AUGUST 19, 2016

MSBA 6120 SUMMER 2016 PROJECT

Fuel Efficiency of Vehicles

Submitted by:

AAYUSH AGRAWAL

WENRUI LIU

BYRON TANG

RUSSELL NOUR

Contents

Executive Summary

In this report, we examined factors that affected a vehicle's fuel efficiency in miles per gallon. Data on vehicles from 2011 to 2016 were obtained from the United States federal government. Relationships between vehicle attributes and fuel efficiency in the data were described and a model was developed to explain the variation in fuel efficiency as a result of those attributes. The model shows that fuel efficiency is affected by a number of factors such as engine size (displacement and number of cylinders), vehicle class, the type of transmission, and the number of gears. *(Appendix 1.1 - 1.2)*

Introduction

For both environmental and financial reasons, fuel efficiency is often an important factor when consumers decide to purchase a vehicle, especially when gas prices are high. Furthermore, fuel efficiency is tested and regulated by federal and state governments in order to limit the environmental impacts of cars on the roads.

Fuel efficiency is measured in terms of distance travelled per amount of gasoline used. In the US, it is measured by miles per gallon or MPG. The Environmental Protection Agency (EPA) requires automakers to report MPG ratings for city driving, highway driving, and combined *(Appendix 2.1 - 2.3)*. The MPG rating of a vehicle depends on many factors. This exercise used government data to describe factors affecting a vehicle's MPG rating and find a model to predict MPG based on vehicle attributes. Fuel economy of a vehicle depends on various attributes. The purpose of this exercise is to find and study some of these attributes to determine their effect upon fuel efficiency in automobiles.

The combined miles per gallon reflects fuel efficiency and is our response variable - in other words, we described, tested, and modelled the relationships between combined MPG and the rest of the variables. We hypothesized that the following factors may affect a vehicle's fuel efficiency. These factors were tested and presented in the analyses part of this report.

Response variable

Combined MPG - the overall fuel efficiency of a vehicle (in miles per gallon)

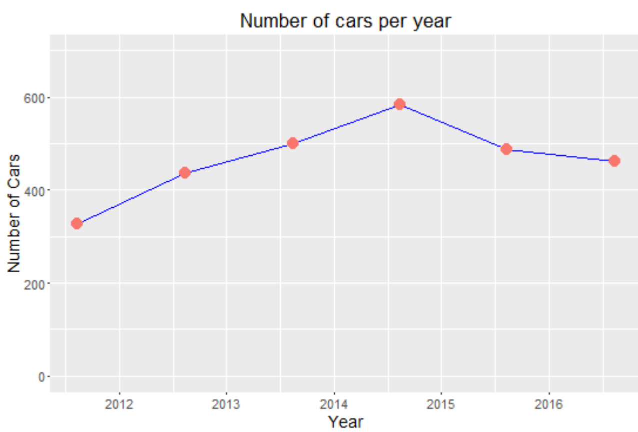
Predictor variables

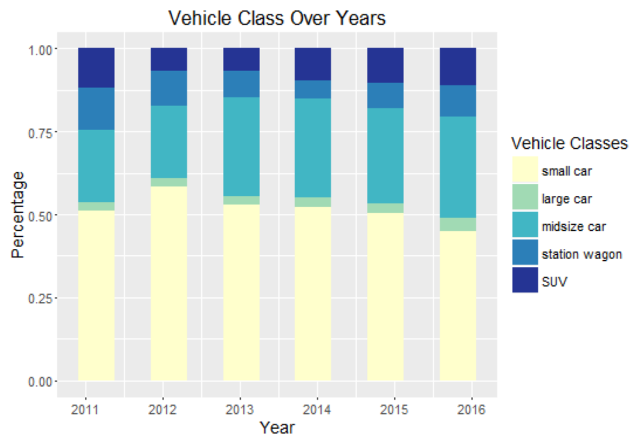
- Engine displacement (indicates engine size and power)
- Drive type - two-wheel drive or four-wheel drive
- Size and type of vehicle (e.g., midsize car)
- Smart Way certification from the EPA for low emissions
- The year in which the car was tested (e.g., 2015)
- Type of transmission (e.g., manual or automatic)
- Number of transmissions or gears in the engine
- Type of fuel the vehicle uses (e.g., gasoline)

