Project Images and Notes

Sunday, June 10, 2018 1:44 PM

1985 Auto Imports Database Analyses

Dataset:

https://www.kaggle.com/ashishpal2702/1985-auto-imports-database-analyses-prediction

Objective:

We would like to examine the effect of car performance attributes on average loss payment per insured vehicle year.

Variable list sensitive to the objective:

Car attributes

1. make: [Manufacturer name eg : alfa-romero, audi, bmw, chevrolet, dodge, honda,isuzu etc.]
2. num-of-doors: [four, two].
3. body-style: [hardtop, wagon, sedan, hatchback, convertible]
4. engine-location: [front, rear]
5. wheel-base: [continuous from 86.6 120.9]
6. length: [continuous from 141.1 to 208.1]
7. width: [continuous from 60.3 to 72.3]
8. height: [continuous from 47.8 to 59.8]
9. curb-weight: [continuous from 1488 to 4066]

Performance attributes

1. fuel-type: [diesel, gas]
2. aspiration: [std, turbo]
3. drive-wheels: [4wd, fwd, rwd]
4. engine-type: [dohc, dohcv, l, ohc, ohcf, ohcv, rotor]
5. num-of-cylinders: [eight, five, four, six, three, twelve, two]
6. engine-size: [continuous from 61 to 326]
7. fuel-system: [1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi]
8. bore: [continuous from 2.54 to 3.94]
9. stroke: [continuous from 2.07 to 4.17]
10. compression-ratio: [continuous from 7 to 23]
11. horsepower: [continuous from 48 to 288]
12. peak-rpm: [continuous from 4150 to 6600]
13. city-mpg: [continuous from 13 to 49]
14. highway-mpg: [continuous from 16 to 54]

Insurance Risk attributes

1. normalized-losses: [average loss payment per insured vehicle year -> continuous from 65 to 256.]
2. symboling: [its assigned insurance risk rating -> [-3, -2, -1, 0, 1, 2, 3]]

Response Variable

1. Price

Data Sets

Original

41 data points were missing from Normalized Losses variable. We chose to remove this variable from the dataset in order to preserve as many records as possible. This variable is not reliable enough to keep in the analysis.

```
/*Import dataset with formatted columns*/
data auto;
    infile '/home/carollr0/DataSets/Automobile_data.csv' dlm=',' firstobs=2;
    input symboling normalizedlosses
        make $  fueltype $  aspiration $ numofdoors $
bodystyle $ drivewheels  $  enginelocation $ wheelbase length width height
        curbweight enginetype $ numofcylinders $ enginesize fuelsystem $ bore stroke
        compressionratio horsepower peakrpm citympg highwaympg price;
run;
```

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
5	aspiration	Char	5	$5.	$5.
7	body-style	Char	11	$11.	$11.
19	bore	Char	4	$4.	$4.
24	city-mpg	Num	8	BEST12.	BEST32.
21	compression-ratio	Num	8	BEST12.	BEST32.
14	curb-weight	Num	8	BEST12.	BEST32.
8	drive-wheels	Char	3	$3.	$3.
9	engine-location	Char	5	$5.	$5.
17	engine-size	Num	8	BEST12.	BEST32.
15	engine-type	Char	5	$5.	$5.
18	fuel-system	Char	4	$4.	$4.
4	fuel-type	Char	6	$6.	$6.
13	height	Num	8	BEST12.	BEST32.
25	highway-mpg	Num	8	BEST12.	BEST32.
22	horsepower	Char	3	$3.	$3.
11	length	Num	8	BEST12.	BEST32.
3	make	Char	13	$13.	$13.
2	normalized-losses	Char	3	$3.	$3.
16	num-of-cylinders	Char	6	$6.	$6.
6	num-of-doors	Char	4	$4.	$4.
23	peak-rpm	Char	4	$4.	$4.
26	price	Char	5	$5.	$5.
20	stroke	Char	4	$4.	$4.
1	symboling	Num	8	BEST12.	BEST32.
10	wheel-base	Num	8	BEST12.	BEST32.
12	width	Num	8	BEST12.	BEST32.

```
/*Descriptive Statistics for Numeric Variables*/
ods noproctitle;
ods graphics / imagemap=on;

proc means data=auto2 chartype mean std min max n nmiss vardef=df;
    var symboling wheelbase length width height curbweight
        enginesize bore stroke compressionratio horsepower peakrpm citympg highwaympg
        price;
run;
```

Variable	Mean	Std Dev	Minimum	Maximum	N	N Miss
symboling	0.8341463	1.2453068	-2.0000000	3.0000000	205	0
wheelbase	98.7565854	6.0217757	86.6000000	120.9000000	205	0
length	174.0492683	12.3372885	141.1000000	208.1000000	205	0
width	65.9078049	2.1452039	60.3000000	72.3000000	205	0
height	53.7248780	2.4435220	47.8000000	59.8000000	205	0
curbweight	2555.57	520.6802035	1488.00	4066.00	205	0
enginesize	126.9073171	41.6426934	61.0000000	326.0000000	205	0
bore	3.3297512	0.2735387	2.5400000	3.9400000	201	4
stroke	3.2554229	0.3167175	2.0700000	4.1700000	201	4
compressionratio	10.1425366	3.9720403	7.0000000	23.0000000	205	0
horsepower	104.2561576	39.7143688	48.0000000	288.0000000	203	2
peakrpm	5125.37	479.3345598	4150.00	6600.00	203	2
citympg	25.2195122	6.5421417	13.0000000	49.0000000	205	0
highwaympg	30.7512195	6.8864431	16.0000000	54.0000000	205	0
price	13207.13	7947.07	5118.00	45400.00	201	4

