Statistical Modeling and Analysis Results for Car Prices, Class Project for STAT 611

Giuliamaria Menara

May 9, 2021

Introduction

This report summerizes the statistical modeling and analysis results for the data set *Vehicle Data.xlsx*, containing car specifications. Analysis of the data is limited to knowledge and techniques learned in STAT611.

The purpose of this report is to document the analysis made to understand the relationship between the Manufacturer's Suggested Retail Price (MSRP) and 11 descriptive variables. In order to describe this relationship we will build models to predict MSRP given the car specifications in the dataset. Then the best model will be selected using different selection techniques.

This report is organized as follows:

- Section 1 contains a description of the dataset.
- Section 2 focuses on an exploratory analysis, needed to create a model suited to the dataset.
- In Section 3 we describe the model selection techniques performed for the predictive analysis.
- Section 4 contains the final remarks on the analysis.

At the end, an Appendix with the full SAS code is included.

1. Data Description

The dataset *Vehicles Data.xlsx* containes 428 observatios for the following variables:

Name	Type	Range
MSPR	quantitative	12280 – 192465
Engine Size	quantitative	1.3 – 8.3
Number of Cylinders	quantitative	-1 – 12
Horsepower	quantitative	73 – 500
MPG City	quantitative	10 – 60
MPG Highway	quantitative	12 – 66
Weight	quantitative	1850 – 7190
Wheelbase	quantitative	89 – 144
Length	quantitative	143 – 237
Width	quantitative	64 – 81
Drive Wheels	categorical	AWD, FWD, RWD
Vehicle Type	categorical	Minivan, Pickup, Sedan, Sports car, SUV, Wagon

Table 1: variables description

MSPR is the quantity we are interested in predicting, the others are the independent variables.

2. Eploratory Analysis

To create a model suited to the dataset, the data was examined and transformed in various ways.

<u>2.1 Imputation of missing values</u>

The first problem to adress was to fill in the 46 missing values. Those missing values were distributed across the following variables: MPG_City (14), MPG_Hwy (14), Weight (2), Wheelbase (2), Length (7), and Width (9). Imputation was performed by assigning to the missing value the average value of the variable after diving the cars by type. Table 2 below idetifies the inputed variables.

Variable	Vehicle type	Value imputed
MPG City	Sedan	21.76695
	Sports car	18.59574
	SUV	16.20339
	Wagon	20.96552

	Minivan	17.9
	Pickup	16.69565
MPG Highway	Sedan	29.36864
	Sports car	25.7234
	SUV	20.62712
	Wagon	27.7931
	Pickup	21.17391
Weight	Sedan	3319.687
Wheelbase	Sedan	107.177
Length	Sedan	186.1152
	Pickup	208.4737
Width	Sedan	70.42387
	Sports car	70.87234
	Pickup	74.26316

Table 2: imputed quatities

<u>2.2 Relationship between variables and MSRP</u>

Once our dataset is complete, we proceed by investigating the relationship between single preditors and the variable of interest, MSRP. There seems to be a linear dependence of MSRP from each predictor, and we do not observe major problems in the dataset. This been said, we point out that the plots show some unusual observations, so we will be looking for outliers, and maybe leverage/influential points.

