. In here, we are seeking how much mileage will impact the price, so we take the mileage as independent variable and price as dependent variable.

We obtained the estimated regression and then we do residual analysis to check whether linear model is appropriate or not in this case.

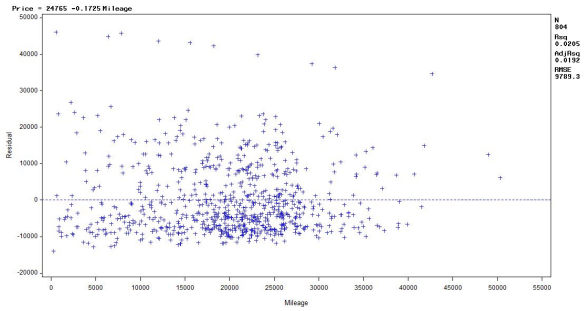

Figure 12 (residual vs mileage)

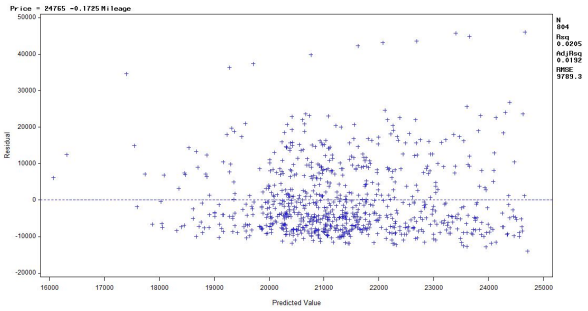

Figure 13 (residual vs predicted value)

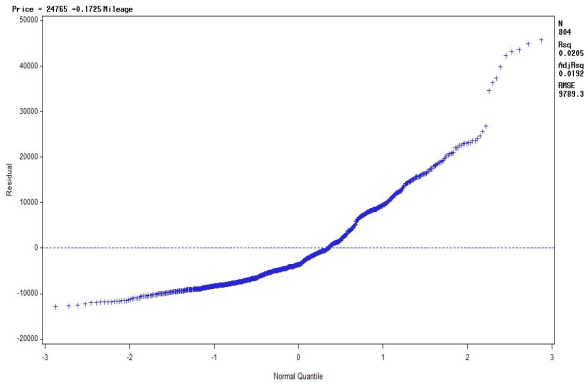

Figure 14 (Q-Q Plot for price)

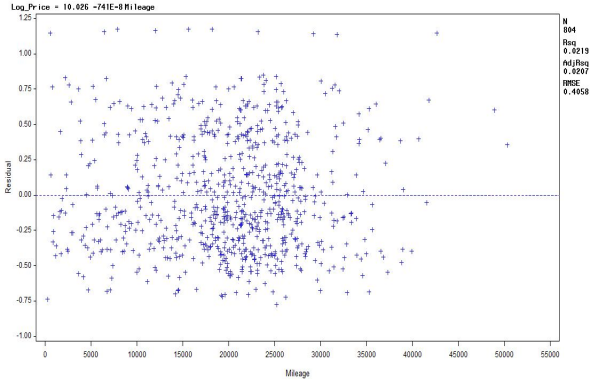

Figure 15 (residual vs mileage for log price)

From Figure 12 and 13, the residual plot for Mileage and Prediction values, we can see that there is no certain pattern in both residual plots, so we proceed to the Q-Q plot for more information. We find that for Q-Q plot, figure 14, all points are not lying in a straight line which implies residuals do not form a normal distribution, so we conclude that it is not appropriate to use this linear model to describe the relationship between Price and Mileage. Next we will try transformation to see if that will work better.

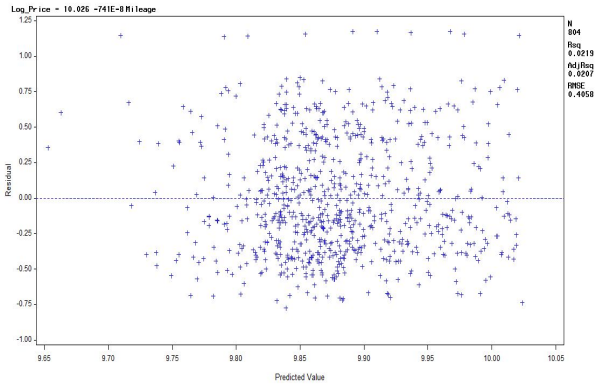

Figure 16 (residual vs predicted values for log price)

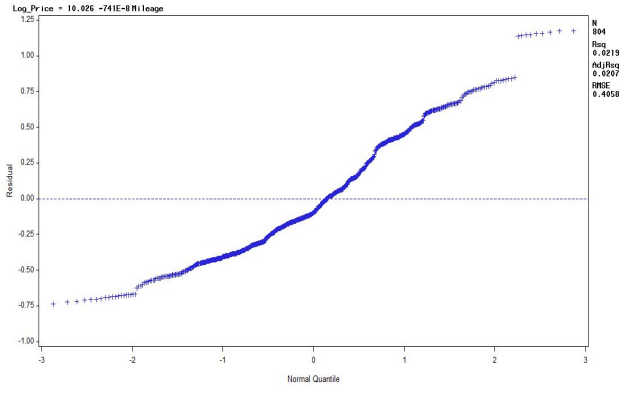

Figure 17 (Q-Q Plot for log price)

To compare results, we did two types transformations: log_price and inverse_price.