Cleaned Data - removed all rows/records with missing data

```
/*Delete rows with missing data - equates to 10 records removed */
data auto_clean;
    set auto;

    if nmiss(of _numeric_, 1) + cmiss(of _character_, '?') then
        delete;
run;

/*Descriptive Statistics for Numeric Variables*/
```

```
ods noproctitle;
ods graphics / imagemap=on;

proc means data=WORK.AUTO_CLEAN chartype mean std min max n nmiss vardef=df;
    var symboling wheelbase length width height curbweight
        enginesize bore stroke compressionratio horsepower peakrpm citympg highwaympg
        price;
run;
```

Variable	N	N Miss	Mean	Std Dev	Minimum	Maximum
symboling	195	0	0.7948718	1.2306123	-2.0000000	3.0000000
wheelbase	195	0	98.8964103	6.1320383	86.6000000	120.9000000
length	195	0	174.2569231	12.4764434	141.1000000	208.1000000
width	195	0	65.8861538	2.1324839	60.3000000	72.0000000
height	195	0	53.8615385	2.3967778	47.8000000	59.8000000
curbweight	195	0	2559.00	524.7157994	1488.00	4066.00
enginesize	195	0	127.9384615	41.4339159	61.0000000	326.0000000
bore	195	0	3.3293846	0.2718657	2.5400000	3.9400000
stroke	195	0	3.2503077	0.3141145	2.0700000	4.1700000
compressionratio	195	0	10.1949744	4.0621088	7.0000000	23.0000000
horsepower	195	0	103.2717949	37.8697302	48.0000000	262.0000000
peakrpm	195	0	5099.49	468.2713809	4150.00	6600.00
citympg	195	0	25.3743590	6.4013819	13.0000000	49.0000000
highwaympg	195	0	30.8410256	6.8293151	16.0000000	54.0000000
price	195	0	13248.02	8056.33	5118.00	45400.00

Deleted Data -

```
/*Data set created - Deleted rows with missing data*/
data auto_clean_missing;
   set auto_clean;
   if nmiss(of _numeric_, 1) + cmiss(of _character_, '?') then
      output;
run;

./*Analyze the deleted records*/
/*Descriptive Statistics for Numeric Variables*/
ods noproctitle;
ods graphics / imagemap=on;

proc means data=auto_clean_missing chartype mean std min max n nmiss vardef=df;
    var symboling wheelbase length width height curbweight
        enginesize bore stroke compressionratio horsepower peakrpm citympg highwaympg
        price;
run;
```

Variable	N	N Miss	Mean	Std Dev	Minimum	Maximum
symboling	10	0	1.6000000	1.3498971	0	3.0000000
wheelbase	10	0	96.0300000	1.6492759	94.5000000	99.5000000
length	10	0	170.0000000	8.6912472	155.9000000	181.5000000
width	10	0	66.3300000	2.4653375	63.6000000	72.3000000
height	10	0	51.0600000	1.7933209	49.6000000	55.2000000
curbweight	10	0	2488.60	452.8291804	1874.00	3366.00
enginesize	10	0	106.8000000	42.7597942	70.0000000	203.0000000
bore	6	4	3.3416667	0.3535770	3.0300000	3.9400000
stroke	6	4	3.4216667	0.3871649	3.1100000	3.9000000
compressionratio	10	0	9.1200000	0.8430105	7.0000000	10.0000000
horsepower	8	2	128.2500000	71.3016730	70.0000000	288.0000000
peakrpm	8	2	5756.25	282.1315549	5400.00	6000.00
citympg	10	0	22.2000000	8.7279882	16.0000000	38.0000000
highwaympg	10	0	29.0000000	8.1240384	22.0000000	43.0000000
price	6	4	11878.33	2399.51	9295.00	15645.00

We compared the descriptive stats for Original dataset against the Cleaned dataset and the deleted record dataset. The means of the deleted items are with in range of the original data. We conclude it is safe to delete the 10 records that contain missing data points.

Exploratory Analysis

Scatter Plot Matrix of all numerical variables

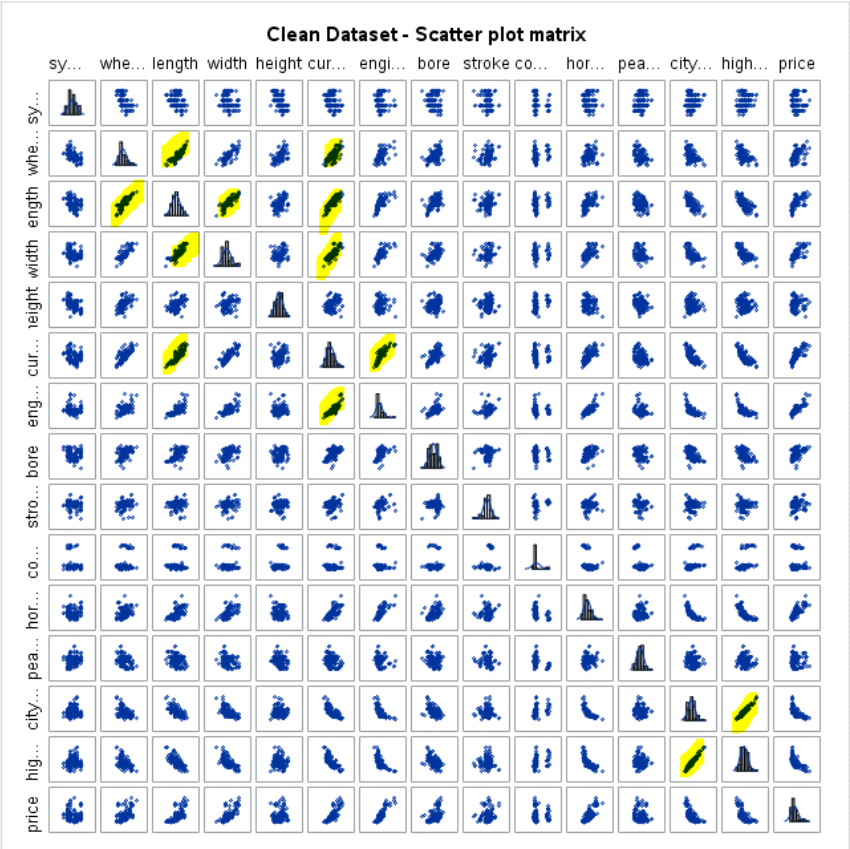
```
/*Plot all numeric variables against each other*/
options validvarname=any;
ods noproctitle;
ods graphics / imagemap=on;

/* Scatter plot matrix macro */
%macro scatterPlotMatrix(xVars=, title=, groupVar=);
   proc sgscatter data=WORK.AUTO_Clean;
      matrix &xVars / %if(&groupVar ne %str()) %then
         %do;
            group=&groupVar legend=(sortorder=ascending) %end;
```

