Multiple Regression Analysis for the Blue Book Cost of a Used Car: SAS Application

Natalia Irena Gust-Bardon

December, 2018

1. Problem Description

The purpose of this project is to obtain a prediction equation for the price of used cars. The data set consists of 630 observations, one dependent variable (price) and the following 11 predictor variables:

(1) mileage: number of miles the car has been driven
(2) make: manufacturer of the car (Buick, Cadillac, Chevrolet, and Pontiac)
(3) model: specific model for each car
(4) trim: specific type of a car model
(5) type: body type (e.g. sedan, wagon)
(6) cylinder: number of cylinders in the engine (4, 6 or 8)
(7) liter: measure of the engine size
(8) doors: number of doors: 2, 4
(9) cruise control: 1=yes, 0=no
(10) sound: upgrader speakers (1=yes, 0=no)
(11) leather seats: 1=yes, 0=no

Before starting the entire process of model building, I have made the following steps:

The TRIM predictor variable comprises 38 categories. Therefore, I have decided to group them together into three following levels:

A. Basic: basic equipment
B. Entry: basic equipment plus some special features
C. High: upgraded equipment and additional features.

The MODEL predictor variable comprises 25 categories. Thus, I have decided to group them together into three following classes:

A. Regular Class: up to 20.000 USD
B. Middle Class: up to 30.00 USD
C. Upper Class: starting from 30.000 USD.

This report presents the set of activities allowing me to build a multivariate regression model, including:

- checking for the violations of model assumptions;
- data transformation;
- variable selection techniques;
- incorporation of categorical explanatory variables;
- application of *F*-tests.

After the entire process of model building, I was able to develop the prediction equation that explains about 99% of the variability in new data.

2. Investigation of the Data

The data set consists of four quantitative variables (Mileage, Cylinder, Liter, and Doors) and seven categorical variables (Make, Model, Trim, Type, Cruise, Sound, Leather).

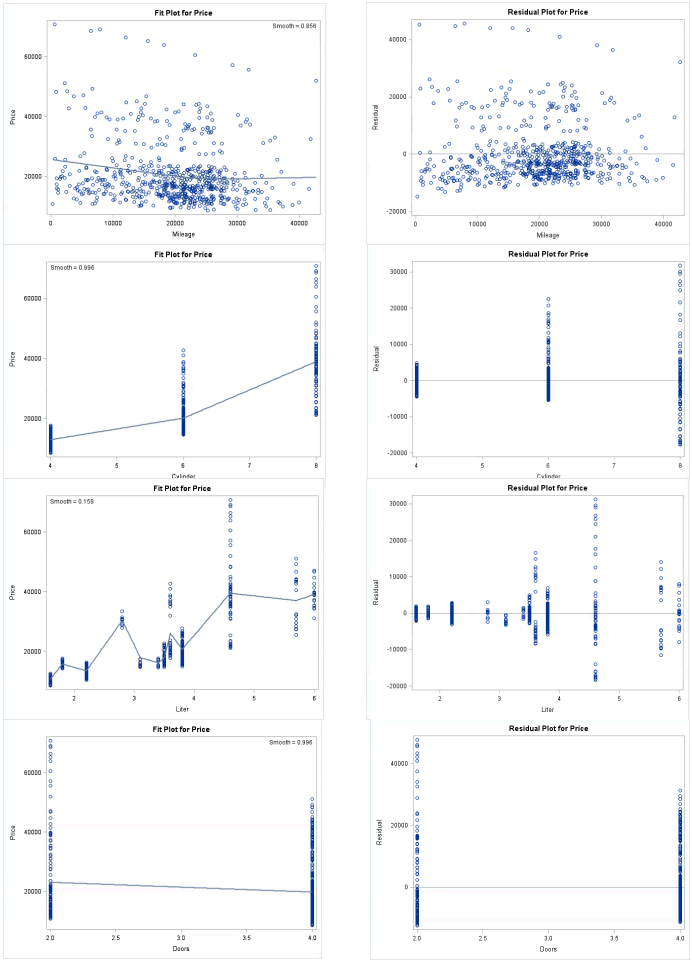


Figure 2.1 Quantitative variables

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
Price	Price	630	20569.46	10215.42	8638.93	70755.47
Mileage	Mileage	630	19579.08	8124.58	266.0000000	42691.00
Cylinder	Cylinder	630	5.5873016	1.3885166	4.0000000	8.0000000
Liter	Liter	630	3.2650794	1.1423567	1.6000000	6.0000000
Doors	Doors	630	3.5555556	0.8321401	2.0000000	4.0000000

Table 2.1 Summary statistics: Quantitative variables

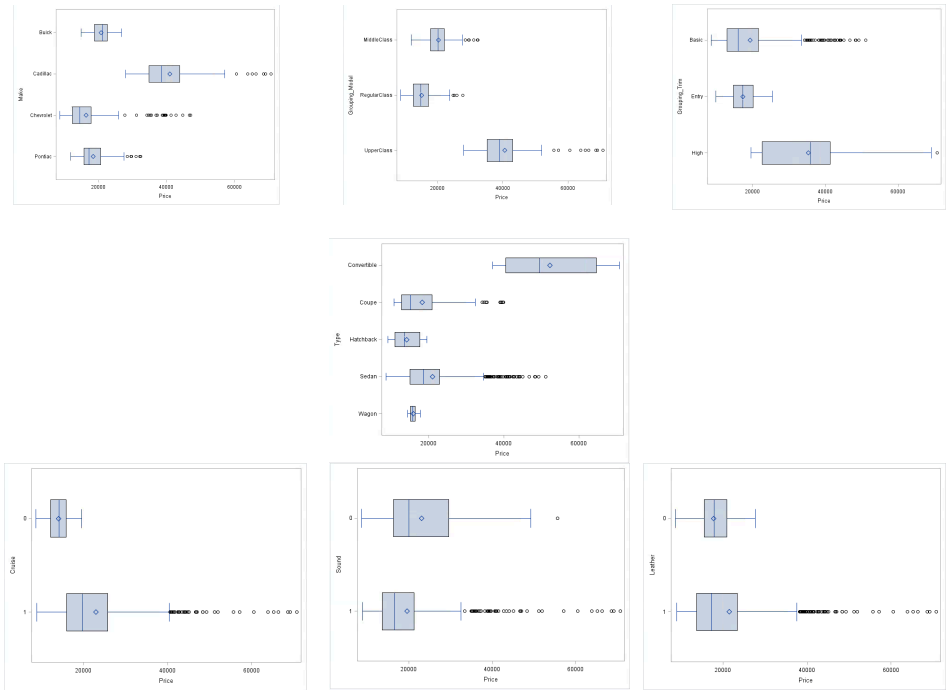


Figure 2.2 Categorical variables

Based on the investigation of the data, I do not see a need for any preliminary transformations. I would rather consider adding some interactions to the model (e.g. Make and Trim, or Make and Model).

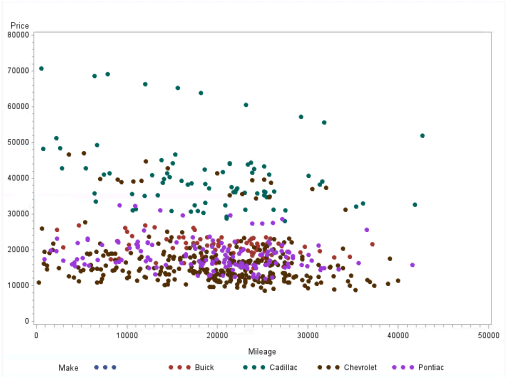
Based on **Figure 2.1** presenting the plot of Price and Mileage, we can see 10 observations (top of the plot) that look like possible outliers. I presume that majority of them come from Cadillac category of Make variable. When it comes to the plot of Price and Cylinder, we can notice outliers for engines with 6 and 8 cylinders. Here, I also assume that most of them come from Cadillac category. Looking at the plot of Price and Liter, we see that most of the outliers are from the engine size of 3.6, 4.6, and 5.7 liters. This, in my opinion is also connected with Cadillac category. The last plot of Price and Doors displays 11 observations that should be carefully scrutinized as possible outliers. In this case, I also believe that they come from Cadillac category.