Firstly, we took logarithm of price. According to the Figure 15 and 16, the residual plots for Mileage and prediction log_price value, there is still no certain patterns in both residual plots. While from the Figure 17, the Q-Q plot, we can see that the points roughly form a straight line, so we can conclude that it is somehow appropriate to use linear model to describe the relationship between Log_price and Mileage. But we want to know if there is other better fitted regression, so we try the inverse of price.

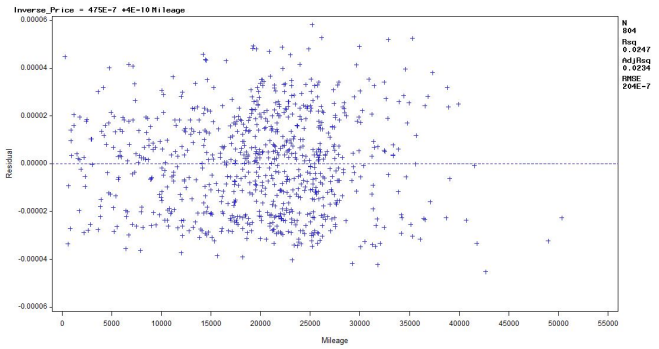

Figure 18 (Residual vs Mileage for inverse price)

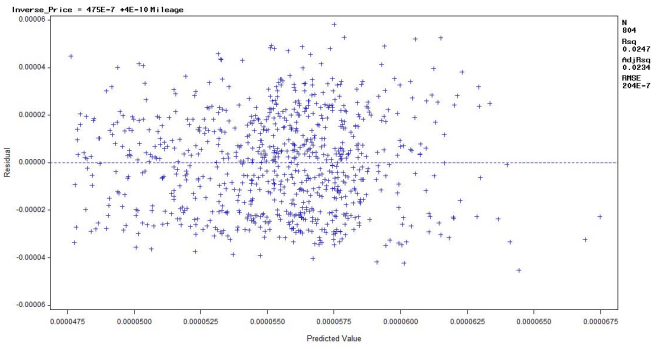

Figure 19 (Residual vs predict value for inverse price)

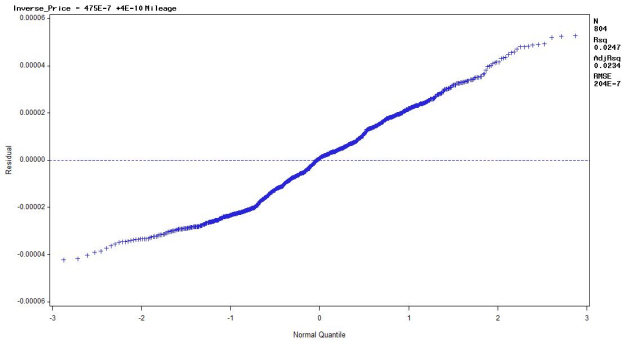

Figure 20 (Q-Q plot for inverse price)

Then, we took inverse of the price. According to the Figure 18 and 19 which is residual plot for the Mileage and residual plot for predict values, we can conclude that the there is no certain pattern in the plot. Then according to the Q-Q plot, the residual points in the plots roughly form a straight which means the residuals come from normal distribution, so we can roughly conclude a linear regression relation between inverse price and mileage is appropriate. And the line is more straighter than other Q-Q plot, so we think the inverse transformation for price will be better. And according to the

parameter estimate table, we have P-value smaller than a=0.05, so we can conclude an estimate regression function: $1/price = 0.00004752 + 3.96413 * 10^{-10} * Mileage$. So when Mileage increase by 1, the inverse Price will increase by $3.96413 * 10^{-10}$.

III-2. Multiple linear Regression Analysis

a. Residual analysis

In our analysis, most independent variables are category variables. For the variable Make, Type, Cylinder, Liter, those category variables have more than 2 categories, so we need to create indicators(dummy variables) for each categories in each category variables. After we create the indicators, we have the multiple linear model is : Inverse Price ~ Mileage + Make_Buick + Make_Cadillac + Make_Chevrolet + Make_Pontiac + Make_SAAB + Make_Saturn + Type_Sedan + Type_Convertible + Type_Hatchback + Type_Coupe + Type_Wagon + Cylinder_4 + Cylinder_6 + Cylinder_8 + Liter1_2 + Liter2_3 + Liter3_4 + Liter4_5 + Liter5_6 + Doors + Cruise + Sound + Leather.