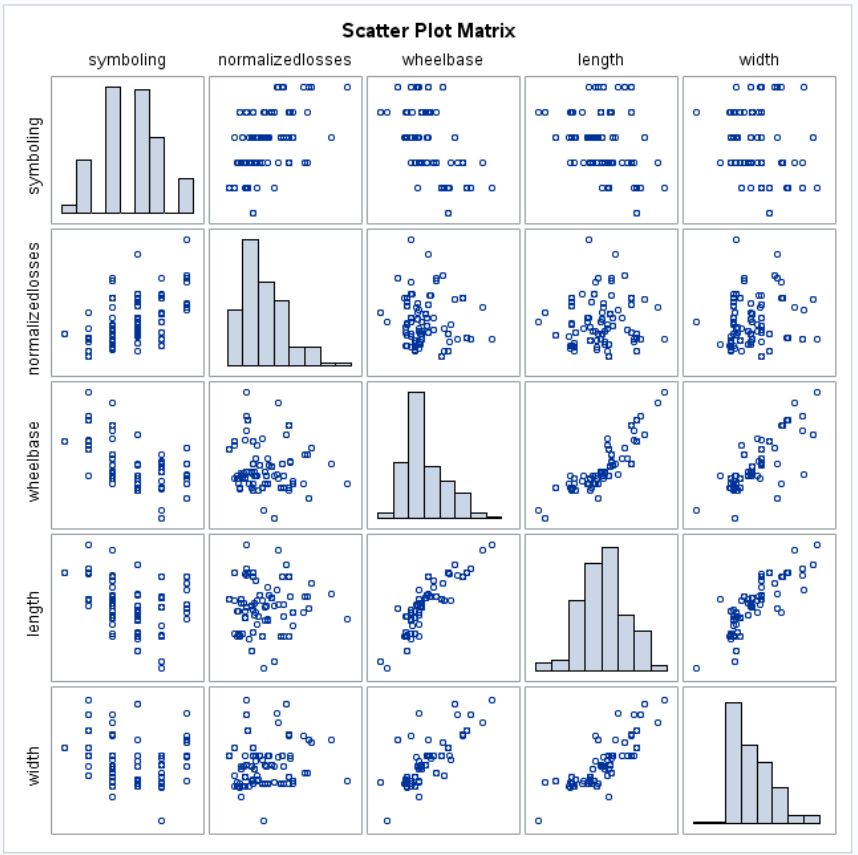
```
        diagonal=(histogram normal);
        title &title;
    run;

    title;
%mend scatterPlotMatrix;

%scatterPlotMatrix(xVars=symboling  wheelbase length width
    height curbweight enginesize bore stroke compressionratio horsepower peakrpm
    citympg highwaympg price, title="Clean Dataset - Scatter plot matrix",
    groupVar=);
```




```
proc corr data=auto_clean plots=matrix (histogram);
run;
```