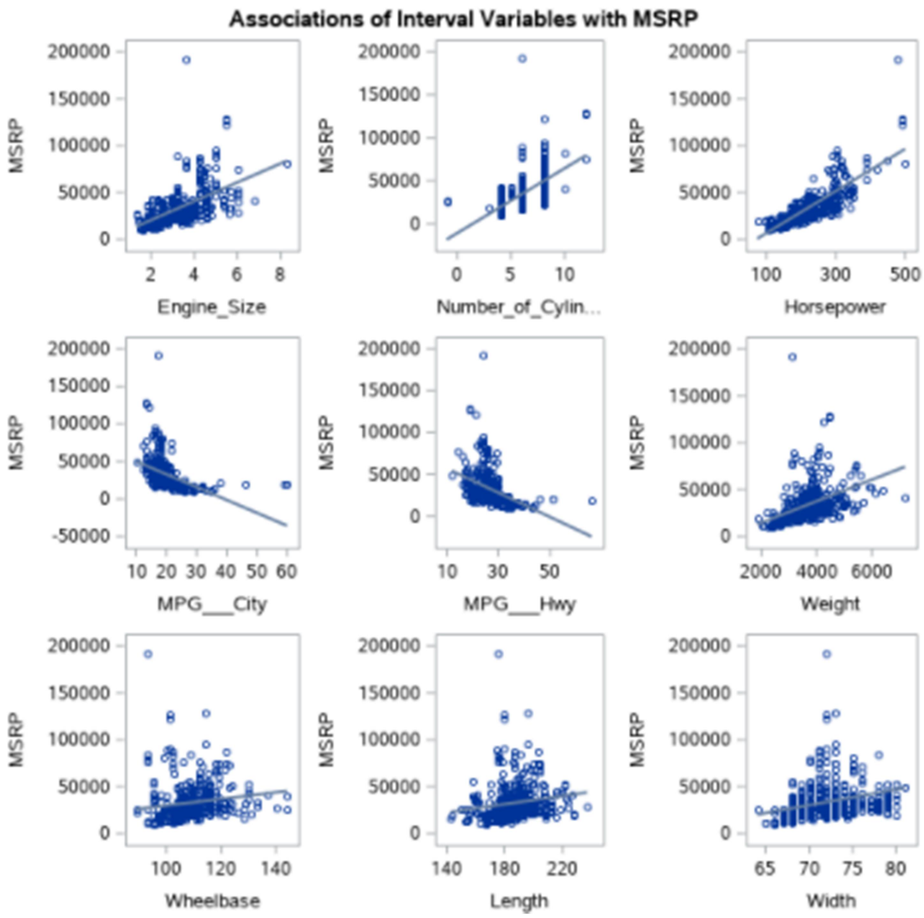

Figure 1

2.3 Adequecy of the fit correlation between predictors

The adequacy of the fit was examined using an ANOVA, which yielded an F-statistic of 133.72, a p-value less then 0.0001 and an Adjusted R^2 of 0.7367.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	1.196669E11	13296321624	133.72	<.0001
Error	418	41564724085	99437139		
Corrected Total	427	1.612316E11			

Root MSE	9971.81723	R-Square	0.7422
Dependent Mean	32775	Adj R-Sq	0.7367
Coeff Var	30.42521		

Table 3: ANOVA to determine accuracy of the fit

From here we can infer that there is not enough evidence to support that there is no relatioship between MSRP and the predictors.

Next, the VIF statistic is used to determine whether there is a correlation between the variables.

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	39916	16328	2.44	0.0149	0
Engine_Size	Engine Size	1	-4218.29918	1284.12092	-3.28	0.0011	8.70235
Number_of_Cylinders	Number of Cylinders	1	2319.83408	711.05590	3.26	0.0012	5.71749
Horsepower	Horsepower	1	243.28216	12.28330	19.81	<.0001	3.34345
MPG___City	MPG - City	1	-247.77664	309.93389	-0.80	0.4245	10.89756
MPG___Hwy	MPG - Hwy	1	666.51632	303.67338	2.19	0.0287	12.55100
Weight	Weight	1	7.56660	1.67712	4.51	<.0001	6.95553
Wheelbase	Wheelbase	1	-556.37364	133.21651	-4.18	<.0001	5.25902
Length	Length	1	23.06882	77.63995	0.30	0.7665	5.37842
Width	Width	1	-611.84845	274.16157	-2.23	0.0262	4.01314

Based on the VIF values in Table 4 and correlation values computed in SAS (we omit this table here), MPG_City and MPG_Hwy are highly correlated. Given these signs of multicollinearity and considering the fact that MPG_City and MPG_Hwy are not meaningful predictors (we can infer this from the p-values in Table 4), it seems we should consider excluinge these variables from the model. Length and Width are also not highly significant but, since they do not seem problematic, for now we keep them in the model.