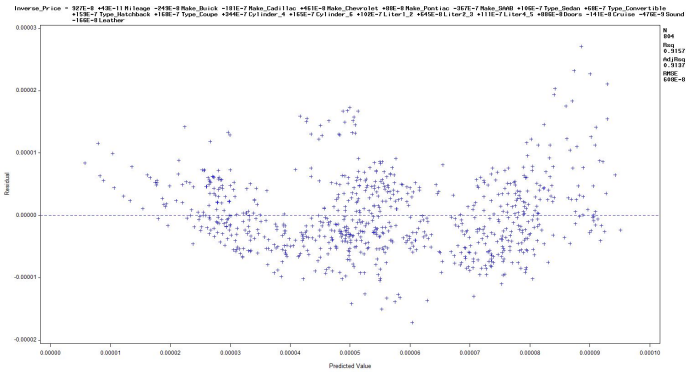

Figure 21 (Residual plot vs Predict value for Multiple)

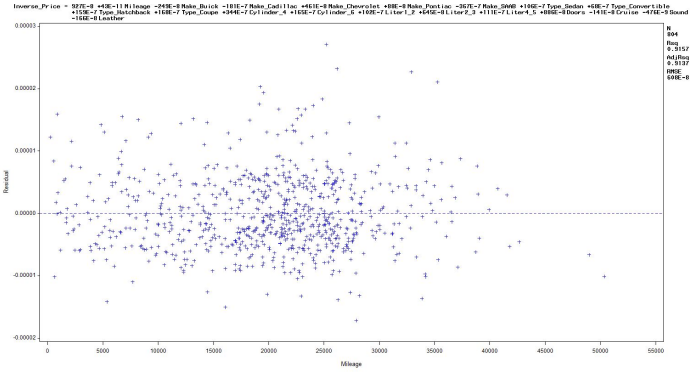

Figure 22 (Residual plot vs Mileage for Multiple)

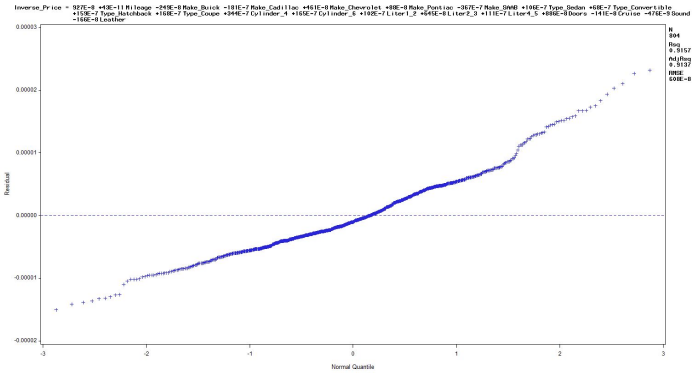

Figure 23 (Q-Q plot for Multiple)

First we will do the residual analysis for the multiple linear regression model. Because except for Mileage variables, other variable is indicators so we will not show the residual plot. In the Figure 21 which is residual plot vs predicted value and Figure 22 which is residual plot vs Mileage, it is hard to find a certain pattern in both plot. Then we take look at Q-Q plot, we can find that the residual points roughly form an approximate straight line, so we can conclude that the residual points may come from normal distribution and the linear relation between independent variables and dependent variables is appropriate.

b. Best Model Selection

After we analyze the residual for the regression line, we need to find out the best variable subset for the models. We will use four criterions including adjusted R square, C(p), AIC(P) and SBC(P) to select the best models.

Adjusted R square is use to consider the trade-off between the # of predictors and # of observations. A good subset of predictors should has the maximum value or close to the maximum. C(P) is use to consider the total mean squared error for the n fitted value. A good subset of predictors should have C(P) close to p which is the # of coefficients. For both AIC and SBC, we are trying to find the subset that has small values.

17	0.9156	0.9138	16.8010	-19296.432	-19212	Mileage Make_Buick Make_Cadillac Make_Chevrolet Make_SAAB Type_Sedan Type_Convertible Type_Wagon Cylinder_4 Cylinder_6 Liter1_2 Liter3_4 Liter5_6 Doors Cruise Sound Leather

Table 2 (best model selection result 17(1))