Analyses

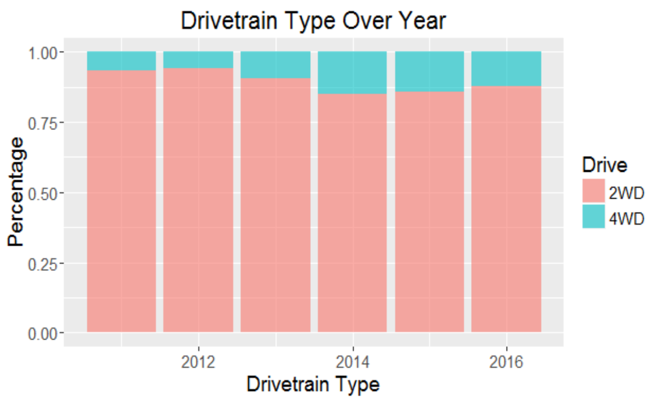
Vehicle Market Overview

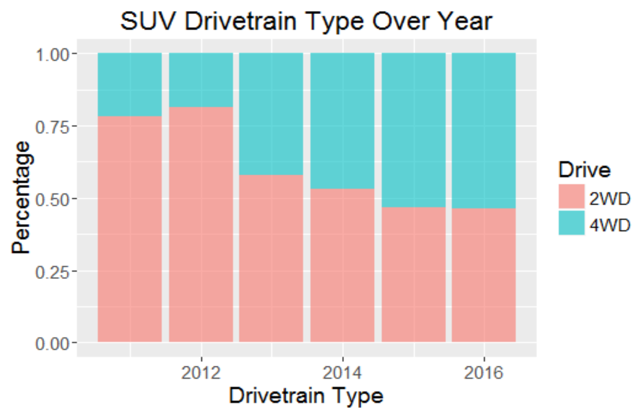
New vehicle launches grew steadily in the past years but peaked in year 2014. Small sedans are consistently the largest share in the market. However, the share of medium size sedans continues to grow and take market share from small sedans. *(Appendix 3)*





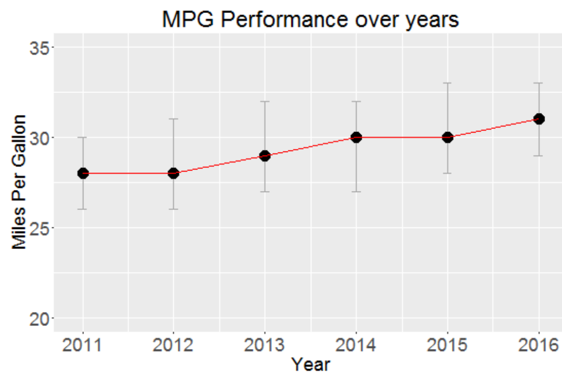
Two-wheel drive is the most common drivetrain type, but more and more four-wheel drive vehicles are coming to the market in recent years. It is a trend that manufacturers are adopting four-wheel drive in their SUVs to meet the needs for various road conditions.





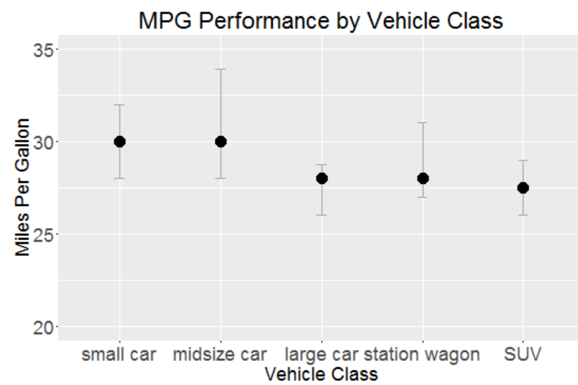
Miles per Gallon by Year

The median combined miles per gallon increases slightly every year. The difference is not statistically significant from year to year, but median combined MPG in 2015 and 2016 were significantly higher when compared to 2011. The chart shows the trend that combined MPG increases by year. *(Appendix 4.1)*



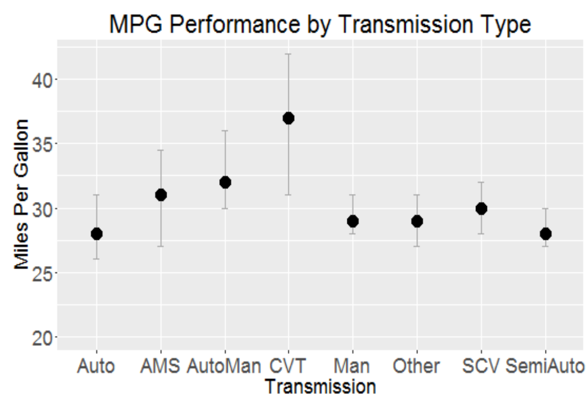
Miles per Gallon by Vehicle Class

Small and medium size cars have the highest median of combined MPG ratings. For both small and midsize cars, the median of combined MPG rating is significantly better than other types of cars, including large cars, station wagons, and SUVs. *(Appendix 4.2)*



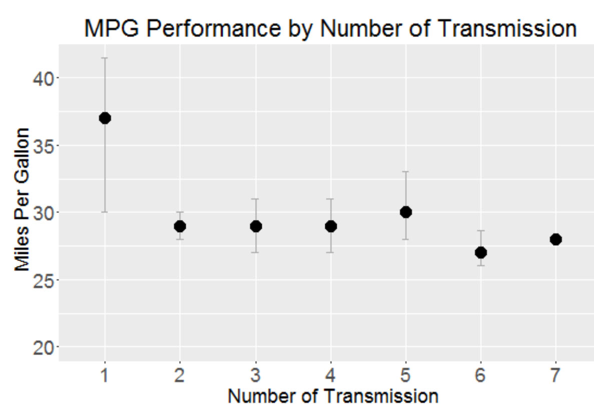
Miles per Gallon by Transmission Type

Among all transmission types, continuously variable transmission (CVT) is the most efficient with a significantly higher MPG, getting on average an additional 5 to 8 miles per gallon than other transmission types. *(Appendix 4.3)*