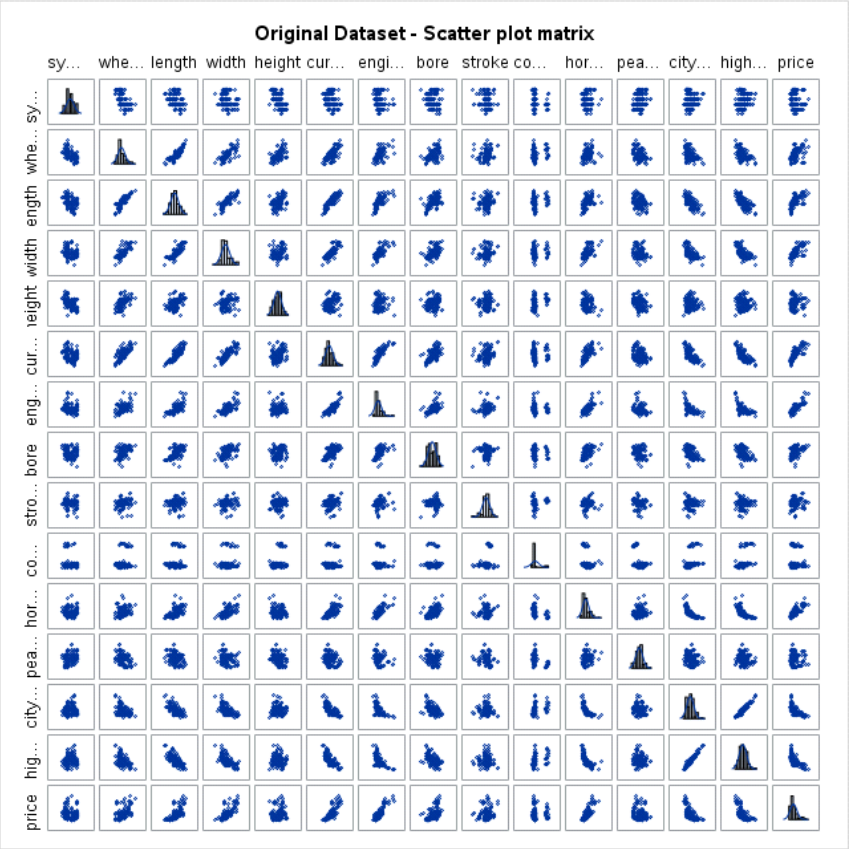


Pearson Correlation Coefficients, N = 195
Prob > |r| under H0: Rho=0

	symboling	wheelbase	length	width	height	curbweight	enginesize	bore	stroke	compressionratio	horsepower	peakrpm	citympg	highwaympg	price
symboling	1.00000	-0.53557 <.0001	-0.36306 <.0001	-0.24858 0.0005	-0.51754 <.0001	-0.23035 0.0012	-0.06828 0.3429	-0.14582 0.0419	-0.01197 0.8681	-0.18126 0.0112	0.07265 0.3128	0.23060 0.0012	0.01176 0.8704	0.07951 0.2692	-0.08412 0.2423
wheelbase	-0.53557 <.0001	1.00000	0.87922 <.0001	0.81901 <.0001	0.59250 <.0001	0.78272 <.0001	0.56970 <.0001	0.49823 <.0001	0.17172 0.0164	0.24773 0.0005	0.37554 <.0001	-0.35233 <.0001	-0.49913 <.0001	-0.56635 <.0001	0.58579 <.0001
length	-0.36306 <.0001	0.87922 <.0001	1.00000	0.85808 <.0001	0.49622 <.0001	0.88166 <.0001	0.68748 <.0001	0.60944 <.0001	0.11866 0.0985	0.16017 0.0253	0.58381 <.0001	-0.28099 <.0001	-0.68966 <.0001	-0.71932 <.0001	0.69533 <.0001
width	-0.24858 0.0005	0.81901 <.0001	0.85808 <.0001	1.00000	0.31583 <.0001	0.86731 <.0001	0.74032 <.0001	0.54431 <.0001	0.18643 0.0091	0.19100 0.0075	0.61678 <.0001	-0.25163 0.0004	-0.64710 <.0001	-0.69222 <.0001	0.75427 <.0001
height	-0.51754 <.0001	0.59250 <.0001	0.49622 <.0001	0.31583 <.0001	1.00000	0.30773 <.0001	0.03129 0.6642	0.18928 0.0080	-0.05552 0.4407	0.26116 0.0002	-0.08441 0.2407	-0.26408 0.0002	-0.10237 0.1544	-0.15119 0.0349	0.13829 0.0539
curbweight	-0.23035 0.0012	0.78272 <.0001	0.88166 <.0001	0.86731 <.0001	0.30773 <.0001	1.00000	0.85757 <.0001	0.64581 <.0001	0.17279 0.0157	0.15538 0.0301	0.76029 <.0001	-0.27894 <.0001	-0.77217 <.0001	-0.81271 <.0001	0.83573 <.0001
enginesize	-0.06828 0.3429	0.56970 <.0001	0.68748 <.0001	0.74032 <.0001	0.03129 0.6642	0.85757 <.0001	1.00000	0.58309 <.0001	0.21199 0.0029	0.02462 0.7327	0.84269 <.0001	-0.21901 0.0021	-0.71062 <.0001	-0.73214 <.0001	0.88894 <.0001
bore	-0.14582 0.0419	0.49823 <.0001	0.60944 <.0001	0.54431 <.0001	0.18928 0.0080	0.64581 <.0001	0.58309 <.0001	1.00000	-0.06679 0.3535	0.00306 0.9662	0.56853 <.0001	-0.27766 <.0001	-0.59195 <.0001	-0.60004 <.0001	0.54687 <.0001
stroke	-0.01197 0.8681	0.17172 0.0164	0.11866 0.0985	0.18643 0.0091	-0.05552 0.4407	0.17279 0.0157	0.21199 0.0029	-0.06679 0.3535	1.00000	0.19988 0.0051	0.10004 0.1641	-0.06830 0.3428	-0.02764 0.7013	-0.03645 0.6129	0.09375 0.1924
compressionratio	-0.18126 0.0112	0.24773 0.0005	0.16017 0.0253	0.19100 0.0075	0.26116 0.0002	0.15538 0.0301	0.02462 0.7327	0.00306 0.9662	0.19988 0.0051	1.00000	-0.21440 0.0026	-0.44458 <.0001	0.33141 <.0001	0.26794 0.0002	0.06950 0.3343
horsepower	0.07265 0.3128	0.37554 <.0001	0.58381 <.0001	0.61678 <.0001	-0.08441 0.2407	0.76029 <.0001	0.84269 <.0001	0.56853 <.0001	0.10004 0.1641	-0.21440 0.0026	1.00000	0.10565 0.1416	-0.83412 <.0001	-0.81292 <.0001	0.81103 <.0001
peakrpm	0.23060 0.0012	-0.35233 <.0001	-0.28099 <.0001	-0.25163 0.0004	-0.26408 0.0002	-0.27894 <.0001	-0.21901 0.0021	-0.27766 <.0001	-0.06830 0.3428	-0.44458 <.0001	0.10565 0.1416	1.00000	-0.06949 0.3344	-0.01695 0.8141	-0.10433 0.1466
citympg	0.01176 0.8704	-0.49913 <.0001	-0.68966 <.0001	-0.64710 <.0001	-0.10237 0.1544	-0.77217 <.0001	-0.71062 <.0001	-0.59195 <.0001	-0.02764 0.7013	0.33141 <.0001	-0.83412 <.0001	-0.06949 0.3344	1.00000	0.97235 <.0001	-0.70268 <.0001
highwaympg	0.07951 0.2692	-0.56635 <.0001	-0.71932 <.0001	-0.69222 <.0001	-0.15119 0.0349	-0.81271 <.0001	-0.73214 <.0001	-0.60004 <.0001	-0.03645 0.6129	0.26794 0.0002	-0.81292 <.0001	-0.01695 0.8141	0.97235 <.0001	1.00000	-0.71559 <.0001
price	-0.08412 0.2423	0.58579 <.0001	0.69533 <.0001	0.75427 <.0001	0.13829 0.0539	0.83573 <.0001	0.88894 <.0001	0.54687 <.0001	0.09375 0.1924	0.06950 0.3343	0.81103 <.0001	-0.10433 0.1466	-0.70268 <.0001	-0.71559 <.0001	1.00000