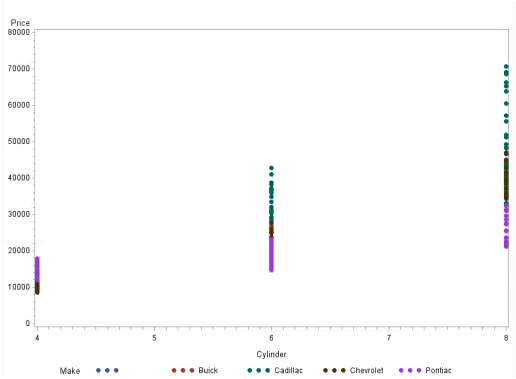
Table 2.1 with summary statistics shows that the range of price if from 8638.93 to 70755.47, and the range of mileage from 266 to 42691. We can conclude that, we have both the luxurious and regular cars in the data set. Looking at the range of mileage we see that cars differ from slightly used (266 miles) to significantly used (42691).

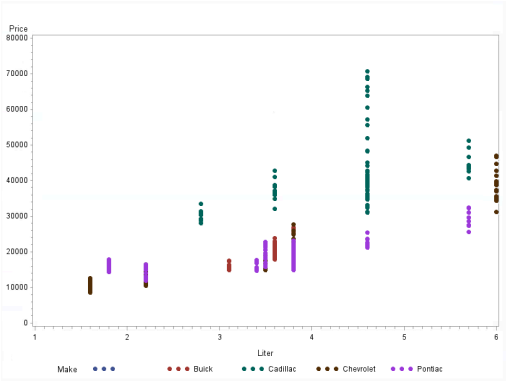
Based on **Figure 2.2** with categorical variables, we can observe outliers in the plot of Price and Make concerning three out of four categories: Cadillac, Chevrolet, and Pontiac. The plot of Price and Trim shows that the outliers appear in the basic level of equipment. One of the reason for this is because for example the trim of Sedan 4D has been classified as a basic level of equipment; however Sedan 4D made by Chevrolet or Buick has a different price range in comparison to Sedan 4D made by Cadillac. Hence, adding interaction term to the model (Make and Trim) seems to be

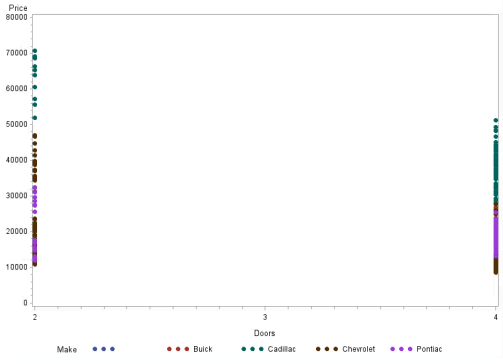
reasonable. The same issue regards the plot of Price and Body Type where Sedan has majority of outliers. When it comes to Cruise, Sound, and Leather variables, possible outliers occur for category 1 (feature included).

In order to check if my assumption about outliers coming mostly from Cadillac is valid, I have prepared the following plots:









We can see that most of the outliers are green dots (Cadillac), which supports my assumption.

3. Specification of the Model

x_1 : mileage

x_2	x_3	x_4	**Make**
1	0	0	Buick
0	1	0	Chevrolet
0	0	1	Pontiac
0	0	0	Cadillac

x_5	x_6	**Model**
1	0	Regular Class
0	1	Middle Class
0	0	Upper Class

x_7	x_8	**Trim**
1	0	Basic
0	1	Entry
0	0	High

x_9	x_{10}	x_{11}	x_{12}	**Type**
1	0	0	0	Convertible
0	1	0	0	Coupe
0	0	1	0	Hatchback
0	0	0	1	Wagon
0	0	0	0	Sedan

x_{13} : cylinder

x_{14} : liter

x_{15} : doors

x_{16} : cruise

x_{17} : sound

x_{18} : leather

I have decided to add interactions to the model: Make × Trim.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} +$$
$$\beta_{11} x_{i11} + \beta_{12} x_{i12} + \beta_{13} x_{i13} + \beta_{14} x_{i14} + \beta_{15} x_{i15} + \beta_{16} x_{i16} + \beta_{17} x_{i17} + \beta_{18} x_{i18} +$$
$$+\beta_{19} x_2 x_7 + \beta_{20} x_2 x_8 + \beta_{21} x_3 x_7 + \beta_{22} x_3 x_8 + \beta_{23} x_4 x_7 + \beta_{24} x_4 x_8 + \epsilon_i$$
$$i = 1,...,n$$

Model assumptions:

1. The mean of response is a linear function of the x_i
2. The errors are independent
3. The errors are normally distributed
4. The errors have equal variance and mean zero

4. Estimation of the Appropriate Model

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	21	62438359771	2973255227	564.77	<.0001
Error	608	3200850394	5264557		
Corrected Total	629	65639210165			

Root MSE	2294.46215	R-Square	0.9512
Dependent Mean	20569	Adj R-Sq	0.9496
Coeff Var	11.15470		

Test for significance of Regression:

$H_0 : \beta_1 = \ldots = \beta_{24} = 0$

$H_1 : \beta_j \neq 0$ for at least one j

$F_0 = 564.77$

p-value=0.001 < $\alpha = 0.05$

We reject the null hypothesis. We can conclude that there is a statistically significant linear association between the price of used cars and at least one of our predictor variables.

Note: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

Doors =	4 * Intercept - 2 * X9 - 2 * X10
X4_X7 =	X4 - X8 + X2_X8 + X3_X8
X4_X8 =	X8 - X2_X8 - X3_X8

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	B	30190	1354.66423	22.29	<.0001
Mileage	Mileage	1	-0.17856	0.01140	-15.66	<.0001
X2		1	-14264	1086.26723	-13.13	<.0001
X3		1	-18412	990.68261	-18.59	<.0001
X4		B	-15875	947.18639	-16.76	<.0001
X5		1	-2268.21260	825.54683	-2.75	0.0062
X6		1	-1337.69262	806.23140	-1.66	0.0976
X7		1	-2406.75697	608.91872	-3.95	<.0001
X8		B	-1938.71408	725.18027	-2.67	0.0077
X9		B	16697	747.01822	22.35	<.0001
X10		B	211.71144	280.35557	0.76	0.4504
X11		1	139.05793	387.94770	0.36	0.7201
X12		1	5611.83227	615.69502	9.11	<.0001
Cylinder	Cylinder	1	-1550.90703	352.48451	-4.40	<.0001
Liter	Liter	1	5636.53619	419.46531	13.44	<.0001
Doors	Doors	0	0	.	.	.
Cruise	Cruise	1	185.38782	260.02008	0.71	0.4761
Sound	Sound	1	120.77111	229.77453	0.53	0.5994
Leather	Leather	1	189.10696	247.88970	0.76	0.4458
X2_X7		1	545.93337	911.10193	0.60	0.5493
X2_X8		B	466.04232	1149.41048	0.41	0.6853
X3_X7		1	3125.39848	843.24590	3.71	0.0002
X3_X8		B	3059.62838	971.77190	3.15	0.0017
X4_X7		0	0	.	.	.
X4_X8		0	0	.	.	.

Table 4.1 Parameter Estimates

SAS gives parameter estimate 0 for Doors and two interactions X4_X7 and X4_X8 because of multicollinearity.

The prediction equation is:

$$\hat{y} = 30190 - 0.17856X_1 - 14264X_2 - 18412X_3 - 15875X_4 - 2268.21X_5 - 1337.69X_6 - 2406.76X_7 - 1938.71X_8 + 16697X_9 + 211.71X_{10}$$
$$+139.06X_{11} + 5611.83X_{12} - 155.90X_{13} + 5636.53X_{14} + 0 \times X_{15} + 185.39X_{16} + 120.77X_{17} + 189.11X_{18} + 545.93X_2X_7 + 466.04X_2X_8$$
$$+3125.39X_3X_7 + 3059.63X_3X_8 + 0 \times X_4X_7 + 0 \times X_4X_8$$

It appears from this analysis that:

- **Mileage** is a significant predictor of used cars price (x1)
- **Make** is a significant predictor of used cars price (x2, x3, x4)
- **Model** is a significant predictor of used cars price if it is **Regular Class** (x5)
- Model is not a significant predictor of used cars price if it is Middle Class (x6)
- **Trim** is a significant predictor of used cars price (x7, x8)
- **Type** is a significant predictor of used cars price if it is **Convertible** (x9) or **Wagon** (x12)
- Type is not a significant predictor of used cars price if it is Coupe (x10) or Hatchback (x11)
- **Cylinder** is a significant predictor of used cars price (x13)
- **Liter** is a significant predictor of used cars price (x14)