We point out that the negative parameter estimate for Engine Size is very suspicious: the plot in Figure 1 "Association of Interval Variables with MSRP" shows that a bigger engine influences positively the price of the car (as one would naturally expect). The VIF value for this variable is 0, so one possible explanation for this estimate is that there are some influential points affecting this parameter. Therefore, we will proceed our analysis by looking for and removing influential points.

2.4 Checking the Plots

The residual plot in Figure 2 has clustered observations and does not resemble a random spread of data. This confirms that the constant variance assumption is not met, and therefore a *log* transform will be applied to price. The RStudent plot shows one clear oulier and a few observations that could be outliers, and the QQ-plot confirms this. The leverage plot displays a couple of unusual observations, but geometrically they seem to "balance" each other. The Cook's D plot shows a single observation with high leverage, but the actual value of the data point is not large.

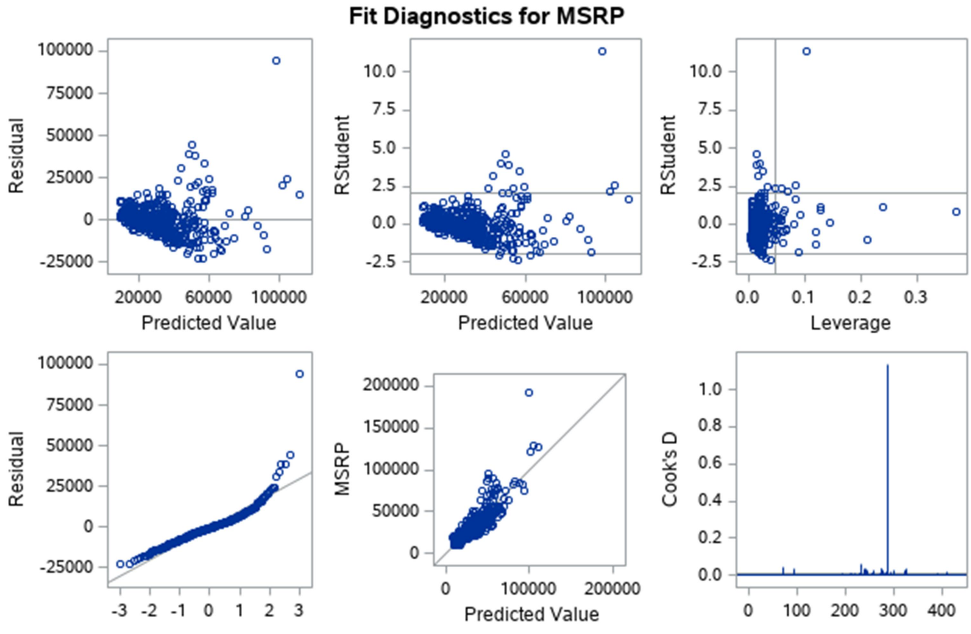

Figure 2

2.5 Model including Log(MSRP)

Figure 3 represents the model with a log transformation performed on MSRP. The residual plot displays a random spread, so the constant variance assumption is met, and the QQ-plot is nearly straight supporting the assumption of normality.