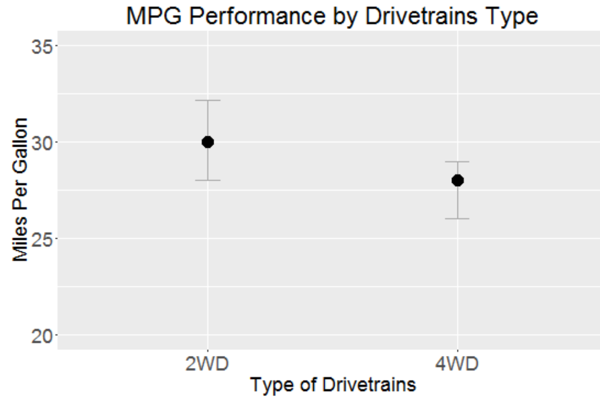
Miles per Gallon by Number of Transmissions

A single transmission delivers the best MPG performance. Vehicles with one transmission get 5 to 9 miles more miles per gallon than those with more transmissions. Variation also exists between different numbers of transmissions but are much less obvious. *(Appendix 4.4)*



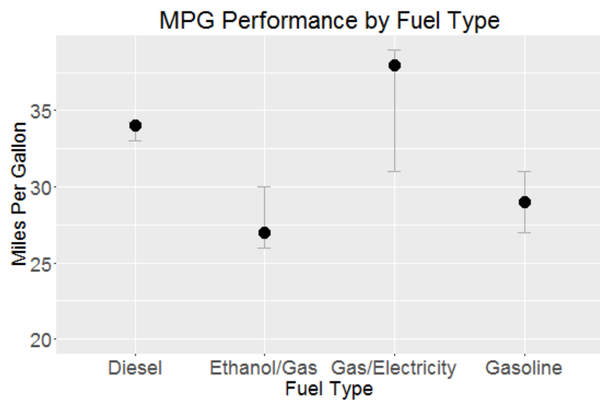
Miles per Gallon by Drivetrain Type

Vehicles with four-wheel drive are less efficient than those with two-wheel drive. *(Appendix 4.5)*



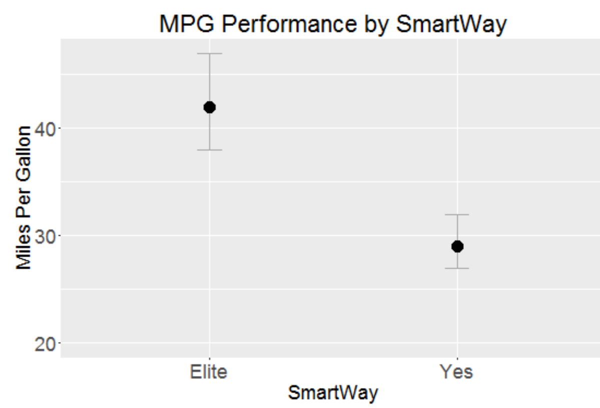
Miles per Gallon by Fuel Type

Although only the fuel efficiency during the gasoline operation of an engine was considered in this study, vehicles that are able to use both gasoline and electricity generally have higher MPG ratings (when using just gasoline) than cars with other fuel types. Diesel vehicles have higher miles per gallon ratings than gasoline vehicles, while vehicles that use a gasoline and ethanol mix have the lowest miles per gallon ratings. *(Appendix 4.6)*



Miles per Gallon and Smart Way

SmartWay Elite certification not only guarantees lower emissions but also signals a significant improvement in fuel efficiency. Elite certificated vehicles on average get 12.5 more miles per gallon than those that are not certified. *(Appendix 4.7)*



Predictive Statistics

(Append 5.1 - 5.5) The data was used to estimate the combined MPG for any vehicle given its attributes. A linear regression model was fit to the data. Predictors that were significant in the model include type of fuel used, engine displacement, vehicle class, year of the car, type of drivetrains, type of transmission, number of transmissions, and the Smart Way certification. The model explains about 60% of the variation in MPG. Considering the complexity behind fuel efficiency, this model is able to provide reasonable estimates. Interpreting the coefficients of the model reveals the following information:

- Vehicles manufactured in recent years have higher MPG ratings
- Larger engine displacement results in a lower MPG
- Diesel vehicles get more miles per gallon than gasoline-powered vehicles
- SUVs and station wagons on average get fewer miles per gallon than smaller vehicles
- Four-wheel drive vehicles tend to have lower MPGs than two-wheel drive vehicles
- Every vehicle in the data has a Smart Way designation, but some are marked as Smart Way Elite, the highest industry benchmark. On average, Smart Way Elite vehicles get more miles per gallon holding all other factors the same. Consumers concerned with fuel efficiency should look for Smart Way Elite certification.