- Doors parameter has been set to 0 because the multicollinearity (x15)
- Cruise is not a significant predictor of used cars price (x16)
- Sound is not a significant predictor of used cars price (x17)
- Leather is not a significant predictor of used cars price (x18)
- X2 and X7 interaction is not a significant predictor of used cars price
- X2 and X8 interaction is not a significant predictor of used cars price
- **X3 and X7** interaction in a significant predictor of used cars price
- **X3 and X8** interaction a significant predictor of used cars price
- X4 and X7 interaction is not a significant predictor of used cars price
- X4 and X8 is not a significant predictor of used cars price

Interpretation of magnitudes of the estimated regression coefficients:

- If the milage increases by one mile, the price decreases by 0.17856 (0.18 USD)
- If a car is made by Buick, the price decreases by 14264 USD in comparison to a car made by Cadillac
- If a car is made by Chevrolet, the price decreases by 18412 USD in comparison to a car made by Cadillac
- If a car is made by Pontiac, the price decreases by 15875 USD in comparison to a car made by Cadillac
- If a car's model is classified as Regular Class, the price decreases by 2268.21 USD in comparison to a car classifies as Upper Class
- If a car's model is classified as Middle Class , the price decreases by 1337.21 USD in comparison to a car classifies as Upper Class
- If a car's trim is classified as Basic, the price decreases by 2406.75 USD in comparison to a car classified as High trim
- If a car's trim is classified as Entry, the price decreases by 1938.71 USD in comparison to a car classified as High trim
- If a car's type is classified as Convertible, the price increases by 16697 USD in comparison to a car classified as Sedan
- If a car's type is classified as Coupe, the price increases by 211.71 USD in comparison to a car classified as Sedan
- If a car's type is classified as Hatchback, the price increases by 139.07 USD in comparison to a car classified as Sedan
- If a car's type is classified as Wagon, the price increases by 5636.53 USD in comparison to a car classified as Sedan
- If the number of cylinders in the engine increases by one, the price decreases by 1550.90 USD
- If the engine size increases by one liter, the price increases by 5636.54 USD
- If a car has a cruise control, the price increases by 185.38 USD (in comparison to a car without it)

- If a car has upgrader speakers, the price increases by 120.77 USD (in comparison to a car without upgraded speakers)
- If a car has leather seats, the price increases by 189.10 USD
- If a car is made by Buick and its trim is classified as Basic, the price increases by 535.93 USD
- If a car is made by Buick and its trim is classified as Entry, the price increases by 466.04 USD
- If a car is made by Chevrolet and its trim is classified as Basic, the price increases by 3125.39 USD
- If a car is bade by Chevrolet and its trim is classified as Entry, the price increases by 3059.62 USD

This preliminary model explains 95% (R-Square=0.95) of variations in the price of used cars. However, further tests of model adequacy are required. Moreover, from the investigation stage, I can presume that some of the estimated coefficients can be inflated by the existence of correlation among predictor variables (e.g. relationship between Cylinder and Liter).

At this stage of the analysis, I have decided to remove the interactions X4_X7 and X4_X8, and Doors variable based on the SAS output.

5. Assessment of the Chosen Prediction Equation

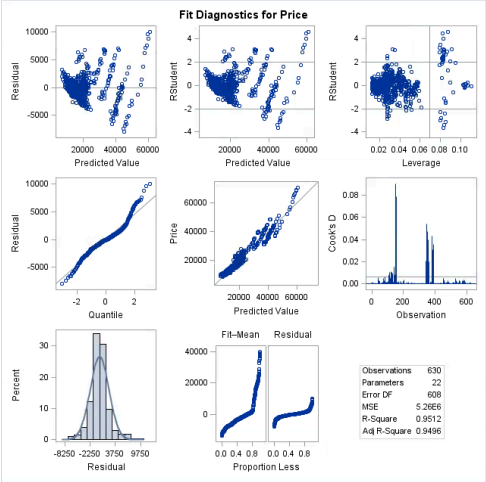


Figure 5.1 Fit diagnostics for Price

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	21	3.849832E16	1.833253E15	29.12	<.0001
Error	608	3.828301E16	6.296548E13		
Corrected Total	629	7.678133E16			

Table 5.1 B-P Test

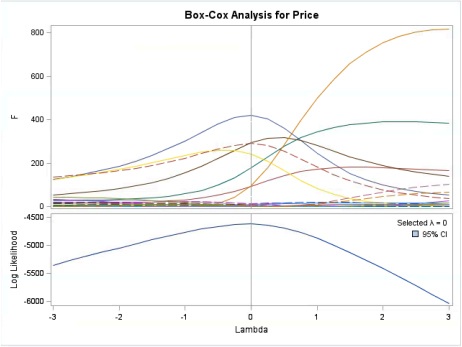
The Breusch-Pagan test for constancy of error variance:

$H_0 : \gamma_1 = \ldots = \gamma_{24} = 0$ (constant variance)

$H_1 : \gamma_1 \neq \ldots \neq \gamma_{24} \neq 0$ for at least one j (non constant variance)

P-value=0.001 < $\alpha = 0.05$

We reject the null hypothesis. The variance of the residuals is not constant, which can be also proved by Figure 5.1.



Since the variance of the residuals is not constant, we should consider the transformation of Y values. The Box-Cox analysis suggests the log(Y) transformation.

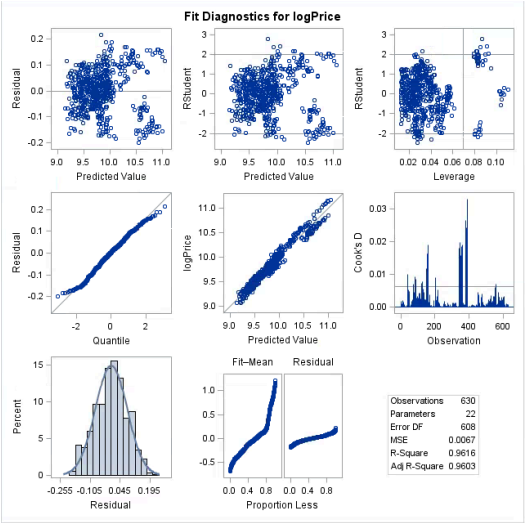


Figure 5.2 Fit Diagnostics for logPrice

Figure 5.2 presents the fit diagnostics after the log(Y) transformation. We can see that variance of the residuals has been improved. Normality has been also improved.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.991926	Pr < W	0.0017
Kolmogorov-Smirnov	D	0.030563	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.113062	Pr > W-Sq	0.0792
Anderson-Darling	A-Sq	0.93773	Pr > A-Sq	0.0189

In the case of large sample, we use the **Kolmogorov-Smirnov test** for normality.

H_0 : the errors follow a normal distribution

H_1 : the errors do not follow a normal distribution

P-value=0.15 > $\alpha = 0.05$. We fail to reject the null hypothesis. The errors follow a normal distribution.

Diagnostics for Leverage and Influence

Measure	Criteria	Observations
Outlier	|R|>3	none
h_{ii}	$\frac{2p}{n} = 0.079$	#11-20, 151-160, 341-360, 381-390
D_i	$\frac{4}{n} = 0.0063$	#41, 75, 81, 82, 85, 88, 90, 121, 125, 128, 151-160, 201, 342, 344-360, 381-390, 560
|DFFITS|	$2\sqrt{\frac{p}{n}} = 0.398$	#41, 42, 75, 81, 82, 83, 85, 88, 90, 121, 122, 125, 128, 151-159, 341-360, 381-390, 555
|DFBETAs|	$\frac{2}{\sqrt{n}} = 0.07968$	Intercept: # 81-86, 88, 90, 106, 107, 108, 110, 112, 121-139, 148, 150, 481, 547, 552, 553, 554, 555, 559, 560 x1: # 42, 42, 80, 81, 82, 90, 120, 121, 122, 130, 131, 140, 141, 150-153, 158, 159, 160, 201, 203, 219, 221, 280, 320, 341, 342, 349-352, 360, 361, 380. 382, 388, 390, 451, 452, 454, 531, 571 x2: # 81-90, 96, 97, 106, 108, 110, 151-160, 341-360, 381-390 x3: # 81-91, 95, 96, 97, 100, 128, 215, 341-360, 381-390 x4: # 81-90, 131-139, 151-160, 341-360, 381-390, 454, 527, 547, 572, 574 x5: # 81-91, 151-160, 341-360, 381-390 x6: # 81, 82, 89, 151-160, 341-360, 381-390 x7: #81-91, 101-131,153-156, 348, 352, 388 x8: # 81-90, 95, 101-131, 144-148 x9: # 101-131, 151-160, 341-360, 381-390
Covratio	$1 \pm \frac{3p}{n}$	>1.119: # 11-20 <0.881: # none

Observations 151-160, 341-360, and 381-390 are most influential.