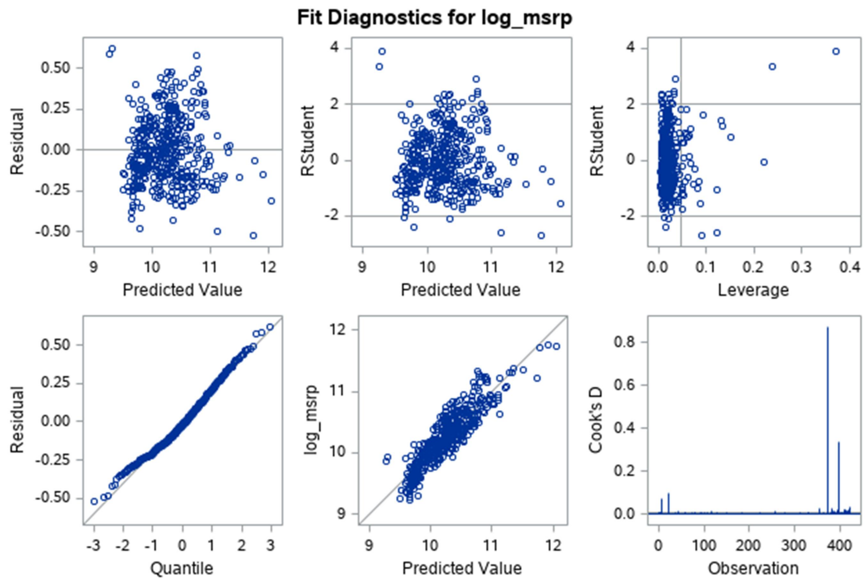

Figure 3

2.6 Delete outliers and influential points

To detect outliers more precisely we will rely on the R-student: we organize the data from the lowest to the highest studentized residual and we will discard as outliers all observations for which the RStudent absolute value is bigger then 3. After this process, 4 observations are deleted, and thus we are left with 424 observations.

With a similar process (using the statistics RStudent, Cook's distance and DFBetas) we are also able to delete influential points and we are left with 415 observations.

3. Model Selection

The updated dataset *Vehicle Data.xlsx* (recall that we imputed missing values, removed outliers and transformed MSRP) was used to generate the model using three model selection techniques: Backward, Forward and Stepwise selection. Table 5 contains the statistical output of the implemented model selection techniques.

	Backward	Forward	Stepwise
Steps	3	7	7
Variables	Engine size, Number of Cylinders, Horsepower, Weight, Width, Drive Wheels, Vehicle Type	Engine size, Number of Cylinders, Horsepower, Weight, Width, Drive Wheels, Vehicle Type	Engine size, Number of Cylinders, Horsepower, Weight, Width, Drive Wheels, Vehicle Type
F-value	182.22	182.22	182.22
R^2	0.8447	0.8447	0.8447
Adj R^2	0.8401	0.8401	0.8401
Root MSE	0.19312	0.19312	0.19312

Table 5

After a few tries, we saw that adding a metric to evaluate the model (like Mallow's Cp or Adjusted R^2) did not improve the proceudre outcome.

Based upon the analysis, we have that the methods select the same model, and thus the obtained statistics are the same. With an Adjusted R^2 value of 0.8401, we can say that the selected model fits the data.

4. Conclusion

To conclude, we fit the model with the selected variables and we obtain the parameter estimantes in Table 6.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Engine_Size	1	41.73820973	41.73820973	1119.15	<.0001
Number_of_Cylinders	1	6.22329545	6.22329545	166.87	<.0001
Horsepower	1	26.45138615	26.45138615	709.26	<.0001
Weight	1	1.57293763	1.57293763	42.18	<.0001
Width	1	0.47770041	0.47770041	12.81	0.0004
Drive_Wheels	2	1.30743186	0.65371593	17.53	<.0001
Vehicle_Type	5	3.78007048	0.75601410	20.27	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Engine_Size	1	0.92594682	0.92594682	24.83	<.0001
Number_of_Cylinders	1	0.42225384	0.42225384	11.32	0.0008
Horsepower	1	9.29457318	9.29457318	249.22	<.0001
Weight	1	3.33245458	3.33245458	89.35	<.0001
Width	1	0.53902473	0.53902473	14.45	0.0002
Drive_Wheels	2	1.54287896	0.77143948	20.69	<.0001
Vehicle_Type	5	3.78007048	0.75601410	20.27	<.0001

Table 6

All the variables are significant and the parameter for Engine Size became positive, accordinly to the relationship described in Figure 1.