R-Square Selection Method

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	SBC	Variables in Model
17	0.9156	0.9138	16.8010	-19296.432	-19212	Mileage Make_Buick Make_Cadillac Make_Chevrolet Make_SAAB Type_Sedan Type_Convertible Type_Wagon Cylinder_4 Cylinder_8 Liter1_2 Liter2_3 Liter4_5 Doors Cruise Sound Leather
17	0.9156	0.9138	16.8010	-19296.432	-19212	Mileage Make_Buick Make_Cadillac Make_Chevrolet Make_SAAB Type_Sedan Type_Convertible Type_Wagon Cylinder_4 Cylinder_8 Liter1_2 Liter2_3 Liter5_6 Doors Cruise Sound Leather
17	0.9156	0.9138	16.8010	-19296.432	-19212	Mileage Make_Buick Make_Cadillac Make_Chevrolet Make_SAAB Type_Sedan Type_Convertible Type_Wagon Cylinder_4 Cylinder_8 Liter1_2 Liter3_4 Liter4_5 Doors Cruise Sound Leather
18	0.9157	0.9137	18.2700	-19294.976	-19206	Mileage Make_Buick Make_Cadillac Make_Chevrolet Make_SAAB Make_Saturn Type_Sedan Type_Convertible Type_Wagon Cylinder_6 Liter1_2 Liter2_3 Liter3_4 Liter4_5 Doors Cruise Sound Leather
18	0.9157	0.9137	18.2700	-19294.976	-19206	Mileage Make_Buick Make_Chevrolet Make_Pontiac Make_SAAB Make_Saturn Type_Sedan Type_Convertible Type_Wagon Cylinder_4 Liter1_2 Liter2_3 Liter3_4 Liter4_5 Doors Cruise Sound Leather
18	0.9157	0.9137	18.2700	-19294.976	-19206	Mileage Make_Buick Make_Cadillac Make_Pontiac Make_SAAB Make_Saturn Type_Sedan Type_Convertible Type_Wagon Cylinder_4 Liter1_2 Liter2_3 Liter3_4 Liter4_5 Doors Cruise Sound Leather

Table 3 (best model selection result 17 and 18)

R-Square Selection Method

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	SBC	Variables in Model
18	0.9157	0.9137	18.2700	-19294.976	-19206	Mileage Make_Buick Make_Chevrolet Make_Pontiac Make_SAAB Make_Saturn Type_Sedan Type_Convertible Type_Wagon Cylinder_4 Cylinder_6 Liter1_2 Liter2_3 Liter4_5 Doors Cruise Sound Leather
19	0.9157	0.9137	20.0000	-19293.253	-19199	Mileage Make_Buick Make_Chevrolet Make_Pontiac Make_SAAB Make_Saturn Type_Sedan Type_Convertible Type_Hatchback Type_Coupe Cylinder_4 Liter1_2 Liter2_3 Liter3_4 Liter4_5 Doors Cruise Sound Leather
19	0.9157	0.9137	20.0000	-19293.253	-19199	Mileage Make_Buick Make_Chevrolet Make_Pontiac Make_SAAB Make_Saturn Type_Sedan Type_Convertible Type_Hatchback Type_Wagon Cylinder_4 Liter1_2 Liter2_3 Liter3_4 Liter4_5 Doors Cruise Sound Leather
19	0.9157	0.9137	20.0000	-19293.253	-19199	Mileage Make_Buick Make_Cadillac Make_Chevrolet Make_Pontiac Make_Saturn Type_Convertible Type_Hatchback Type_Coupe Type_Wagon Cylinder_6 Liter1_2 Liter2_3 Liter3_4 Liter4_5 Doors Cruise Sound Leather
19	0.9157	0.9137	20.0000	-19293.253	-19199	Mileage Make_Buick Make_Cadillac Make_Chevrolet Make_Pontiac Make_Saturn Type_Convertible Type_Hatchback Type_Coupe Type_Wagon Cylinder_4 Cylinder_6 Liter1_2 Liter2_3 Liter4_5 Doors Cruise Sound Leather

Table 4 (best model selection result 18 and 19)

According to the selection rule of the four criterions, we have selected the results based on each criterions. And the we will take the common models that each results have. Then selection we have is below:

For adj R: 0.9138: 17(1), 17(2), 17(3), 17(4),

0.9137: 18(1), 18(2), 18(3), 18(4), 19(1), 19(2), 19(3), 19(4)

For C(p): 20: 19(1), 19(2), 19(3), 19(4)

17: 17(1), 17(2), 17(3), 17(4),

For AIC: -19297.720 : 16(1), 16(2), 16(3), 16(4),

-19296.432 : 17(1), 17(2), 17(3), 17(4),

From the results, we can find the common subset of predictors is 17(1), 17(2), 17(3), 17(4). Then we will do the residual plot analysis and check vif for each variables in each models to see which model is better.

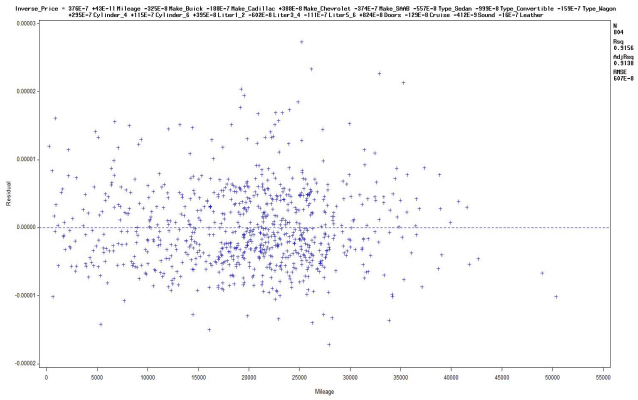

Figure 24 (Residual vs Mileage for model 17(1))