Observations 151-160: Cadillac, the most expensive trim

Observations 341-360: Chevrolet, the most expensive trim

Observations 381-390: Chevrolet, price range 15,000 - 27,000 USD

Multicollinearity Diagnostics

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	9.92105	0.04823	205.70	<.0001	0
Mileage	Mileage	1	-0.00000833	4.059556E-7	-20.52	<.0001	1.02533
X2		1	-0.37437	0.03867	-9.68	<.0001	15.65394
X3		1	-0.47046	0.03527	-13.34	<.0001	29.35475
X4		1	-0.57822	0.03372	-17.15	<.0001	19.47606
X5		1	-0.09416	0.02939	-3.20	0.0014	19.97319
X6		1	-0.00420	0.02870	-0.15	0.8836	15.32582
X7		1	-0.05859	0.02168	-2.70	0.0071	10.62038
X8		1	-0.02189	0.02582	-0.85	0.3968	12.84324
X9		1	0.26096	0.02660	9.81	<.0001	2.05267
X10		1	-0.01660	0.00998	-1.66	0.0969	1.45033
X11		1	-0.03633	0.01381	-2.63	0.0087	1.55192
X12		1	0.33983	0.02192	15.50	<.0001	2.05732
Cylinder	Cylinder	1	-0.04494	0.01255	-3.58	0.0004	28.62010
Liter	Liter	1	0.25429	0.01493	17.03	<.0001	27.43370
Cruise	Cruise	1	0.01370	0.00926	1.48	0.1395	1.57014
Sound	Sound	1	0.02030	0.00818	2.48	0.0133	1.27636
Leather	Leather	1	0.03114	0.00883	3.53	0.0005	1.39326
X2_X7		1	-0.08016	0.03244	-2.47	0.0137	7.25821
X2_X8		1	-0.05441	0.04092	-1.33	0.1842	2.46967
X3_X7		1	-0.08568	0.03002	-2.85	0.0045	17.92307
X3_X8		1	-0.06291	0.03460	-1.82	0.0695	16.28630

Table 5.2 Variance Inflation Factor

at VIF=29.35 $R_j^2 = 1 - \frac{1}{29.35} = 0.966$

at VIF=28.60 $R_j^2 = 1 - \frac{1}{28.60} = 0.965$

at VIF=27.43 $R_j^2 = 1 - \frac{1}{27.43} = 0.963$

Coefficients for X_3, Cylinder, Liter, are highly inflated.

Mileage	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	Cylinder	Liter	Cruise	Sound	Leather	X2_X7	X2_X8	X3_X7	X3_X8
0.00133	0.00007425	0.00010337	0.00008808	0.00013486	0.00014458	0.00023454	0.00019311	0.00013599	0.00149	0.00073606	0.00016211	0.00001985	0.00003707	0.00132	0.00183	0.00150	0.00011678	0.00007763	0.00013892	0.00011816
0.00007458	0.00350	0.00078882	0.00073298	0.00046078	0.00228	0.00022690	0.00062472	0.00007265	0.00110	0.02160	0.00092619	0.00000582	0.00001867	0.00200	0.00028638	0.00045288	0.00862	0.00091086	0.00057986	0.00216
0.00001875	0.00019843	0.00012582	0.00030544	0.00018040	0.00121	0.00278	0.00795	0.00001950	0.00135	0.01076	0.00247	7.811922E-7	0.00000264	0.00006184	0.00000678	0.00015608	0.00005812	0.00490	0.00467	0.00500
0.00002911	0.00732	0.00015900	0.00709	0.00016806	0.00097178	0.00001192	0.00010019	0.00006123	0.00828	0.00948	0.03534	9.597668E-7	0.00000213	0.00001232	0.00020930	0.00013901	0.01177	0.01371	0.00020657	0.00023193
0.00012163	0.00127	0.00010328	0.00089573	0.00184	0.00219	0.00019602	0.00002708	0.03100	0.04894	0.01570	0.14223	0.00000439	0.00003504	0.00201	0.00002576	0.00030713	0.00003198	0.03125	0.00012571	4.907499E-7
0.00022040	0.00004088	0.00005540	0.00008772	0.00000181	0.00282	0.00058045	0.00014667	0.24690	0.03517	0.00259	0.00265	0.00002654	0.00006608	0.00138	0.00011434	0.00191	0.00470	0.02723	0.00050529	0.00068326
0.00000348	0.00000988	0.00000147	0.00002309	0.00001154	1.38755E-8	0.00012649	0.00082109	0.02454	0.05964	0.07035	0.01475	0.00000157	5.554489E-7	0.00032236	0.00003004	0.00017470	0.01150	0.19601	0.00064584	0.00093471
0.00039671	0.00180	0.00021547	0.00215	0.00032067	0.00164	0.00048588	0.00006776	0.02590	0.19713	0.17762	0.12010	0.00002323	0.00003875	0.00164	0.00154	0.00037008	0.00807	0.00911	0.00114	0.00692
0.00149	0.00016914	0.00000690	0.00044678	0.00064127	0.00415	0.00036085	0.00137	0.08814	0.26006	0.42929	0.02315	0.00003237	0.00005154	0.00739	0.00469	0.01053	0.00233	0.04606	0.00069154	0.00112
0.00003844	0.00040580	0.00481	0.00315	0.00119	0.01830	0.00010156	0.00822	0.08560	0.15815	0.08932	0.00003190	0.00016061	0.00038306	0.01657	0.00419	0.07014	0.00690	0.00262	0.01374	0.00007190
0.00986	0.00210	0.00221	0.01527	0.00387	0.00354	0.00025411	0.00045948	0.01224	0.02142	0.00099141	0.11769	0.00012080	0.00033107	0.06782	0.19448	0.04754	0.00485	0.00764	0.00087551	0.00560
0.00589	0.03538	2.401318E-7	0.00681	0.01404	0.00366	0.02333	0.00419	0.00363	0.06046	0.00569	0.02442	0.00001118	0.00001976	0.01509	0.15975	0.02365	0.01455	0.14569	0.00223	0.01351
0.00541	0.00264	0.00012788	0.00010850	0.00036836	0.00014833	0.00003197	0.01383	0.00323	0.00143	0.03079	0.00001965	0.00000497	0.00000907	0.15446	0.37877	0.34006	0.04660	0.01479	0.00358	0.00058861
0.59040	0.00361	0.00032762	0.00004920	6.256176E-9	0.00144	0.00097175	0.02436	0.00061678	0.00074574	0.00557	0.05483	0.00000488	0.00002763	0.20569	0.00007987	0.03164	0.03243	0.00007783	0.00889	0.00687
0.13138	0.00077583	0.00027539	0.02599	0.01374	0.00573	0.06246	0.01752	0.03454	0.00212	0.03304	0.33381	0.00019520	0.00039113	0.01203	0.18190	0.07068	0.00045577	0.02478	0.01897	0.04792
0.19607	0.08271	0.00138	0.00620	0.00123	0.04354	0.00440	0.01387	0.01238	0.03293	0.00364	0.07747	0.00080470	0.00326	0.27494	0.01457	0.01913	0.20610	0.02253	0.00204	0.00173
0.00005820	0.02435	0.00122	0.00177	0.00396	0.04524	0.00025846	0.05104	0.02971	0.02988	0.00000771	0.00182	0.00281	0.00670	0.20323	0.02552	0.32983	0.06426	0.00403	0.00118	0.02785
0.00001109	0.08582	0.09191	0.01311	0.01637	0.01049	0.18122	0.23153	0.21051	0.00000171	0.00118	0.00000556	0.00012777	0.00430	0.00018528	0.00632	0.01192	0.19222	0.17017	0.08111	0.02824
0.03974	0.11328	0.00327	0.35682	0.10990	0.22408	0.00617	0.03930	0.00479	0.05199	0.00100	0.00241	0.00129	0.02725	0.03054	0.01044	0.00281	0.01677	0.01155	0.25190	0.31048
0.00452	0.23915	0.41309	0.00343	0.09906	0.07893	0.35887	0.38073	0.00331	0.00051096	0.00418	0.00085976	0.01012	0.00086066	0.00042911	0.00000159	0.00145	0.23140	0.18636	0.46201	0.43937
0.01274	0.11354	0.09780	0.40520	0.69075	0.53039	0.26386	0.19371	0.18264	0.01000	0.00168	0.04330	0.00041858	0.06523	0.00226	0.00157	0.00702	0.05914	0.06968	0.07466	0.08275
0.00020499	0.28185	0.38202	0.15027	0.04177	0.01910	0.09307	0.00993	0.00002790	0.01719	0.08480	0.00154	0.98381	0.89098	0.00060894	0.01367	0.02859	0.07712	0.01083	0.07011	0.01786