Initial look at Correlated variables

Highway MPG and City MPG
Length and Wheelbase
Width and Length
Wheelbase and Cure weight
Width and Curb weight
Engine Size and Curb weight
Engine Size and Horsepower



Residual Plots, Outliers and Leverage

Finding Subset of Variables

Look at VIF - variable selection

```
Proc reg data=auto_clean plots(label)=(rstudentbyleverage cooksd);
/*class     make fueltype aspiration  numofdoors  bodystyle   drivewheels enginelocation enginetype  numofcylinders fuelsyste m;*/
   Model price = symboling wheelbase   length  width   height  curbweight    enginesize    bore    stroke  compressionratio
   horsepower  peakrpm citympg highwaympg /VIF ;
   run;
quit;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-62068	16179	-3.84	0.0002	0
wheel-base	1	70.46712	103.12725	0.68	0.4953	7.79297
length	1	-89.73375	57.01356	-1.57	0.1173	9.86017
width	1	620.84626	256.36169	2.42	0.0164	5.82404
height	1	319.93882	141.95448	2.25	0.0254	2.25580
curb-weight	1	1.71246	1.72944	0.99	0.3234	16.04739
engine-size	1	126.67481	15.05815	8.41	<.0001	7.58580
bore	1	-918.71093	1206.84422	-0.76	0.4475	2.09777
stroke	1	-2962.97261	793.68468	-3.73	0.0003	1.21121
compression-ratio	1	239.72476	85.31439	2.81	0.0055	2.34042
horsepower	1	38.01528	18.10543	2.10	0.0371	9.16112
peak-rpm	1	2.08565	0.67290	3.10	0.0022	1.93482
city-mpg	1	-308.03512	181.91022	-1.69	0.0921	26.42459
highway-mpg	1	283.95609	163.94654	1.73	0.0850	24.42898

Correlation Analysis

7 With Variables:	stroke compression-ratio horsepower peak-rpm city-mpg highway-mpg price
9 Variables:	symboling normalized-losses wheel-base length width height curb-weight engine-size bore

Pearson Correlation Coefficients
Number of Observations

	symboling	normalized-losses	wheel-base	length	width	height	curb-weight	engine-size	bore
stroke	-0.00896 201	0.06563 160	0.16148 201	0.12974 201	0.18296 201	-0.05700 201	0.16893 201	0.20667 201	-0.05591 201
compression-ratio	-0.17852 205	-0.13265 164	0.24979 205	0.15841 205	0.18113 205	0.26121 205	0.15136 205	0.02897 205	0.00520 201
horsepower	0.07162 203	0.29577 164	0.35230 203	0.55500 203	0.64248 203	-0.11071 203	0.75103 203	0.81077 203	0.57727 199
peak-rpm	0.27457 203	0.26460 164	-0.36105 203	-0.28732 203	-0.21996 203	-0.32227 203	-0.26631 203	-0.24462 203	-0.26427 199
city-mpg	-0.03582 205	-0.25850 164	-0.47041 205	-0.67091 205	-0.64270 205	-0.04864 205	-0.75741 205	-0.65366 205	-0.59458 201
highway-mpg	0.03461 205	-0.21077 164	-0.54408 205	-0.70466 205	-0.67722 205	-0.10736 205	-0.79746 205	-0.67747 205	-0.59457 201
price	-0.08239 201	0.20325 164	0.58464 201	0.69063 201	0.75127 201	0.13549 201	0.83441 201	0.87234 201	0.54344 197

Residuals - Price (response) vs all numeric variables