In addition, some findings were counterintuitive:

- It is generally believed that manual transmission is more fuel efficient than automatic transmissions, but our analysis showed no significant advantage. The most efficient transmission type is continuously variable transmission (CVT) which are often used in hybrid vehicles.
- An increase in the number of transmissions leads to a lower MPG

Conclusions

In this exercise, we studied factors affecting fuel economy by analyzing different vehicle features. From the regression analysis and model, we can conclude that the fuel type, engine displacement, vehicle class, year of the car, type of drivetrain, type of transmission, number of transmissions, and SmartWay certification, are statistically significant. Our results confirm trends in fuel economy observed in the U.S. automobile industry. However, there are additional factors that were not captured in our analysis such as weight, design specification, aerodynamic efficiency, etc. Some of the assumptions behind the statistical techniques we employed may not be valid and can be looked at in more detail in further studies. Automakers and consumers can use this model to benchmark car models based on specific features and to determine the fuel efficiency of any vehicle of interest.

Appendix

Code File - http://rpubs.com/aayushmnit/stats_final

1.1 Data Processing

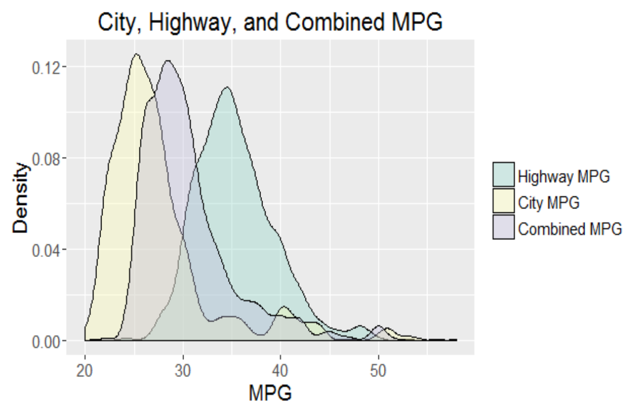
- Collected data for different years from 2011 - 2016, from the www.fueleconomy.gov website and collated them using R code
- Removed some of the unnecessary columns such as Sales Area, Standards, Standard Description, Underhood ID's and source file names
- Separated transmission type and number of transmissions from the trans column, giving CVT transmission a 1 in transmission type variable based on business understanding
- Removed hydrogen and electricity operated cars as we are interested in gasoline based cars (including hybrids) or vehicles with an MPG rating
- Took only MPG of gasoline engines from the hybrid fuel cars, for making "Apple to Apple comparisons" and keeping the response variable in Miles per Gallon
- For duplicate entries we took a mean of numerical variables to gauge the average values of our variable of interest (Combined MPG)

1.2 Variable Dictionary

Variable Name	**Description**	**Type of Data**
Model	Vehicle make and model (e.g., Honda Civic).	Nominal
Displ	Engine displacement. Indicates engine size and power.	Interval
Cyl	Number of cylinders. Indicates power of the engine.	Interval
Drive	2-wheel drive or 4-wheel drive	Nominal
Veh.Class	Size and type of vehicle (e.g., midsize car)	Nominal
SmartWay	Certification from the EPA for low emissions	Nominal
year	The year in which the car was tested (e.g., 2015)	Interval
Transmission_type	Type of transmission (e.g., manual or automatic)	Nominal

Transmission_number	Number of transmissions or gears in the engine.	Interval
Fuel	Type of fuel the vehicle uses (e.g., gasoline)	Nominal
Hwy.MPG.mean	Highway MPG. The fuel efficiency of a vehicle (in miles per gallon) while driving on a highway	Interval
City.MPG.mean	City MPG. The fuel efficiency of a vehicle (in miles per gallon) while driving on in city conditions	Interval
Cmb.MPG.mean	Combined MPG. The overall fuel efficiency of a vehicle (in miles per gallon)	Interval

2.1 Distribution of Miles per Gallon Ratings



The combined MPG is calculated from highway MPG and city MPG. The mean and median of combined miles per gallon in the sample is 30.5 and 29.0, suggesting that there are some extreme values skewing the mean higher. As we can see from the distribution of the three variables, the distribution of highway MPG resembles a bell curve, whereas the distribution of city MPG is significantly right-skewed, which explained right-skewed shape of the combined MPG distribution. The reason behind is the relative stable speed on highway versus the unpredicted speed and conditions in city that largely drag down the fuel efficiency.

2.2 Testing of relationship between MPG and Highway MPG

Significance test by simple linear regression

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-4.4071	0.5663	-7.782	1.27E-14	***
data1$Hwy.MPG.mean	0.9867	0.0159	62.044	<2e-16	***

2.3 Testing of relationship between MPG and City MPG

Significance test by simple linear regression

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8.158635	0.135401	60.26	<2e-16	***
data1$City.MPG.mean	0.808725	0.004798	168.56	<2e-16	***

3.1 Residual table by year and vehicle type

Year	Large car	Midsize car	Small car	Station wagon	SUV
2011	-0.23	-1.59	-0.11	2.10	1.11
2012	-0.29	-1.67	1.42	1.09	-1.34
2013	-0.39	0.70	0.32	-0.24	-1.51
2014	0.05	0.79	0.15	-2.09	0.26
2015	-0.23	0.41	-0.27	-0.57	0.61
2016	1.10	0.90	-1.50	0.42	0.96

4.1 Testing of relationship between MPG and year

ANOVA test

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
data1$year	5	1367	273.39	12.18	1.30E-11	***
Residuals	1608	36080	22.44			