Table 5.3 Eigensystem Analysis

Correlation of Estimates																					
Variable	Label	Intercept	Mileage	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	Cylinder	Liter	Cruise	Sound	Leather	X2_X7	X2_X8	X3_X7	X3_X8
Intercept	Intercept	1.0000	-0.1428	-0.3516	-0.3864	-0.1349	-0.1753	-0.1418	-0.5135	-0.2832	-0.1999	-0.0972	0.2063	-0.0785	-0.7637	0.5732	-0.1119	-0.1133	-0.2173	0.3269	0.2046	0.2663	0.1554
Mileage	Mileage	-0.1428	1.0000	0.0018	-0.0029	0.0236	-0.0207	-0.0709	0.0390	0.0454	0.0164	0.0238	0.0085	-0.0616	-0.0188	0.0178	-0.0278	0.0103	-0.0633	0.0157	-0.0505	-0.0137	-0.0082
X2		-0.3516	0.0018	1.0000	0.6885	0.6091	-0.6679	-0.5668	0.3288	0.2103	-0.0588	-0.0585	-0.1481	-0.0478	0.4682	-0.5339	-0.0227	0.0858	0.1433	-0.5797	-0.3651	-0.1230	-0.0459
X3		-0.3864	-0.0029	0.6885	1.0000	0.5419	-0.5657	-0.4658	0.2812	0.1581	-0.2889	-0.0129	-0.1714	-0.0559	0.5552	-0.6161	-0.0071	0.0851	0.1411	-0.2809	-0.1242	-0.5610	-0.4223
X4		-0.1349	0.0236	0.6091	0.5419	1.0000	-0.7684	-0.8222	-0.2365	-0.3840	-0.2967	-0.1273	-0.1565	-0.3331	0.3677	-0.4281	0.0938	0.0696	0.0340	0.1727	0.2227	0.2788	0.4203
X5		-0.1753	-0.0207	-0.6679	-0.5657	-0.7684	1.0000	0.8767	0.1609	0.1405	0.3826	0.1524	0.0564	0.0831	-0.1782	0.3432	0.0155	-0.1477	-0.0031	0.0102	-0.0669	-0.1750	-0.1897
X6		-0.1418	-0.0709	-0.5668	-0.4658	-0.8222	0.8767	1.0000	0.1399	0.1180	0.3379	0.0499	0.0691	0.2659	-0.1134	0.2278	-0.0845	-0.0697	0.1371	-0.1558	-0.0576	-0.2128	-0.2469
X7		-0.5135	0.0390	0.3288	0.2812	-0.2365	0.1609	0.1399	1.0000	0.7966	0.4212	0.0816	-0.0749	-0.0195	0.2302	-0.1808	-0.0051	-0.0949	0.0191	-0.6485	-0.5039	-0.5903	-0.4877
X8		-0.2832	0.0454	0.2103	0.1581	-0.3840	0.1405	0.1180	0.7966	1.0000	0.3513	0.1082	-0.0018	0.1810	0.0358	-0.0192	-0.0250	-0.0660	-0.0270	-0.5114	-0.6292	-0.4788	-0.6654
X9		-0.1999	0.0164	-0.0588	-0.2889	-0.2967	0.3826	0.3379	0.4212	0.3513	1.0000	0.1035	-0.0013	-0.0275	-0.0140	0.0389	0.0310	-0.1797	-0.0252	-0.2340	-0.2103	0.0045	0.0004
X10		-0.0972	0.0238	-0.0585	-0.0129	-0.1273	0.1524	0.0499	0.0816	0.1082	0.1035	1.0000	0.1787	-0.0094	0.1416	-0.1325	-0.0280	-0.0778	-0.1111	0.0199	-0.0713	-0.2005	-0.2092
X11		0.2063	0.0085	-0.1481	-0.1714	-0.1565	0.0564	0.0691	-0.0749	-0.0018	-0.0013	0.1787	1.0000	0.1104	-0.2974	0.2794	0.1235	0.0470	-0.0758	0.0432	0.0067	0.0055	-0.1292
X12		-0.0785	-0.0616	-0.0478	-0.0559	-0.3331	0.0831	0.2659	-0.0195	0.1810	-0.0275	-0.0094	0.1104	1.0000	-0.0411	0.0931	0.0311	0.1734	0.1493	-0.1192	-0.1167	0.0116	-0.1806
Cylinder	Cylinder	-0.7637	-0.0188	0.4682	0.5552	0.3677	-0.1782	-0.1134	0.2302	0.0358	-0.0140	0.1416	-0.2974	-0.0411	1.0000	-0.9427	-0.0149	0.1118	0.1345	-0.2230	-0.0604	-0.2095	-0.0867
Liter	Liter	0.5732	0.0178	-0.5339	-0.6161	-0.4281	0.3432	0.2278	-0.1808	-0.0192	0.0389	-0.1325	0.2794	0.0931	-0.9427	1.0000	-0.0420	-0.0993	-0.1608	0.2284	0.0507	0.2259	0.1069
Cruise	Cruise	-0.1119	-0.0278	-0.0227	-0.0071	0.0938	0.0155	-0.0845	-0.0051	-0.0250	0.0310	-0.0280	0.1235	0.0311	-0.0149	-0.0420	1.0000	0.0005	0.0528	0.0745	0.0159	0.0777	0.0988
Sound	Sound	-0.1133	0.0103	0.0858	0.0851	0.0696	-0.1477	-0.0697	-0.0949	-0.0660	-0.1797	-0.0778	0.0470	0.1734	0.1118	-0.0993	0.0005	1.0000	-0.1424	-0.0496	-0.0252	0.0134	-0.0404
Leather	Leather	-0.2173	-0.0633	0.1433	0.1411	0.0340	-0.0031	0.1371	0.0191	-0.0270	-0.0252	-0.1111	-0.0758	0.1493	0.1345	-0.1608	0.0528	-0.1424	1.0000	-0.0866	0.0425	-0.0695	-0.0454
X2_X7		0.3269	0.0157	-0.5797	-0.2809	0.1727	0.0102	-0.1558	-0.6485	-0.5114	-0.2340	0.0199	0.0432	-0.1192	-0.2230	0.2284	0.0745	-0.0496	-0.0866	1.0000	0.5831	0.4201	0.3662
X2_X8		0.2046	-0.0505	-0.3651	-0.1242	0.2227	-0.0669	-0.0576	-0.5039	-0.6292	-0.2103	-0.0713	0.0067	-0.1167	-0.0604	0.0507	0.0159	-0.0252	0.0425	0.5831	1.0000	0.3083	0.4253
X3_X7		0.2663	-0.0137	-0.1230	-0.5610	0.2788	-0.1750	-0.2128	-0.5903	-0.4788	0.0045	-0.2005	0.0055	0.0116	-0.2095	0.2259	0.0777	0.0134	-0.0695	0.4201	0.3083	1.0000	0.8392
X3_X8		0.1554	-0.0082	-0.0459	-0.4223	0.4203	-0.1897	-0.2469	-0.4877	-0.6654	0.0004	-0.2092	-0.1292	-0.1806	-0.0867	0.1069	0.0988	-0.0404	-0.0454	0.3662	0.4253	0.8392	1.0000

Table 5.4 Correlation Matrix

The Eigensystem Analysis also indicates multicollinearity for X_3, Cylinder, Liter (Table 5.3).

The correlation matrix (Table 5.4) indicates high correlation for:

- x6 and x5 (0.8767)
- x4 and x5 (-0.7684)
- x4 and x6 (-0.822)
- x8 and x7 (0.7966)
- x3_x8 and x3_x7 (0.8392)
- Cylinder and Liter (-0.9427)

Based on the multicollinearity diagnostics, I have decided to remove Cylinder, X5, and the interaction X3_X7 from the model. I have decided to keep X3 in the model, because the removal of this variable would affect the normality (Figure 5.3).