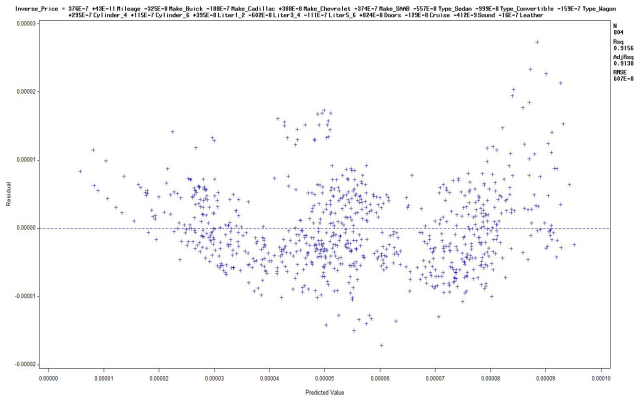

Figure 25 (Residual vs predict value for model 17(1))

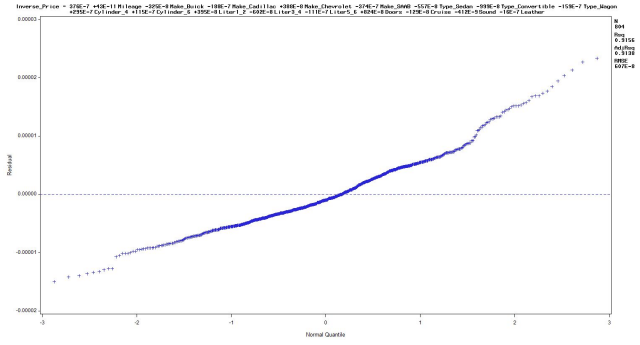

Figure 26 (Q-Q plot for model 17(1))

Table 5 (VIF table for model 17(1))

For the subset 17(1), we have predictors Mileage Make_Buick Make_Cadillac Make_Chevrolet Make_SAAB Type_Sedan Type_Convertible Type_Wagon Cylinder_4 Cylinder_6 Liter1_2 Liter3_4 Liter5_6 Doors Cruise Sound Leather. And we do the residual plot analysis and vif for the best predictors subset 17(1).

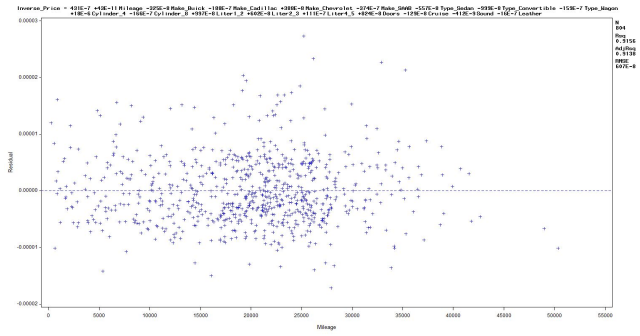

Figure 27 (Residual vs Mileage for Model 17(2))

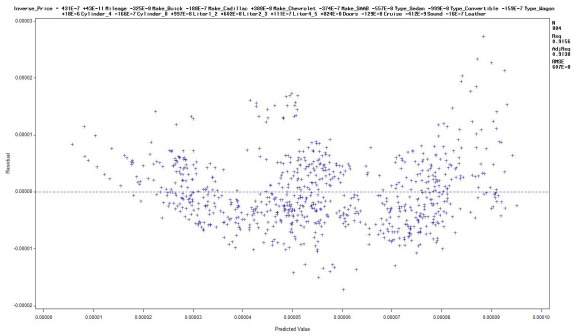

Figure 28 (Residual vs Predict value for Model 17(2))

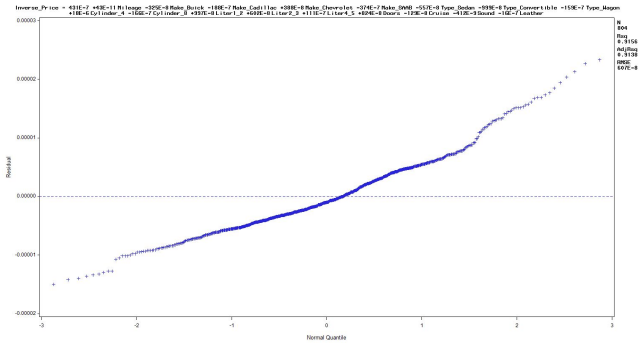

Figure 29 (Q-Q plot for model 17(2))