Tukey HSD test at 95% confidence interval

Tukey multiple comparisons of means 95% family-wise confidence level				
year	**diff**	**lwr**	**upr**	**p adj**
2012-2011	0.4812012	-0.8273133	1.7897160	0.9010468
2013-2011	1.7511545	0.4932161	3.0090930	0.0010489
2014-2011	1.6161990	0.4033719	2.8290260	0.0020604
2015-2011	2.1831840	0.9420622	3.4243060	0.0000086
2016-2011	2.9668898	1.6715033	4.2622760	0.0000000
2013-2012	1.2699534	0.0818518	2.4580550	0.0281268
2014-2012	1.1349979	-0.0052327	2.2752280	0.0518667
2015-2012	1.7019829	0.5317010	2.8722650	0.0005012
2016-2012	2.4856886	1.2580069	3.7133700	0.0000001
2014-2013	-0.1349555	-1.2167711	0.9468600	0.9992527
2015-2013	0.4320295	-0.6814150	1.5454740	0.8786064
2016-2013	1.2157352	0.0421078	2.3893630	0.0372718
2015-2014	0.5669850	-0.4952292	1.6291990	0.6494658
2016-2014	1.3506907	0.2255499	2.4758320	0.0082625
2016-2015	0.7837057	-0.3718788	1.9392900	0.3810355

4.2 Testing of relationship between MPG and vehicle class

ANOVA test

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
data1$Veh.Class	4	2744	686	31.8	<2e-16	***
Residuals	1609	34703	21.6			

Tukey HSD test at 95% confidence interval

Tukey multiple comparisons of means 95% family-wise confidence level				
Class	**diff**	**lwr**	**upr**	**p adj**
midsize car-large car	3.4821018	1.5180929	5.4461107	0.0000139
small car-large car	1.8872332	-0.0336470	3.8081133	0.0569143
station wagon-large car	0.8145033	-1.3466899	2.9756965	0.8418909
SUV-large car	-1.0244565	-3.1586763	1.1097632	0.6845017
small car-midsize car	-1.5948686	-2.3390051	-0.8507322	0.0000001
station wagon-midsize car	-2.6675984	-3.9064344	-1.4287625	0.0000000
SUV-midsize car	-4.5065583	-5.6977142	-3.3154024	0.0000000
station wagon-small car	-1.0727298	-2.2419878	0.0965281	0.0898837
SUV-small car	-2.9116897	-4.0303058	-1.7930736	0.0000000
SUV-station wagon	-1.8389599	-3.3330407	-0.3448790	0.0070968

4.3 Testing of relationship between MPG and transmission type

ANOVA test

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
data1$Transmission_type	7	12467	1781	114.5	<2e-16	***
Residuals	1606	24980	15.6			

Tukey HSD test at 95% confidence interval

Tukey multiple comparisons of means
95% family-wise confidence level

Transmission Type	diff	lwr	upr	p adj
Auto-AMS	-3.4230609	-5.1933504	-1.6527714	0.0000001
AutoMan-AMS	0.5178258	-1.5008123	2.5364640	0.9941936
CVT-AMS	5.0005305	3.2396239	6.7614371	0.0000000
Man-AMS	-2.6120029	-4.2579380	-0.9660678	0.0000437
Other-AMS	-2.1877445	-5.8549583	1.4794694	0.6129687
SCV-AMS	-0.8950094	-2.8327161	1.0426972	0.8566773
SemiAuto-AMS	-3.4215468	-5.0854238	-1.7576697	0.0000000
AutoMan-Auto	3.9408867	2.4071210	5.4746524	0.0000000
CVT-Auto	8.4235914	7.2495686	9.5976141	0.0000000
Man-Auto	0.8110580	-0.1822395	1.8043555	0.2055166
Other-Auto	1.2353164	-2.1890076	4.6596405	0.9579966
SCV-Auto	2.5280515	1.1024833	3.9536196	0.0000023
SemiAuto-Auto	0.0015141	-1.0212392	1.0242674	1.0000000
CVT-AutoMan	4.4827047	2.9597784	6.0056309	0.0000000
Man-AutoMan	-3.1298287	-4.5182144	-1.7414430	0.0000000
Other-AutoMan	-2.7055703	-6.2646390	0.8534984	0.2901802
SCV-AutoMan	-1.4128352	-3.1371378	0.3114673	0.2017475
SemiAuto-AutoMan	-3.9393726	-5.3489822	-2.5297630	0.0000000
Man-CVT	-7.6125334	-8.5890102	-6.6360566	0.0000000
Other-CVT	-7.1882749	-10.6077577	-3.7687922	0.0000000
SCV-CVT	-5.8955399	-7.3094393	-4.4816405	0.0000000
SemiAuto-CVT	-8.4220773	-9.4285023	-7.4156522	0.0000000
Other-Man	0.4242584	-2.9374633	3.7859801	0.9999438
SCV-Man	1.7169935	0.4491520	2.9848349	0.0010799
SemiAuto-Man	-0.8095439	-1.5977075	-0.0213803	0.0391550
SCV-Other	1.2927350	-2.2210631	4.8065332	0.9533389
SemiAuto-Other	-1.2338023	-4.6043449	2.1367403	0.9545752
SemiAuto-SCV	-2.5265373	-3.8175859	-1.2354888	0.0000001