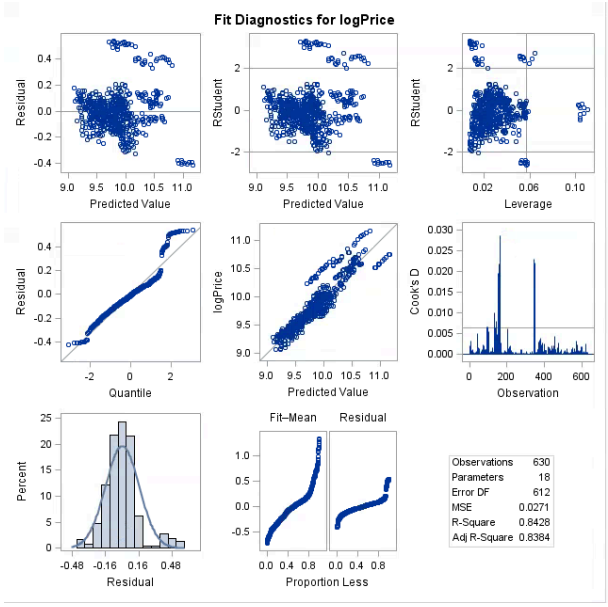


Figure 5.3 Fit Diagnostics after removing X3, Cylinder, X5, and X3_X7

After the changes, the model includes the following variables:

```
X1 X2 X3 X4  X6 X7 X8 X9 X10 X11 X12 X14 X16 X17 X18 X2_X7 X2_X8
```

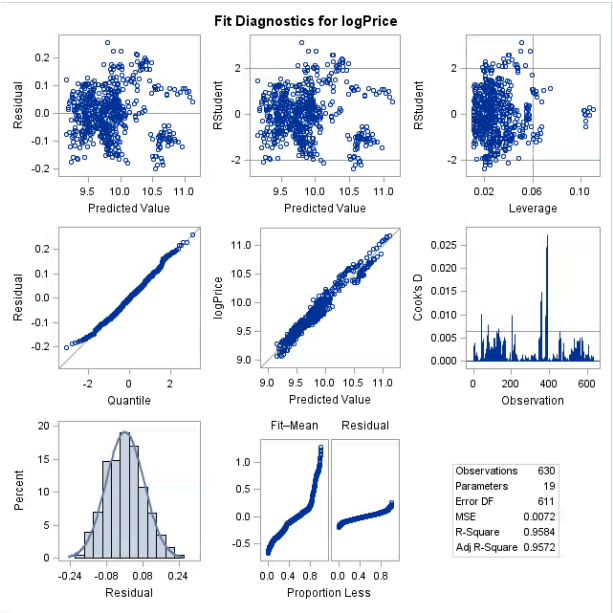


Figure 5.3 Fit Diagnostics for a new model

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.99363	Pr < W	0.0092
Kolmogorov-Smirnov	D	0.030661	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.085929	Pr > W-Sq	0.1794
Anderson-Darling	A-Sq	0.709286	Pr > A-Sq	0.0673

Table 5.5 Test for Normality

In the case of large sample, we use the **Kolmogorov-Smirnov test** for normality.

H_0 : the errors follow a normal distribution

H_1 : the errors do not follow a normal distribution

P-value=0.15 > $\alpha = 0.05$. We fail to reject the null hypothesis. The errors follow a normal distribution.

Based on Figure 5.3 and Table 5.5, I conclude that the new model meets assumptions about constant variance and normality.

Th new prediction equation is:

$$log\text{Prise} = 9.739 - 0.00000844X_1 - 0.42815X_2 - 0.56002X_3 - 0.59462X_4 + 0.07485X_6 - 0.07611X - 7 - 0.05896X_8 + 0.31078X_9$$

$$-0.01094X_{10} - 0.05464X_{11} + 0.34530X_{12} + 0.21913X_{14} + 0.01718X_{16} + 0.01978X_{17} + 0.03454X_{18} - 0.05167X_2X_7 - 0.02012X_2X_8 + 0.02946X_3X_8$$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	101.23275	5.62404	782.79	<.0001
Error	611	4.38981	0.00718		
Corrected Total	629	105.62256			

Root MSE	0.08476	R-Square	0.9584
Dependent Mean	9.83819	Adj R-Sq	0.9572
Coeff Var	0.86156		

Table 5.6 Analysis of Variance of the new model

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	9.73852	0.02809	346.67	<.0001
Mileage	Mileage	1	-0.00000844	4.208973E-7	-20.05	<.0001
X2		1	-0.42815	0.02545	-16.82	<.0001
X3		1	-0.56002	0.01358	-41.24	<.0001
X4		1	-0.59462	0.01948	-30.53	<.0001
X6		1	0.07485	0.01419	5.28	<.0001
X7		1	-0.07611	0.01789	-4.25	<.0001
X8		1	-0.05896	0.02342	-2.52	0.0121
X9		1	0.31078	0.02533	12.27	<.0001
X10		1	-0.01094	0.00998	-1.10	0.2735
X11		1	-0.05464	0.01366	-4.00	<.0001
X12		1	0.34530	0.02265	15.24	<.0001
Liter	Liter	1	0.21913	0.00424	51.74	<.0001
Cruise	Cruise	1	0.01718	0.00957	1.79	0.0732
Sound	Sound	1	0.01978	0.00836	2.37	0.0183
Leather	Leather	1	0.03454	0.00907	3.81	0.0002
X2_X7		1	-0.05167	0.03013	-1.71	0.0869
X2_X8		1	-0.02012	0.04039	-0.50	0.6185
X3_X8		1	0.02946	0.01923	1.53	0.1259

Table 5.7 Parameter Estimates for the new model

6. Selection of Variable Subset

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	101.20155	6.74677	937.01	<.0001
Error	614	4.42101	0.00720		
Corrected Total	629	105.62256			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	9.75068	0.02696	941.76314	130794	<.0001
Mileage	-0.00000846	4.205103E-7	2.91301	404.57	<.0001
X2	-0.44545	0.02049	3.40236	472.53	<.0001
X3	-0.56064	0.01306	13.27490	1843.65	<.0001
X4	-0.60569	0.01740	8.72742	1212.08	<.0001
X6	0.07603	0.01364	0.22376	31.08	<.0001
X7	-0.08572	0.01595	0.20788	28.87	<.0001
X8	-0.05030	0.01692	0.06366	8.84	0.0031
X9	0.30484	0.02447	1.11778	155.24	<.0001
X11	-0.04643	0.01301	0.09170	12.74	0.0004
X12	0.35602	0.02124	2.02228	280.86	<.0001
Liter	0.21831	0.00422	19.30980	2681.79	<.0001
Cruise	0.01583	0.00956	0.01975	2.74	0.0982
Sound	0.02077	0.00828	0.04535	6.30	0.0123
Leather	0.03390	0.00895	0.10321	14.33	0.0002
X2_X7	-0.03355	0.02496	0.01301	1.81	0.1793

Table 6.1 Forward selection

$\alpha = 0.25$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	101.18854	7.22775	1002.49	<.0001
Error	615	4.43402	0.00721		
Corrected Total	629	105.62256			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	9.75575	0.02671	961.60325	133375	<.0001
Mileage	-0.00000843	4.200735E-7	2.90022	402.26	<.0001
X2	-0.46328	0.01563	6.33656	878.88	<.0001
X3	-0.55821	0.01294	13.41680	1860.91	<.0001
X4	-0.59858	0.01659	9.39076	1302.50	<.0001
X6	0.06838	0.01240	0.21911	30.39	<.0001
X7	-0.09430	0.01463	0.29946	41.54	<.0001
X8	-0.05587	0.01641	0.08353	11.59	0.0007
X9	0.29721	0.02381	1.12297	155.76	<.0001
X11	-0.04770	0.01298	0.09731	13.50	0.0003
X12	0.35170	0.02101	2.01974	280.14	<.0001
Liter	0.21850	0.00422	19.36646	2686.13	<.0001
Cruise	0.01676	0.00954	0.02225	3.09	0.0795
Sound	0.02046	0.00828	0.04403	6.11	0.0137
Leather	0.03291	0.00893	0.09795	13.59	0.0002

Table 6.2 Backward Elimination

$\alpha = 0.10$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	101.18854	7.22775	1002.49	<.0001
Error	615	4.43402	0.00721		
Corrected Total	629	105.62256			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	9.75575	0.02671	961.60325	133375	<.0001
Mileage	-0.00000843	4.200735E-7	2.90022	402.26	<.0001
X2	-0.46328	0.01563	6.33656	878.88	<.0001
X3	-0.55821	0.01294	13.41680	1860.91	<.0001
X4	-0.59858	0.01659	9.39076	1302.50	<.0001
X6	0.06838	0.01240	0.21911	30.39	<.0001
X7	-0.09430	0.01463	0.29946	41.54	<.0001
X8	-0.05587	0.01641	0.08353	11.59	0.0007
X9	0.29721	0.02381	1.12297	155.76	<.0001
X11	-0.04770	0.01298	0.09731	13.50	0.0003
X12	0.35170	0.02101	2.01974	280.14	<.0001
Liter	0.21850	0.00422	19.36646	2686.13	<.0001
Cruise	0.01676	0.00954	0.02225	3.09	0.0795
Sound	0.02046	0.00828	0.04403	6.11	0.0137
Leather	0.03291	0.00893	0.09795	13.59	0.0002