Appendix: SAS Code

```
**importing data;
proc import datafile='/folders/myfolders/xlsx datasets/Vehicle Data.xlsx'
                  dbms=xlsx
                  out=vehicle_data
                  replace;
                  getnames=yes;

**imputation of missing values;
data vehicle_data;
      set vehicle_data;

      if MPG___City=' ' and Vehicle_Type='Sedan' then MPG___City=21.76695;
      if MPG___City=' ' and Vehicle_Type='Sports Car' then MPG___City=18.59574;
      if MPG___City=' ' and Vehicle_Type='SUV' then MPG___City=16.20339;
      if MPG___City=' ' and Vehicle_Type='Wagon' then MPG___City=20.96552;
      if MPG___City=' ' and Vehicle_Type='Minivan' then MPG___City=17.9;
      if MPG___City=' ' and Vehicle_Type='Pickup' then MPG___City=16.69565;

      if MPG___Hwy=' ' and Vehicle_Type='Sedan' then MPG___Hwy=29.36864;
      if MPG___Hwy=' ' and Vehicle_Type='Sports Car' then MPG___Hwy=25.7234;
      if MPG___Hwy=' ' and Vehicle_Type='SUV' then MPG___Hwy=20.62712;
      if MPG___Hwy=' ' and Vehicle_Type='Wagon' then MPG___Hwy=27.7931;
      if MPG___Hwy=' ' and Vehicle_Type='Pickup' then MPG___Hwy=21.17391;

      if Weight=' ' and Vehicle_Type='Sedan' then Weight=3319.687;

      if Wheelbase=' ' and Vehicle_Type='Sedan' then Wheelbase=107.177;

      if Length=' ' and Vehicle_Type='Sedan' then Length=186.1152;
      if Length=' ' and Vehicle_Type='Pickup' then Length=208.4737;

      if Width=' ' and Vehicle_Type='Sedan' then Width=70.42387;
      if Width=' ' and Vehicle_Type='Sports Car' then Width=70.87234;
      if Width=' ' and Vehicle_Type='Pickup' then Width=74.26316;

proc print data=vehicle_data;
run;


**Explore quantitative data;

*perform preliminary analysis;
proc sgplot data=vehicle_data;
    vbox MSRP / category=Drive_Wheels
                 connect=mean;
    title "MSRP Differences across Drive_Wheels";
run;

proc sgplot data=vehicle_data;
    vbox MSRP / category=Vehicle_type
                 connect=mean;
    title "MSRP Differences across Vehicle_type";
run;

options nolabel;
proc sgscatter data= vehicle_data;
      plot MSRP*(Engine_Size Number_of_Cylinders Horsepower MPG___City MPG___Hwy Weight
      Wheelbase Length Width) / reg;
      title "Associations of Interval Variables with MSRP";
run;
```

```
*check collinearity;
proc reg data=vehicle_data;
      model MSRP= Engine_Size Number_of_Cylinders Horsepower MPG___City MPG___Hwy
      Weight Wheelbase Length Width/VIF;
run;

proc corr data=vehicle_data
      nosimple
      best=4;
      var Engine_Size Number_of_Cylinders Horsepower MPG___City MPG___Hwy Weight
      Wheelbase Length Width;
      title "Correlations and Scatter Plot Matrix of Predictors";
run;

/*candidates for removal: MPG___City: VIF= 10.89756 and MPG___Hwy: VIF= 12.55100 */

*try log transformation for msrp;
data log_vehicle_data;
      set vehicle_data;
      log_msrp = log(MSRP);
run;

*detect outliers;
proc reg data=log_vehicle_data plots=all;
      model log_msrp= Engine_Size Number_of_Cylinders Horsepower MPG___City MPG___Hwy
      Weight Wheelbase Length Width/r;
      output out=log_vehicle_data predicted=predicted residual=resid student=studresid;
run;

proc sort data=log_vehicle_data; by studresid;
run;

*remove outliers;