4.4 Testing of relationship between MPG and number of transmission

ANOVA test

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
data1$Transmission_number	6	10794	1799.1	108.5	<2e-16	***
Residuals	1607	26652	16.6			

Tukey HSD test at 95% confidence interval

Tukey multiple comparisons of means
95% family-wise confidence level

Transmission Number	diff	lwr	upr	p adj
2-1	-7.5329963	-9.5304201	-5.5355725	0.0000000
3-1	-7.3491136	-8.5023882	-6.1958389	0.0000000
4-1	-7.1539208	-8.0383277	-6.2695138	0.0000000
5-1	-5.1705520	-6.6422940	-3.6988099	0.0000000
6-1	-8.7482337	-10.2090524	-7.2874150	0.0000000
7-1	-8.6492754	-14.0839313	-3.2146194	0.0000583
3-2	0.1838827	-1.8317847	2.1995502	0.9999688
4-2	0.3790755	-1.4957707	2.2539218	0.9969139
5-2	2.3624443	0.1490990	4.5757896	0.0275845
6-2	-1.2152374	-3.4213344	0.9908596	0.6653529
7-2	-1.1162791	-6.7968035	4.5642454	0.9973663
4-3	0.1951928	-0.7296796	1.1200652	0.9960877
5-3	2.1785616	0.6821532	3.6749700	0.0003662
6-3	-1.3991202	-2.8847865	0.0865462	0.0803168
7-3	-1.3001618	-6.7415494	4.1412258	0.9923062
5-4	1.9833688	0.6828415	3.2838961	0.0001464
6-4	-1.5943129	-2.8824657	-0.3061602	0.0049648
7-5	-1.4953546	-6.8861645	3.8954553	0.9830800
6-5	-3.5776817	-5.3221549	-1.8332085	0.0000000
7-5	-3.4787234	-8.9963933	2.0389465	0.5066492
7-6	0.0989583	-5.4158080	5.6137246	1.0000000

4.5 Testing of relationship between MPG and drivetrain

ANOVA test

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)		
data1$Drive	1	1461	1461.4	65.46	1.15E-15	***
Residuals	1612	35986	22.3			

Tukey HSD test at 95% confidence interval

Tukey multiple comparisons of means				
95% family-wise confidence level				
Drive	**diff**	**lwr**	**upr**	**p adj**
4WD-2WD	-3.0302400	-3.7648470	-2.2956330	0.0000000

4.6 Testing of relationship between MPG and fuel type

ANOVA test

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
data1$Fuel	3	2984	994.8	46.47	<2e-16	***
Residuals	1610	34463	21.4			

Tukey HSD test at 95% confidence interval

Tukey multiple comparisons of means				
95% family-wise confidence level				
Fuel	**diff**	**lwr**	**upr**	**p adj**
Ethanol/Gas-Diesel	-5.8111110	-8.0721325	-3.5500900	0.0000000
Gas/Electricity-Diesel	2.2947270	0.2052979	4.3841560	0.0247320
Gasoline-Diesel	-3.5335570	-4.9704590	-2.0966560	0.0000000
Gas/Electricity-Ethanol/Ga	8.1058380	5.7509588	10.4607170	0.0000000
Gasoline-Ethanol/Gas	2.2775540	0.4763239	4.0787830	0.0064253
Gasoline-Gas/Electricity	-5.8282840	-7.4087668	-4.2478020	0.0000000

4.7 Testing of relationship between MPG and Smart Way

ANOVA test

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
data1$SmartWay	1	5619	5619	284.6	<2e-16	***
Residuals	1612	31828	20			