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	0.00004312	0.00000115	37.56	<.0001	0
Mileage	Mileage	1	4.27625E-10	2.62891E-11	16.27	<.0001	1.01022
Make_Buick		1	-0.00000325	8.855859E-7	-3.67	0.0003	1.53090
Make_Cadillac		1	-0.00001880	0.00000114	-16.44	<.0001	2.55420
Make_Chevrolet		1	0.00000388	5.995531E-7	6.48	<.0001	1.87632
Make_SAAB		1	-0.00003737	9.187331E-7	-40.67	<.0001	2.23764
Type_Sedan		1	-0.00000557	7.541807E-7	-7.38	<.0001	2.94940
Type_Convertible		1	-0.00000999	0.00000121	-8.24	<.0001	1.86659
Type_Wagon		1	-0.00001591	0.00000115	-13.81	<.0001	2.12013
Cylinder_4		1	0.00001801	0.00000161	11.17	<.0001	14.14812
Cylinder_8		1	-0.00001655	0.00000115	-14.40	<.0001	3.13525
Liter1_2		1	0.00000997	0.00000171	5.82	<.0001	9.70224
Liter2_3		1	0.00000602	0.00000153	3.93	<.0001	11.26077
Liter4_5		1	0.00001106	0.00000143	7.75	<.0001	3.06362
Doors	Doors	1	0.00000824	8.448225E-7	9.75	<.0001	3.09046
Cruise	Cruise	1	-0.00000129	6.134396E-7	-2.10	0.0364	1.52690
Sound	Sound	1	-4.11736E-7	4.934715E-7	-0.83	0.4043	1.15610
Leather	Leather	1	-0.00000160	5.561347E-7	-2.88	0.0041	1.34677

Table 6 (VIF table for model 17(2))

For the subset 17(2), we have predictors Mileage Make_Buick Make_Cadillac Make_Chevrolet Make_SAAB Type_Sedan Type_Convertible Type_Wagon Cylinder_4 Cylinder_8 Liter1_2 Liter2_3 Liter4_5 Doors Cruise Sound Leather. And we do the residual plot analysis and vif for the best predictors subset 17(2)

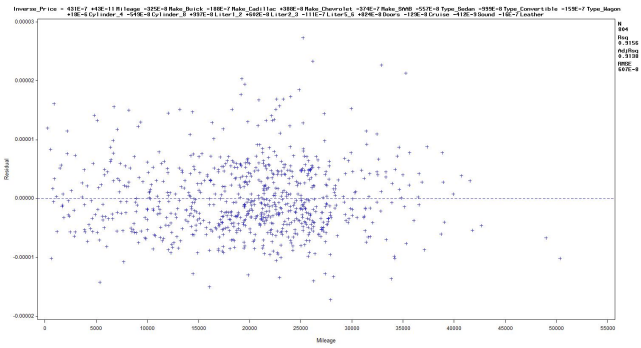

Figure 30 (Residual vs Mileage for model 17(3))

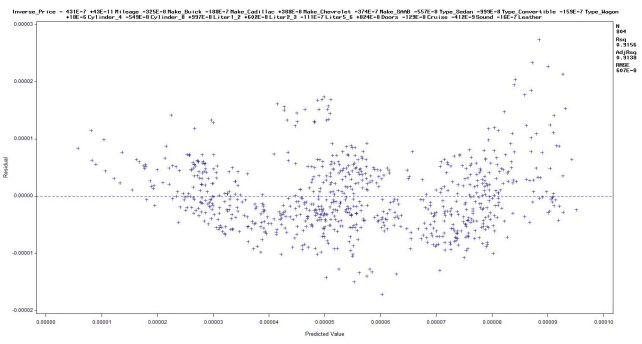

Figure 31 (Residual vs Predict value for model 17(3))

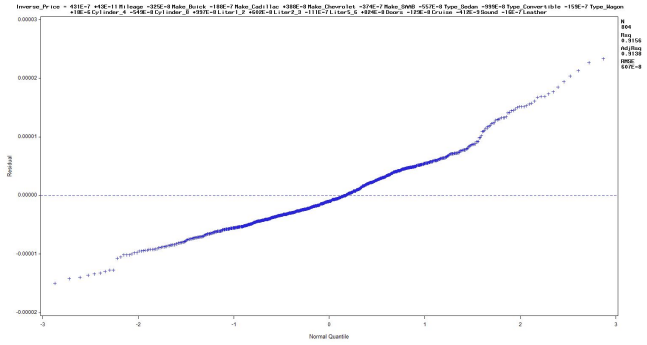

Figure 32 (Q-Q plot for the model 17(3))

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	3.14729E-7	1.851347E-8	501.65	<.0001
Error	786	2.900743E-8	3.69051E-11		
Corrected Total	803	3.437364E-7			

Root MSE	0.00000607	R-Square	0.9156
Dependent Mean	0.00005538	Adj R-Sq	0.9138
Coeff Var	10.96978		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	0.00004312	0.00000115	37.56	<.0001	0
Mileage	Mileage	1	4.27625E-10	2.62891E-11	16.27	<.0001	1.01022
Make_Buick		1	-0.00000325	8.855859E-7	-3.67	0.0003	1.53090
Make_Cadillac		1	-0.00001880	0.00000114	-16.44	<.0001	2.55420
Make_Chevrolet		1	0.00000388	5.995531E-7	6.48	<.0001	1.87632
Make_SAAB		1	-0.00003737	9.187331E-7	-40.67	<.0001	2.23764

Table 7 (VIF table for model 17(3))

For the subset 17(3), we have predictors Mileage Make_Buick Make_Cadillac Make_Chevrolet Make_SAAB Type_Sedan Type_Convertible Type_Wagon Cylinder_4 Cylinder_8 Liter1_2 Liter2_3 Liter5_6 Doors Cruise Sound Leather. And we do the residual plot analysis and vif for the best predictors subset 17(3).