Table 6.3 Stepwise Regression

$\alpha = 0.15$

Number in Model	C(p)	R-Square	Variables in Model
14	15.7960	0.9580	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Cruise Sound Leather
15	15.9889	0.9581	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Cruise Sound Leather X2_X7
14	16.7316	0.9580	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Sound Leather X2_X7
13	16.8861	0.9578	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Sound Leather
16	16.9218	0.9582	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Cruise Sound Leather X2_X7 X2_X8
16	17.1105	0.9582	Mileage X2 X3 X4 X6 X7 X8 X9 X10 X11 X12 Liter Cruise Sound Leather X2_X7
15	17.2600	0.9581	Mileage X2 X3 X4 X6 X7 X8 X9 X10 X11 X12 Liter Cruise Sound Leather
15	17.7669	0.9580	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Sound Leather X2_X7 X2_X8
15	17.7799	0.9580	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Cruise Sound Leather X2_X8
15	17.8880	0.9580	Mileage X2 X3 X4 X6 X7 X8 X9 X10 X11 X12 Liter Sound Leather X2_X7
17	18.0000	0.9583	Mileage X2 X3 X4 X6 X7 X8 X9 X10 X11 X12 Liter Cruise Sound Leather X2_X7 X2_X8
14	18.4048	0.9578	Mileage X2 X3 X4 X6 X7 X8 X9 X10 X11 X12 Liter Sound Leather
16	18.8836	0.9581	Mileage X2 X3 X4 X6 X7 X8 X9 X10 X11 X12 Liter Sound Leather X2_X7 X2_X8
14	18.8859	0.9578	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Sound Leather X2_X8
15	19.1044	0.9579	Mileage X2 X3 X4 X6 X7 X9 X11 X12 Liter Cruise Sound Leather X2_X7 X2_X8
16	19.2533	0.9581	Mileage X2 X3 X4 X6 X7 X8 X9 X10 X11 X12 Liter Cruise Sound Leather X2_X8
13	19.9109	0.9576	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Cruise Leather
16	20.0131	0.9580	Mileage X2 X3 X4 X6 X7 X9 X10 X11 X12 Liter Cruise Sound Leather X2_X7 X2_X8
14	20.2874	0.9577	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Cruise Leather X2_X7
14	20.3640	0.9577	Mileage X2 X3 X4 X6 X7 X9 X11 X12 Liter Sound Leather X2_X7 X2_X8

Table 6.4 C(p) selection method

Number in Model	Adjusted R-Square	R-Square	Variables in Model
16	0.9571	0.9582	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Cruise Sound Leather X2_X7 X2_X8
15	0.9571	0.9581	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Cruise Sound Leather X2_X7
17	0.9571	0.9583	Mileage X2 X3 X4 X6 X7 X8 X9 X10 X11 X12 Liter Cruise Sound Leather X2_X7 X2_X8
16	0.9571	0.9582	Mileage X2 X3 X4 X6 X7 X8 X9 X10 X11 X12 Liter Cruise Sound Leather X2_X7
14	0.9571	0.9580	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Cruise Sound Leather
15	0.9570	0.9581	Mileage X2 X3 X4 X6 X7 X8 X9 X10 X11 X12 Liter Cruise Sound Leather
14	0.9570	0.9580	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Sound Leather X2_X7
15	0.9570	0.9580	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Sound Leather X2_X7 X2_X8
15	0.9570	0.9580	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Cruise Sound Leather X2_X8
16	0.9570	0.9581	Mileage X2 X3 X4 X6 X7 X8 X9 X10 X11 X12 Liter Sound Leather X2_X7 X2_X8
15	0.9570	0.9580	Mileage X2 X3 X4 X6 X7 X8 X9 X10 X11 X12 Liter Sound Leather X2_X7
16	0.9570	0.9581	Mileage X2 X3 X4 X6 X7 X8 X9 X10 X11 X12 Liter Cruise Sound Leather X2_X8
13	0.9569	0.9578	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Sound Leather
16	0.9569	0.9580	Mileage X2 X3 X4 X6 X7 X9 X10 X11 X12 Liter Cruise Sound Leather X2_X7 X2_X8
15	0.9569	0.9579	Mileage X2 X3 X4 X6 X7 X9 X11 X12 Liter Cruise Sound Leather X2_X7 X2_X8
14	0.9569	0.9578	Mileage X2 X3 X4 X6 X7 X8 X9 X10 X11 X12 Liter Sound Leather
14	0.9568	0.9578	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Sound Leather X2_X8
15	0.9568	0.9578	Mileage X2 X3 X4 X6 X7 X8 X9 X10 X11 X12 Liter Sound Leather X2_X8
15	0.9568	0.9578	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Cruise Leather X2_X7 X2_X8
14	0.9568	0.9577	Mileage X2 X3 X4 X6 X7 X8 X9 X11 X12 Liter Cruise Leather X2_X7

Table 6.5 R-square

Based on the variable selection process presented in Tables 6.1, 6.2, 6.3, 6.4, and 6.5, I have decided to eliminate X10 and interactions from the model: X2_X7, X2_X8, X3_X8. The variable selection process yields the following model:

$$log(\text{Price})_i = \beta_0 + \beta_1 \text{Mileage}_i + \beta_2 \text{Buick}_i + \beta_3 \text{Chevrolet}_i + \beta_4 \text{Pontiac}_i + \beta_5 \text{MiddleClassModel}_i + \beta_6 \text{BasicTrim}_i + \beta_7 \text{EntryTrim}_i$$
$$+\beta_8 \text{Convertible}_i + \beta_9 \text{Hatchback}_i + \beta_{10} \text{Wagon}_i + \beta_{11} \text{Liter}_i + \beta_{12} \text{Cruise}_i + \beta_{13} \text{Sound}_i + \beta_{14} \text{Leather}_i + \epsilon_i$$

Root MSE	0.08491	R-Square	0.9580
Dependent Mean	9.83819	Adj R-Sq	0.9571
Coeff Var	0.86307		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	9.75575	0.02671	365.21	<.0001
Mileage	Mileage	1	-0.00000843	4.200735E-7	-20.06	<.0001
X2		1	-0.46328	0.01563	-29.65	<.0001
X3		1	-0.55821	0.01294	-43.14	<.0001
X4		1	-0.59858	0.01659	-36.09	<.0001
X6		1	0.06838	0.01240	5.51	<.0001
X7		1	-0.09430	0.01463	-6.44	<.0001
X8		1	-0.05587	0.01641	-3.40	0.0007
X9		1	0.29721	0.02381	12.48	<.0001
X11		1	-0.04770	0.01298	-3.67	0.0003
X12		1	0.35170	0.02101	16.74	<.0001
Liter	Liter	1	0.21850	0.00422	51.83	<.0001
Cruise	Cruise	1	0.01676	0.00954	1.76	0.0795
Sound	Sound	1	0.02046	0.00828	2.47	0.0137
Leather	Leather	1	0.03291	0.00893	3.69	0.0002

Table 6.6 Parameter Estimates after the variable selection process

Table 6.6 presents the parameter estimates after the variable selection process. I have decided to remove Cruise variable from the model as its p-value=0.0795 $> \alpha = 0.05$. The reduction yields the following model:

$$log(\text{Price})_i = \beta_0 + \beta_1 \text{Mileage}_i + \beta_2 \text{Buick}_i + \beta_3 \text{Chevrolet}_i + \beta_4 \text{Pontiac}_i + \beta_5 \text{MiddleClassModel}_i + \beta_6 \text{BasicTrim}_i + \beta_7 \text{EntryTrim}_i$$

$$+\beta_8 \text{Convertible}_i + \beta_9 \text{Hatchback}_i + \beta_{10} \text{Wagon}_i + \beta_{11} \text{Liter}_i + \beta_{12} \text{Sound}_i + \beta_{13} \text{Leather}_i + \epsilon_i$$

Validation of the Regression Model

In order to validate the regression model, I have split the data into estimation (70%) and prediction (30%) sets. Table 6.7 presents the parameter estimates for the estimation set.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	72.47719	5.57517	790.26	<.0001
Error	434	3.06180	0.00705		
Corrected Total	447	75.53899			