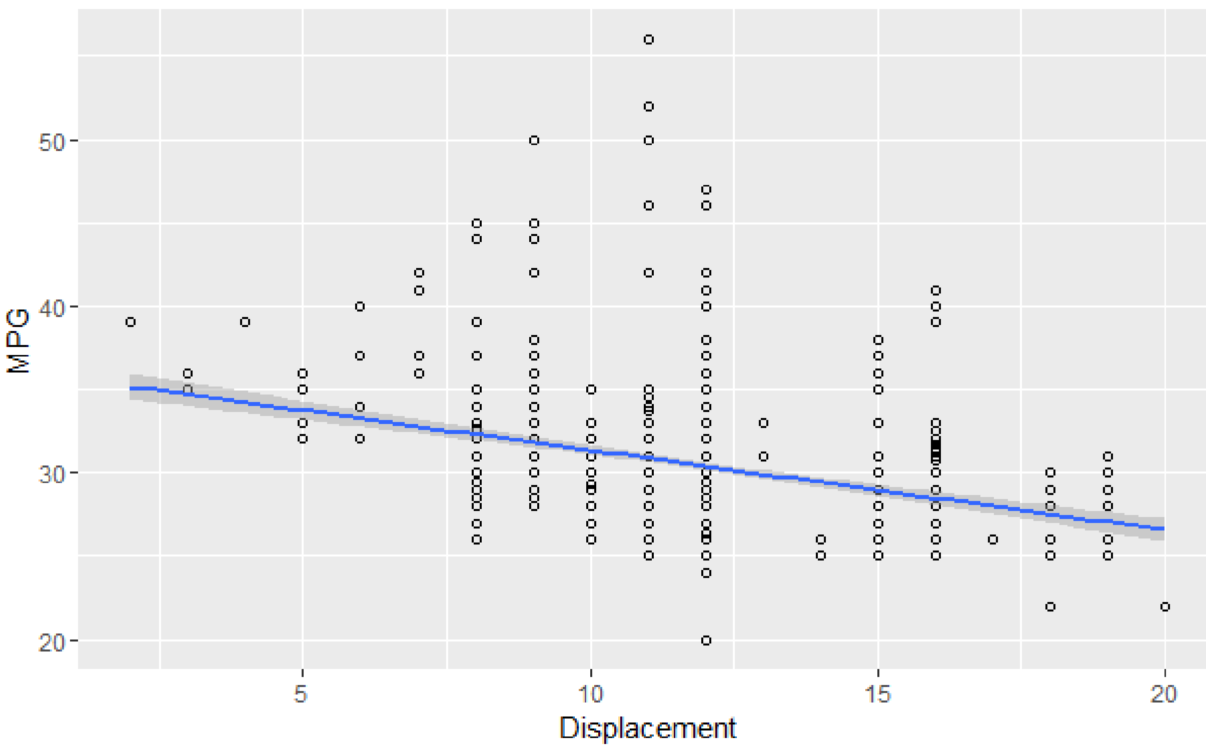
Tukey HSD test at 95% confidence interval

Tukey multiple comparisons of means				
95% family-wise confidence level				
SmartWay	**diff**	**lwr**	**upr**	**p adj**
Yes-Elite	-12.8105200	-14.2999500	-11.3210900	0.0000000

4.8 Testing of relationship between MPG and Displacement

Significant test by simple linear regression

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	36.14021	0.48662	74.27	<2e-16	***
data1$Displ	-0.47804	0.04005	-11.94	<2e-16	***