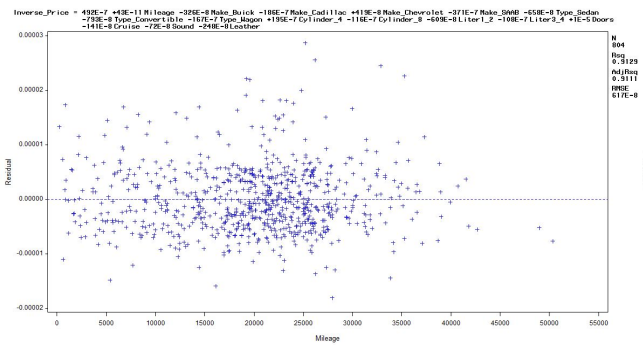

Figure 33 (Residual vs Mileage for model 17(4))

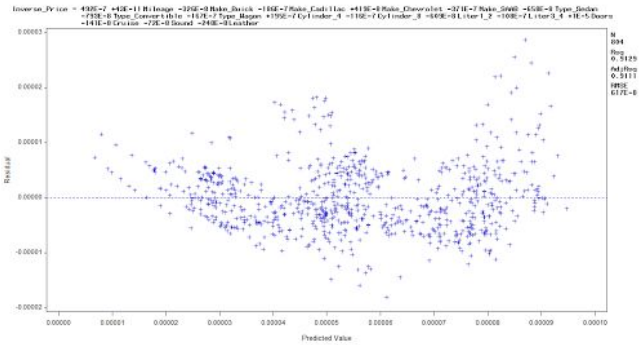

Figure 34 (Residual vs Predict value for model 17(4))

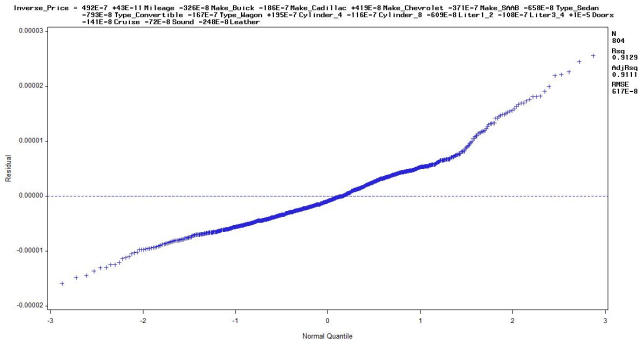

Figure 35 (Q-Q plot for model 17(4))

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	0.00004914	0.00000182	27.02	<.0001	0
Mileage	Mileage	1	4.27625E-10	2.62891E-11	16.27	<.0001	1.01022
Make_Buick		1	-0.00000325	8.855859E-7	-3.67	0.0003	1.53090
Make_Cadillac		1	-0.00001880	0.00000114	-16.44	<.0001	2.55420
Make_Chevrolet		1	0.00000388	5.995531E-7	6.48	<.0001	1.87632
Make_SAAB		1	-0.00003737	9.187331E-7	-40.67	<.0001	2.23764
Type_Sedan		1	-0.00000557	7.541807E-7	-7.38	<.0001	2.94940
Type_Convertible		1	-0.00000999	0.00000121	-8.24	<.0001	1.86659
Type_Wagon		1	-0.00001591	0.00000115	-13.81	<.0001	2.12013
Cylinder_4		1	0.00001801	0.00000161	11.17	<.0001	14.14812
Cylinder_8		1	-0.00002257	0.00000176	-12.84	<.0001	7.32922
Liter1_2		1	0.00000395	7.829532E-7	5.04	<.0001	2.02674
Liter3_4		1	-0.00000602	0.00000153	-3.93	<.0001	11.77420
Liter4_5		1	0.00001106	0.00000143	7.75	<.0001	3.06362
Doors	Doors	1	0.00000824	8.448225E-7	9.75	<.0001	3.09046
Cruise	Cruise	1	-0.00000129	6.134396E-7	-2.10	0.0364	1.52690
Sound	Sound	1	-4.11736E-7	4.934715E-7	-0.83	0.4043	1.15610
Leather	Leather	1	-0.00000160	5.561347E-7	-2.88	0.0041	1.34677

Table 8 (VIF table for model 17(4))

For the subset 17(4), we have predictors Mileage Make_Buick Make_Cadillac Make_Chevrolet Make_SAAB Type_Sedan Type_Convertible Type_Wagon Cylinder_4

Cylinder_8 Liter1_2 Liter3_4 Liter4_5 Doors Cruise Sound Leather. And we do the residual plot analysis and vif for the best predictors subset 17(4).