```
Proc reg data=auto  plots(label)=(rstudentbyleverage cooksd);
Model price = 'wheel-base'n  length  width   height 'curb-weight'n 'engine-size'n
bore stroke
 'compression-ratio'n horsepower  'peak-rpm'n 'city-mpg'n 'highway-mpg'n; run;
quit;
```

Model: MODEL1
Dependent Variable: price

Number of Observations Read	205
Number of Observations Used	195
Number of Observations with Missing Values	10

Analysis of Variance

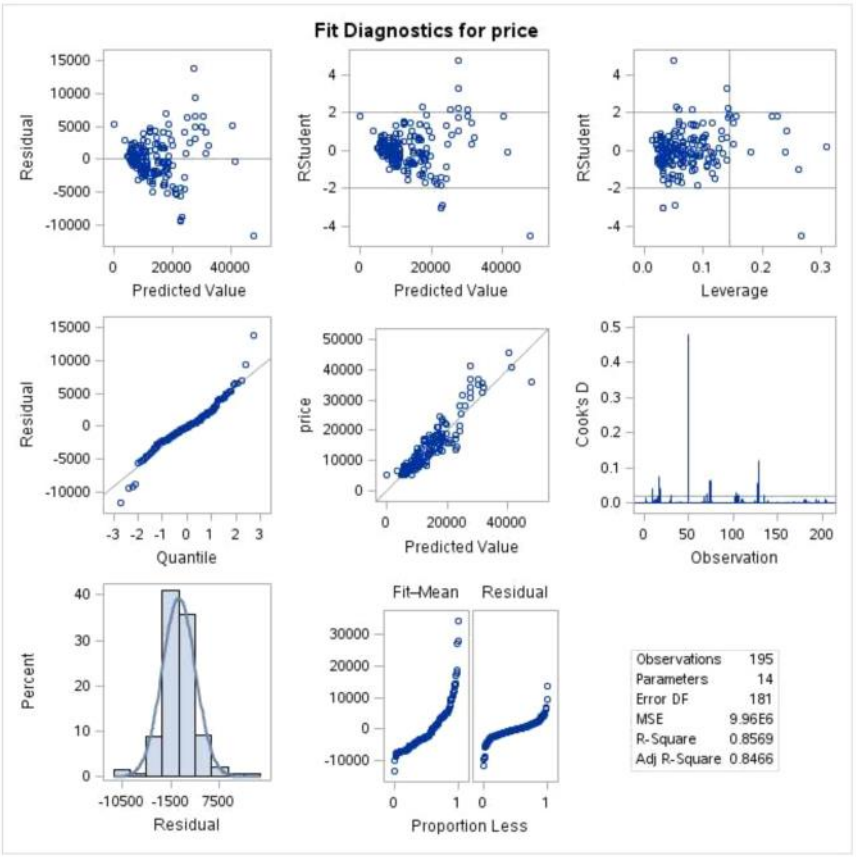
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	10789551512	829965501	83.37	<.0001
Error	181	1801912675	9955319		
Corrected Total	194	12591464187			

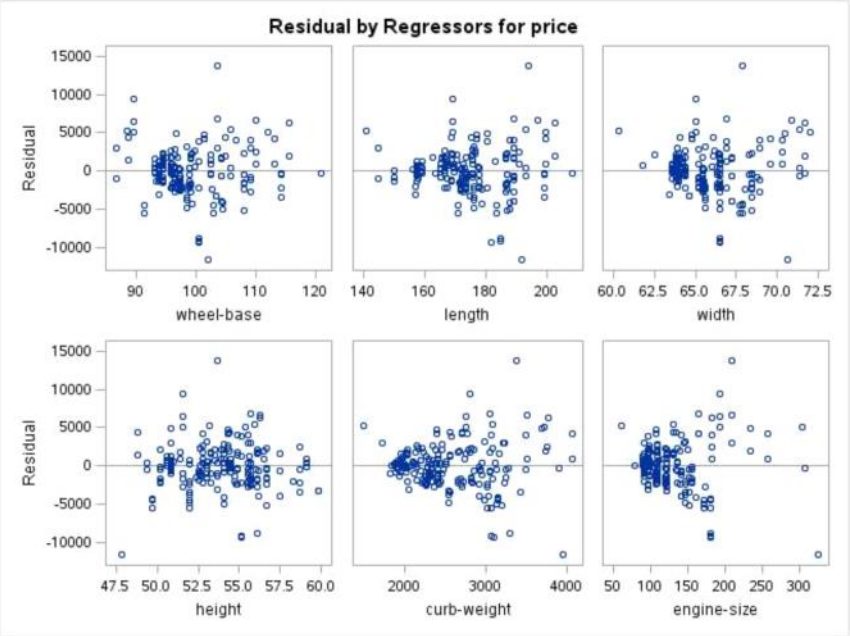
Root MSE	3155.20501	R-Square	0.8569
Dependent Mean	13248	Adj R-Sq	0.8466
Coeff Var	23.81644		

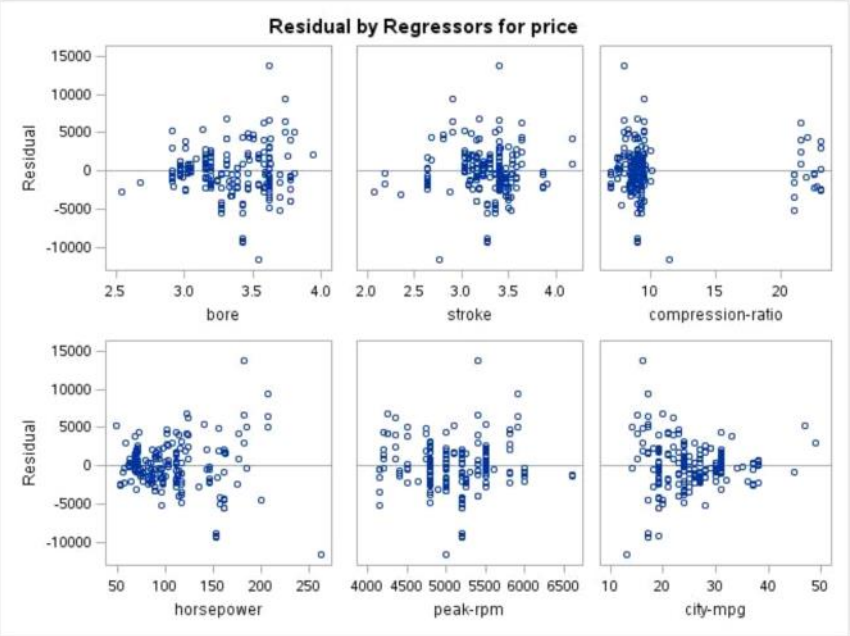
Parameter Estimates

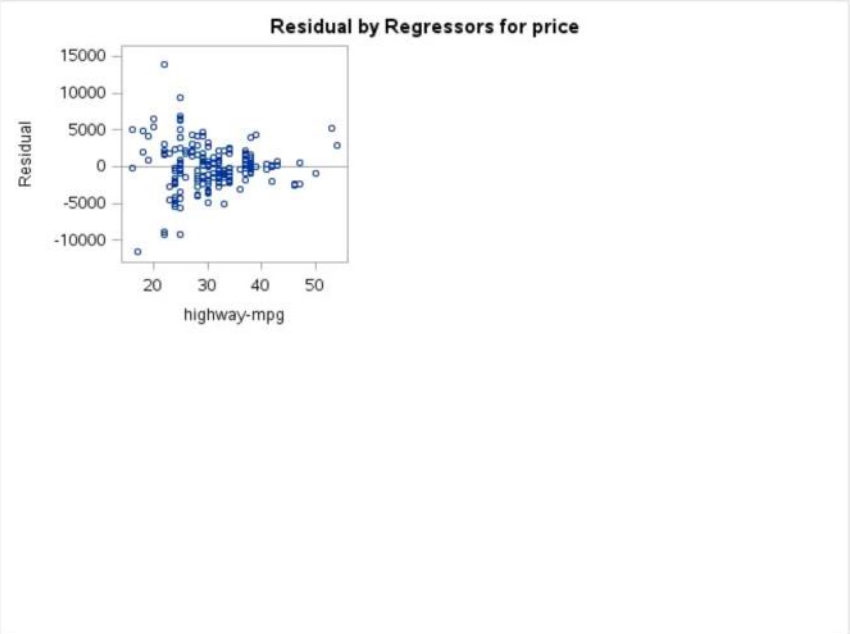
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-62068	16179	-3.84	0.0002
wheel-base	1	70.46712	103.12725	0.68	0.4953
length	1	-89.73375	57.01356	-1.57	0.1173
width	1	620.84626	256.36169	2.42	0.0164
height	1	319.93882	141.95448	2.25	0.0254
curb-weight	1	1.71246	1.72944	0.99	0.3234