Root MSE	0.08399	R-Square	0.9595
Dependent Mean	9.83329	Adj R-Sq	0.9583
Coeff Var	0.85417		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	9.74297	0.03151	309.18	<.0001
Mileage	Mileage	1	-0.00000811	4.957853E-7	-16.36	<.0001
X2		1	-0.46049	0.01837	-25.07	<.0001
X3		1	-0.55392	0.01529	-36.23	<.0001
X4		1	-0.59293	0.01988	-29.83	<.0001
X6		1	0.07008	0.01435	4.89	<.0001
X7		1	-0.09736	0.01764	-5.52	<.0001
X8		1	-0.05482	0.01955	-2.80	0.0053
X9		1	0.31850	0.02740	11.62	<.0001
X11		1	-0.04666	0.01503	-3.10	0.0020
X12		1	0.35578	0.02491	14.28	<.0001
Liter	Liter	1	0.22186	0.00494	44.88	<.0001
Sound	Sound	1	0.02541	0.00971	2.62	0.0091
Leather	Leather	1	0.03054	0.01050	2.91	0.0038

Table 6.7 Parameter Estimates: Estimation Set

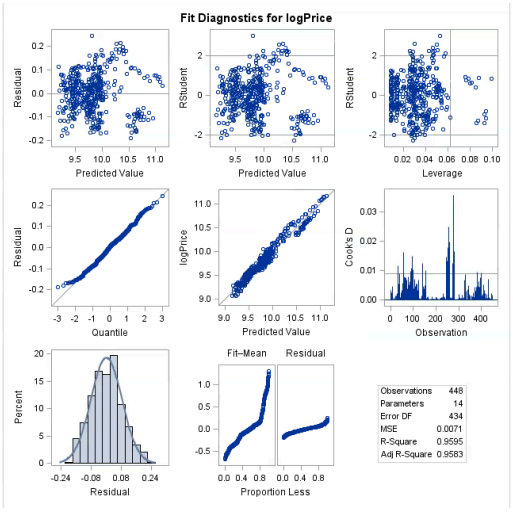


Figure 6.1 Fit Diagnostics

Test for significance of Regression:

$H_0 : \beta_1 = \ldots = \beta_{13} = 0$
$H_1 : \beta_j \neq 0$ for at least one j
$F_0 = 790.26$

p-value=0.001 < $\alpha = 0.05$ We reject the null hypothesis.

$R^2 = 0.9595$, which means that 95% of variation in the price of used cars is explain by the model.

Figure 6.1 shows that the model meets the assumptions of the constant variance and normality.

The prediction equation is:

$log(\text{Price}) = 9.74297 - 0.00000811\text{Mileage} - 0.46049\text{Buick} - 0.55392\text{Chevrolet} - 0.59293\text{Pontiac} + 0.07008\text{MiddleClassModel}$
$-0.09736\text{BasicTrim} - 0.05482\text{EntryTrim} + 0.31850\text{Convertible} - 0.04666\text{Hatchback} + 0.35578\text{Wagon} + 0.22186\text{Liter}$
$+0.02541\text{Sound} + 0.03054\text{Leather}$

After using this model to predict the observations in the prediction data set, I have obtained the following results:

logPrice	Predicted_val	Prediction_error
9.7593	9.8622	-0.10289
9.8921	9.8929	-0.00073
10.0150	10.0338	-0.01886
10.0768	10.0513	0.02550
10.0400	10.0206	0.01939
9.9652	9.9278	0.03736
10.0439	10.0343	0.00962
9.9516	9.9058	0.04581
9.9476	9.9410	0.00656
9.9378	10.1045	-0.16675
9.7842	9.9206	-0.13641
9.9968	9.9794	0.01739
9.9025	9.9264	-0.02387
9.8854	9.9002	-0.01477
10.0614	9.9754	0.08593
9.9850	9.9475	0.03747
10.1955	10.0259	0.16954
10.0583	9.9339	0.12432
10.0608	9.9229	0.13786
10.0502	9.8885	0.16169
9.9775	9.8531	0.12441
10.7525	10.8417	-0.08921
10.6949	10.7677	-0.07277
10.6895	10.7512	-0.06172
10.6988	10.7481	-0.04938
10.6776	10.7618	-0.08418
10.3556	10.2335	0.12210
10.3347	10.2016	0.13304
10.3350	10.1778	0.15717
10.4934	10.6159	-0.12249

Table 6.8 Example of Prediction Errors: 30 out of 182

We can see that the predicted values correspond closely to the observed values. The sum of squares of the prediction error is $\sum e_i^2 = 1.42611$, and the approximate R^2 for the prediction is:

$$R^2_{\text{pred}} = 1 - \frac{\sum e_i^2}{SS_T} = 1 - \frac{1.42611}{17689.00} = 0.99$$

We might expect this model to explain about 99% of the variability in new data.

7. Conclusion

The aim of this project was to develop the prediction equation for the Blue Book cost of a used car. The data set consists of 630 observations, one dependent variable and 11 predictor variables. The explanatory variables include both quantitative and qualitative regressors. Two of the categorical predictors (Trim and Model) include many categories (sometimes over 20). Therefore, in order to make the model more compact, I have decided to group these categories into three classes which have been introduced in Section 1 (Problem Description).

This report presents the set of activities allowing me to build a multivariate regression model, including:

- checking for the violations of model assumptions;
- data transformation;
- variable selection techniques;
- incorporation of categorical explanatory variables;
- application of *F*-tests
- validation of the final regression model.

Base on the results of the Breusch-Pagan test showing non constant variance, I have decided to make a log(Y) transformation. No other transformations were needed. Four of the qualitative variables have more than two levels, for that reason I have presented detailed tables in Section 3 (Specification of the Model) showing that a qualitative variable with a levels can be represented by a-1 indicator variables, each taking on the values 0 and 1.

After the implementation of variable selection techniques, some predictors have been removed. At the beginning of the analysis, I have included some interactions to the model. However, they happened to be either insignificant or highly correlated. As a result, the final model does not include any interactions.

After the entire process of model building, I was able to develop the prediction equation that explains about 99% of the variability in new data.

Since the dependent variable in the final model is log-transformed, the interpretation of effect of changes in X_i on Y is the following:

- each one unit increase in X_i results in $(e^{\hat{\beta}_i} - 1) \times 100$ percentage change in Y.

$$log(\text{Price}) = 9.74297 - 0.00000811\text{Mileage} - 0.46049\text{Buick} - 0.55392\text{Chevrolet} - 0.59293\text{Pontiac} + 0.07008\text{MiddleClassModel} - 0.09736\text{BasicTrim} - 0.05482\text{EntryTrim} + 0.31850\text{Convertible} - 0.04666\text{Hatchback} + 0.35578\text{Wagon} + 0.22186\text{Liter} + 0.02541\text{Sound} + 0.03054\text{Leather}$$

The interpretation of the prediction equation gives the following conclusions:

- an increase of one mileage results in $(e^{0.00000811} - 1) \times 100 = 0.000811$ percentage decrease in the price (an increase of 10,000 mileages results in 8.11% decrease in the price)
- if a car is made by Buick, the price is $(e^{0.46049} - 1) \times 100 = 58.48$ % lower than Cadillac
- if a car is made by Chevrolet, the price is $(e^{0.55392} - 1) \times 100 = 74$ % lower than Cadillac
- if a car is made by Pontiac, the price is $(e^{0.59293} - 1) \times 100 = 80.90$ % cheeper than Cadillac

- if a car is classified as middle class model, the price is $(e^{0.07008} - 1) \times 100 = 7.26\,\%$ higher than a car classified as upper class model
- if a car's trim is classified as "Basic", the price is $(e^{0.09736} - 1) \times 100 = 10.22\,\%$ lower than a car with a "High" trim (upgraded equipment and additional features)
- if a car's trim is classified as "Entry", the price is $(e^{0.05482} - 1) \times 100 = 5.63\,\%$ lower than a car with a "High" trim (upgraded equipment and additional features)
- for a convertible car, the price is $(e^{0.03185} - 1) \times 100 = 37.56\,\%$ higher than for a sedan
- for a hatchback, the price is $(e^{0.04666} - 1) \times 100 = 4.78\,\%$ lower than for a sedan
- for a wagon, the price is $(e^{0.35578} - 1) \times 100 = 42.73\,\%$ higher than for a sedan
- an increase of an engine size by one liter results in $(e^{0.22186} - 1) \times 100 = 24.83\,\%$ increase in the price
- the price of a car with upgrader speakers is higher by $(e^{0.02541} - 1) \times 100 = 2.57\,\%$
- the price of a car with leather seats is higher by $(e^{0.03054} - 1) \times 100 = 3.10\,\%$