Then we compared the residual plot and Q-Q plot for both three models, we find that there is no certain pattern in each residual plot and the residual plot for three model is quite similar. What's more for the Q-Q plot, we can also find in three Q-Q plot the residual points roughly form a straight line and the line is pretty similar for each models. So it is hard to tell which model is better from residual plots. Then we turn to VIF table,

VIF is variance inflation factor, which represents how much the variance of estimated coefficients are inflated when the predictors are correlated with each others. And the rule is when VIF is equal to 1 then the predictor are not correlated with other predictors. If the VIF is greater than 1 and less than 5, then the predictor are moderately correlated with other predictors. And if the VIF is greater than 5, then predictor are highly correlated with other predictors. Compared with four VIF tables, we find that the the sum of VIF for model 17(1) is 63.40897, model 17(2) is 64.57563, model 17(3) is 64.38238 and model 17(4) is 61.60753. So the model 17(4) has smallest VIF value. So we can conclude that the predictors in fourth model is less correlated with other predictors in other models. So we have the best model is Mileage Make_Buick Make_Cadillac Make_Chevrolet Make_SAAB Type_Sedan Type_Convertible Type_Wagon Cylinder_4 Cylinder_8 Liter1_2 Liter3_4 Liter4_5 Doors Cruise Sound Leather. Then according to the table 8, we have the estimated regression function is :

$$1/Price = 4.914 * 10^{-5} + 4.27625 * 10^{-10} Mileage - 0.00000325 Make\_Buick - 0.00001880 Make\_Cadillac$$
$$+ 0.00000388 Make\_Chevrolet - 0.00003737 Make\_SAAB - 0.00000557 Type\_Sedan -$$
$$0.00000999 Type\_Convertible - 0.00001591 Type\_Wagon + 0.00001801 Cylinder\_4 -$$
$$0.00002257 Cylinder\_8 + 0.00000395 Liter1\_2 - 0.00000602 Liter3\_4 + 0.00001106 Liter4\_5 +$$
$$0.00000824 Doors - 0.00000129 Cruise - 4.11736 * 10^{-7} Sound - 0.00000160 Leather + e$$

IV. Conclusion

Interpret Function

For our multiple linear estimated regression function, the model means that when all other independent variables constant, an unit increases in made by Buick have a 0.00000325 decrease in the inverse of price. Similarity, an unit increases in made by Cadillac has a 0.00001880 decrease in the inverse of price. For the cars make by Chevrolet, an unit has a 0.00000388 decrease in the inverse of price. For the cars make by SAAB, an unit has a 0.00003737 decrease in the inverse of price. For the cars which have sedan type, an unit has a 0.00000557 decrease in the inverse of price. For the cars which have convertible type, an unit has a 0.00000999 decrease in the inverse of price. For the cars which have wagon type, an unit has a 0.00001591 decrease in the inverse of price. For the cars which have 4 cylinders, an unit has a 0.00001801 increase in the inverse of price. For the cars which have 8 cylinders, an unit has a 0.00002257 decrease in the inverse of price. For the cars which have 1 to 2 liter, an unit has a 0.00000395 increase in the inverse of price. For the cars which have 3 to 4 liter, an unit has a 0.00000602 decrease in the inverse of price. For the cars which have 4 to 5 liter, an unit has a 0.00001106 increase in the inverse of price. For the cars which have doors, an unit has a 0.00000824 increase in the inverse of price. For the cars which have cruise, an unit has a 0.00000129 decrease in the inverse of price. For the car, increase one unit in Leather will decrease 0.00000160 in the inverse of Price. And if we increase one unit in Sound, the inverse of price will also decrease by $4.11736 * 10^{-7}$. For the Cruise predictor, if we increase one unit in it, the inverse of car price will also decrease by 0.00000129. While for the Doors, if we increase one unit in the Doors variable, the inverse of car price will increase by 0.00000824. Then we put them together, if we have a 4 door car which is made by Buick and it is a Sedan with 4 Cylinder and 1.5 liter. And the Equipments the car have is Cruise control, upgraded Sound system and Leather seat. The car has been driven for 20000 mileages. Then the average inverse price will be $7.5771 * 10^{-5}$. So the average price will be 13197.70248.

Limitation

This report does not manage to cover any prediction for the price of used cars that is not a 2005 used one. Additionally, since used car price may change as time changes, the prediction of car price in this report will be effective only the same time as the dataset was created. Inflation and other factors may influence the price of 2005 used cars after the dataset was created, as a result, the prediction may not match the price now for 2005 used cars.

Another bias in this report will be locations because price of the same car varies in different locations. In the dataset, it does not provide either locations or zip code in variables. Therefore, there is no evidence to show where these cars were and our prediction may be accurate only if the 2005 used cars were at random place.