      data log_vehicle_data;
      set log_vehicle_data;
      if studresid > 3 then delete;
run;

proc print data=log_vehicle_data;
run;

*find infulential points;
ods graphics on;
ods output RSTUDENTBYPREDICTED=Rstud
           COOKSDPLOT=Cook
           DFFITSPLOT=Dffits
           DFBETASPANEL=Dfbs;
proc reg data=log_vehicle_data
         plots(only label)=
              (RSTUDENTBYPREDICTED
               COOKSD
               DFFITS
               DFBETAS);
      SigLimit: model log_msrp = Engine_Size Number_of_Cylinders Horsepower  MPG___Hwy
      Weight Wheelbase  Width;
      title 'SigLimit Model - Plots of Diagnostic Statistics';
run;
quit;

proc print data=Rstud;
run;

proc print data=Cook;
run;

proc print data=Dffits;
run;
```

```
proc print data=Dfbs;
run;

data Dfbs01;
      set Dfbs (obs=424);
run;

data Dfbs02;
      set Dfbs (firstobs=425);
run;

data Dfbs2;
      update Dfbs01 Dfbs02;
      by Observation;
run;

data influential; *merge datasets from above;
      merge Rstud
              Cook
              Dffits
              Dfbs2;
      by observation;

      if (ABS(Rstudent)>3) or (Cooksdlabel ne ' ') or Dffitsout then flag=1;
      *flag observations that have exceeded at least one cutpoint;

      array dfbetas{*} _dfbetasout: ;
      do i=2 to dim(dfbetas);
            if dfbetas{i} then flag=1;
      end;

      if ABS(Rstudent)<=3 then RStudent=.;
      *set to missing values of influence statistics for those that have not exceeded
      cutpoints;
      if Cooksdlabel eq ' ' then CooksD=.;

      if flag=1; *subset only observations that have been flagged;
      drop i flag;
run;

title;
proc print data=influential;
      id observation;
      var Rstudent CooksD Dffitsout _dfbetasout:;
run;

*remove influential points;
data vehicle_data_inf;
      set log_vehicle_data;
      if _n_ in (70,161,219,229,247,274,300,323,326) then delete;
run;

proc print data=vehicle_data_inf;
run;

**Model selection;

*backward;
proc glmselect data=vehicle_data_inf plots=all;
      class Drive_Wheels Vehicle_type; /* generates dummy variables internally */
      model log_msrp= Engine_Size Number_of_Cylinders Horsepower MPG___Hwy Weight
      Wheelbase Length Width Drive_Wheels Vehicle_type/ selection=backward;
run;
quit;
```

```
*forward;
proc glmselect data=vehicle_data_inf plots=all;
      class Drive_Wheels Vehicle_type; /* generates dummy variables internally */
      model log_msrp= Engine_Size Number_of_Cylinders Horsepower MPG___Hwy Weight
      Wheelbase Length Width Drive_Wheels Vehicle_type/ selection=forward;
run;
quit;

*stepwise;
proc glmselect data=vehicle_data_inf plots=all;
      class Drive_Wheels Vehicle_type; /* generates dummy variables internally */
      model log_msrp= Engine_Size Number_of_Cylinders Horsepower MPG___Hwy Weight
      Wheelbase Length Width Drive_Wheels Vehicle_type/ selection=stepwise;
run;
quit;


**visualize plots for selected model;
proc glm data=vehicle_data_inf plots=all;
      class Drive_Wheels Vehicle_type;
      model log_msrp= Engine_Size Number_of_Cylinders Horsepower Weight Width
      Drive_Wheels Vehicle_type;
      output out=vehicle_data predicted=predicted residual=resid student=studresid;
run;
quit;
```