engine-size	1	126.67481	15.05815	8.41	<.0001
bore	1	-918.71093	1206.84422	-0.76	0.4475
stroke	1	-2962.97261	793.68468	-3.73	0.0003
compression-ratio	1	239.72476	85.31439	2.81	0.0055
horsepower	1	38.01528	18.10543	2.10	0.0371
peak-rpm	1	2.08565	0.67290	3.10	0.0022
city-mpg	1	-308.03512	181.91022	-1.69	0.0921
highway-mpg	1	283.95609	163.94654	1.73	0.0850

Model: MODEL1
Dependent Variable: price


Fit Diagnostics for price
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
price
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 195
Parameters 14
Error DF 181
MSE 9.96E6
R-Square 0.8569
Adj R-Square 0.8466


Residual by Regressors for price
Residual
15000
10000
5000
0
-5000
-10000
90
100
110
120
wheel-base
140
160
180
200
length
60.0
62.5
65.0
67.5
70.0
72.5
width
47.5
50.0
52.5
55.0
57.5
60.0
height
2000
3000
4000
curb-weight
50
100
150
200
250
300
engine-size


Residual by Regressors for price
Residual
15000
10000
5000
0
-5000
-10000
2.5
3.0
3.5
4.0
bore
2.0
2.5
3.0
3.5
4.0
stroke
10
15
20
compression-ratio
Residual
15000
10000
5000
0
-5000
-10000
50
100
150
200
250
horsepower
4000
4500
5000
5500
6000
6500
peak-rpm
10
20
30
40
50
city-mpg




```
/*Variable Selection techinques - LARS*/
Proc glmselect data=auto seed=12;
   Class make fueltype aspiration  numofdoors  bodystyle   drivewheels enginelocation
enginetype  numofcylinders fuelsystem;
   Model price= symboling wheelbase   length  width   height  curbweight   enginesize
bore    stroke  compressionratio
   horsepower  peakrpm citympg highwaympg / selection=LARS (choose=cv stop=cv)
CVDETAILS;
Run;

quit;

/*Variable Selection techinques - LASSO*/
proc glmselect data=auto plots(stepaxis=number)=(criterionpanel ASEPlot) seed=1;
   partition fraction(test=.5);
   Model price= symboling wheelbase   length  width   height  curbweight   enginesize
bore    stroke  compressionratio
   horsepower  peakrpm citympg highwaympg /
```

```
      selection=lasso(choose=cv stop=cv) CVDETAILS;
run;
```

Data Set	WORK.AUTO
Dependent Variable	price
Selection Method	LASSO
Stop Criterion	Cross Validation
Choose Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	1

Number of Observations Read	205
Number of Observations Used	195
Number of Observations Used for Training	101
Number of Observations Used for Testing	94

Dimensions	
Number of Effects	13
Number of Parameters	13

LASSO Selection Summary

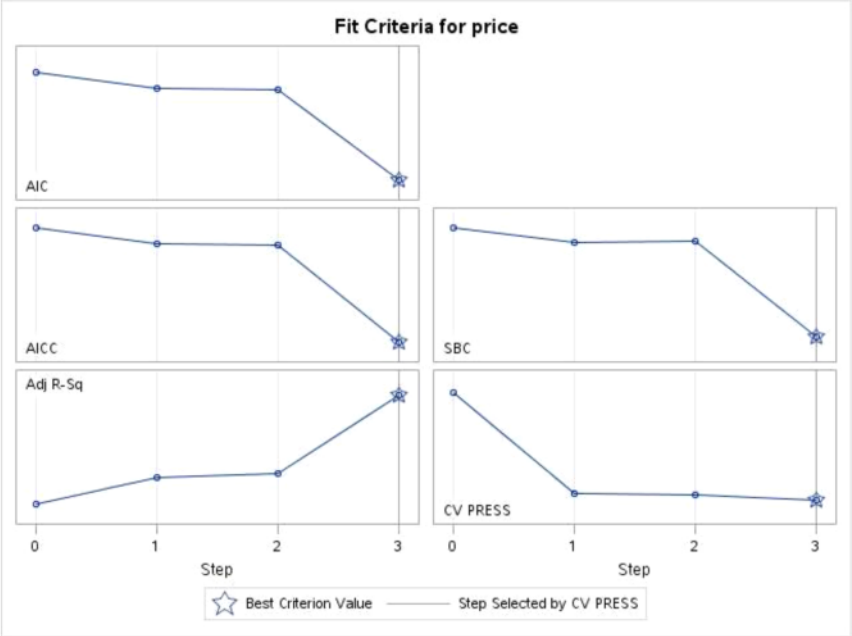
Step	Effect Entered	Effect Removed	Number Effects In	ASE	Test ASE	CV PRESS
0	Intercept		1	47723463.6	83371772.1	4880619142
1	engine-size		2	38325419.7	64811923.4	1387353594
2	curb-weight		3	36828943.7	62144682.9	1343506385
3	width		4	11092522.1	20149985.0	1138369863*

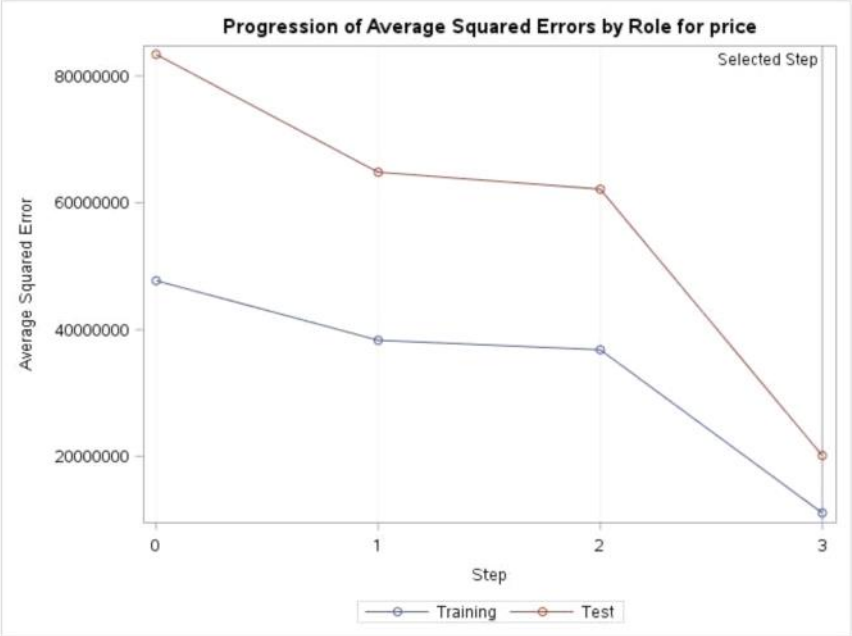
* Optimal Value of Criterion

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details

Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Entry	horsepower	1187674676	>	1138369863


Fit Criteria for price
AIC
AICC
SBC
Adj R-Sq
CV PRESS
0
1
2
3
Step
Best Criterion Value
Step Selected by CV PRESS



Selected Model

The selected model, based on Cross Validation, is the model at Step 3.

Effects:	Intercept width curb-weight engine-size

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	3	3699725089	1233241696	106.77
Error	97	1120344737	11549946	
Corrected Total	100	4820069825		

Root MSE	3398.52111
Dependent Mean	12668
R-Square	0.7676
Adj R-Sq	0.7604
AIC	1749.39996
AICC	1750.03154
SBC	1656.86044
ASE (Train)	11092522
ASE (Test)	20149985
CV PRESS	1138369863

Cross Validation Details

Cross Validation Details			
Index	Observations Fitted	Left Out	CV PRESS
Index	Fitted	Left Out	CV PRESS
1	80	21	119133638
2	82	19	287931227
3	76	25	309180778
4	82	19	180809767
5	84	17	241314454
Total			1138369863

Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	-57100
width	1	842.064938
curb-weight	1	1.752601
engine-size	1	78.413368

/*Variable Selection techinques - LARS*/

```
Proc glmselect data=auto seed =12;
Model price = 'wheel-base'n  length  width   height 'curb-weight'n 'engine-size'n bore
stroke 'compression-ratio'n
horsepower  'peak-rpm'n  'highway-mpg'n / selection=LARS ( choose = cv stop = cv)
CVDETAILS;
Run; quit;
```

Data Set	WORK.AUTO
Dependent Variable	price
Selection Method	LAR
Stop Criterion	Cross Validation
Choose Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	12

Number of Observations Read	205
Number of Observations Used	195

Dimensions	
Number of Effects	13
Number of Parameters	13

LAR Selection Summary

Step	Effect Entered	Number Effects In	CV PRESS
0	Intercept	1	1.29979E10
1	engine-size	2	2683767760
2	curb-weight	3	2494788526
3	horsepower	4	2436104408
4	width	5	2370747538
5	stroke	6	2332828477
6	height	7	2303294552
7	compression-ratio	8	2278796587
8	peak-rpm	9	2159812429
9	bore	10	2157666039*

* Optimal Value of Criterion

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details

Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Entry	length	2160896630	>	2157666039

Selected Model

The selected model, based on Cross Validation, is the model at Step 9.

Effects:	Intercept width height curb-weight engine-size bore stroke compression-ratio horsepower peak-rpm

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	9	10737696713	1193077413	119.07

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Error	185	1853767474	10020365	
Corrected Total	194	12591464187		

Root MSE	3165.49597
Dependent Mean	13248
R-Square	0.8528
Adj R-Sq	0.8456
AIC	3350.15983
AICC	3351.60246
SBC	3185.88983
CV PRESS	2157666039

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	156	39	300697634
2	147	48	505315529
3	155	40	726194735
4	162	33	116069829
5	160	35	509388312
Total			2157666039

Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	-57252
width	1	516.194183
height	1	240.585859
curb-weight	1	0.630809
engine-size	1	123.591466
bore	1	-495.332791
stroke	1	-2492.668400
compression-ratio	1	208.622672
horsepower	1	40.433048
peak-rpm	1	1.873483