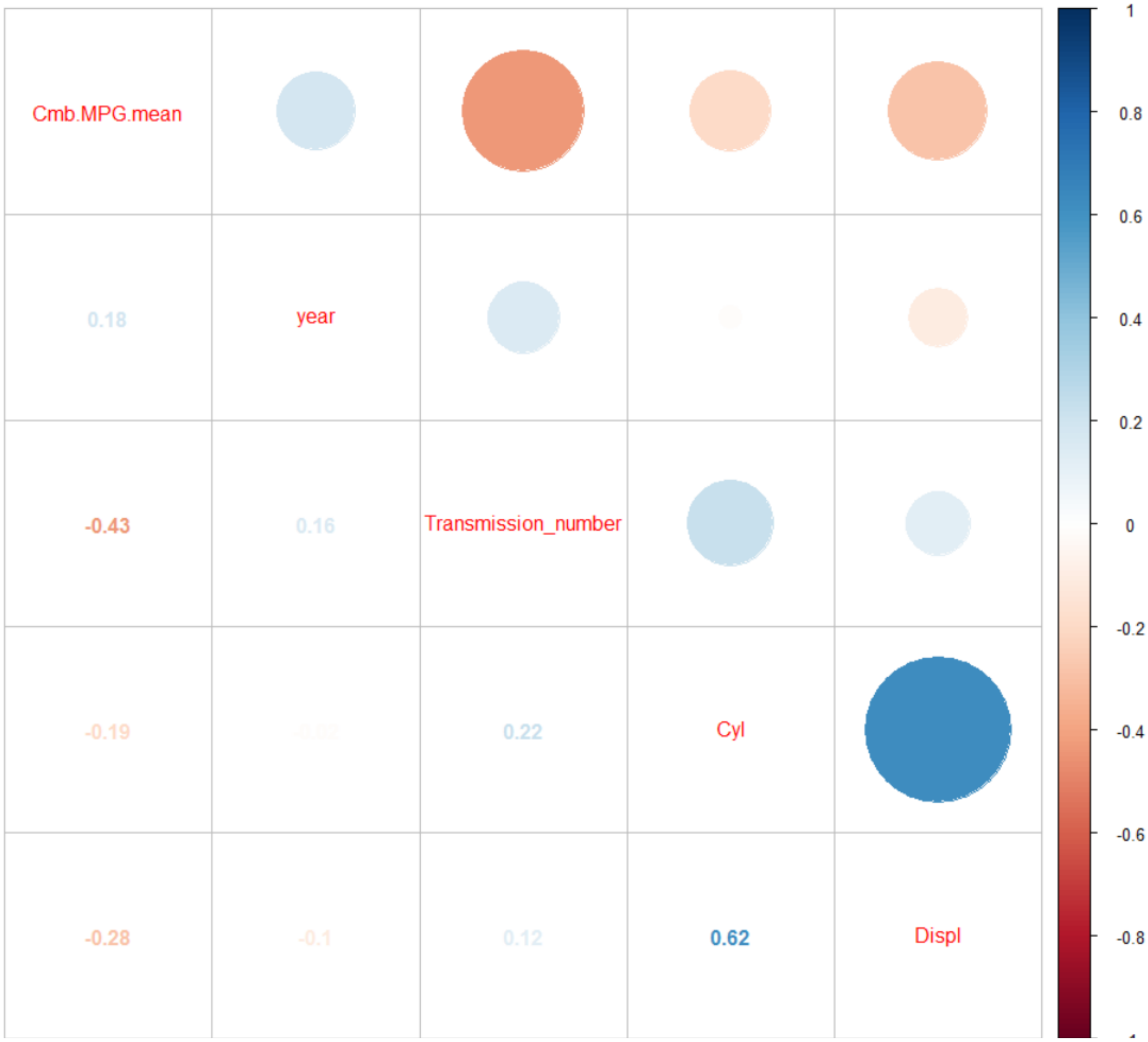


5.1 Regression Output

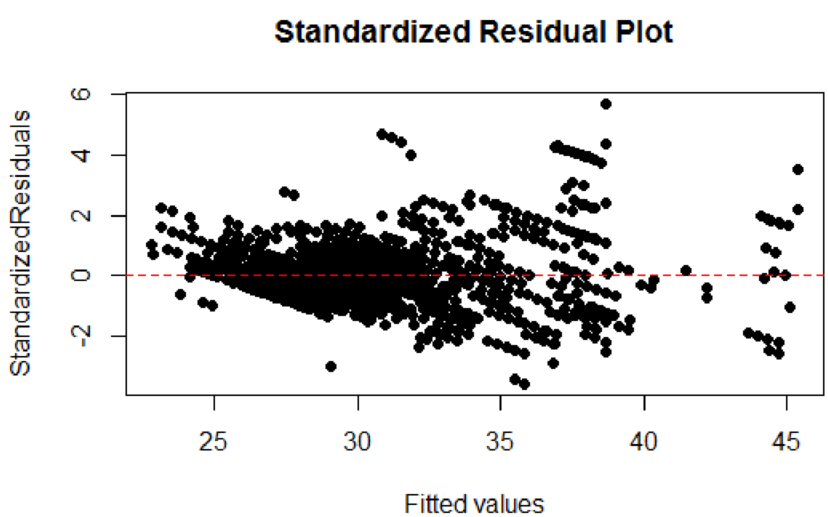
Cmbined MPG = Fuel Type + Displacement + Vehicle Class + Drivetrain + SmartWay + year + Transmission_number + Transmission_type					
Residuals:					
	Min	1Q	Median	3Q	Max
	-10.8541	-1.6747	-0.2647	1.2047	17.3489
Coefficients:					
	Estimate	Std. Error	t value	P-value	
Intercept	-652.15149	106.53766	-6.121	1.17E-09	***
Fuel type - Ethanol/Gas	-5.96909	0.61271	-9.742	< 2e-16	***
Fuel type - Gas/Electricity	-5.66837	0.59622	-9.507	< 2e-16	***
Fuel type - Gasoline	-5.40254	0.39729	-13.598	< 2e-16	***
Displacement	-0.42848	0.03152	-13.593	< 2e-16	***
Vehicle Class - midsize car	0.97853	0.19926	4.911	1.00E-06	***
Vehicle Class - large car	-0.04509	0.48447	-0.093	0.92585	
Vehicle Class - station wagon	-1.40628	0.29124	-4.829	1.51E-06	***
Vehicle Class - SUV	-2.02771	0.31392	-6.459	1.39E-10	***
Drive - 4WD	-2.33181	0.26926	-8.66	< 2e-16	***
SmartWay - Yes	-6.7349	0.58676	-11.478	< 2e-16	***
Year	0.34749	0.05294	6.564	7.06E-11	***
Transmission_number	-0.40742	0.11644	-3.499	0.00048	***
Transmission type - AMS	1.0912	0.4933	2.212	0.02711	*
Transmission type - AutoMan	3.44301	0.40802	8.438	< 2e-16	***
Transmission type - CVT	6.53533	0.44032	14.842	< 2e-16	***
Transmission type - Man	0.18928	0.26441	0.716	0.47419	
Transmission type - Other	2.67494	0.91739	2.916	0.0036	**
Transmission type - SCV	4.58299	0.40428	11.336	< 2e-16	***
Transmission type - SemiAuto	0.10898	0.2844	0.383	0.70162	
Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
Residual standard error: 3.081 on 1594 degrees of freedom					
Multiple R-squared: 0.596		Adjusted R-squared: 0.5912			
F-statistic: 123.8 on 19 and 1594 degrees of freedom, p-value: < 2.2e-16					

5.2 Correlation Test

We checked the collinearity between our interval predictors: year, number of transmissions, number of cylinders, and number of engine displacement, and we find the coefficient between number of cylinders and number of engine displacement is 0.62, which shows some evidence of collinearity. The cylinder variable has a positive coefficient, which means the MPG will increase as number of cylinders' increase, and this is counter-intuitive. We decided to remove the cylinder variable from our regression. The new model's R square and adjusted R square dropped a bit, but the coefficients are more interpretable, so we preferred the model without cylinder variable.

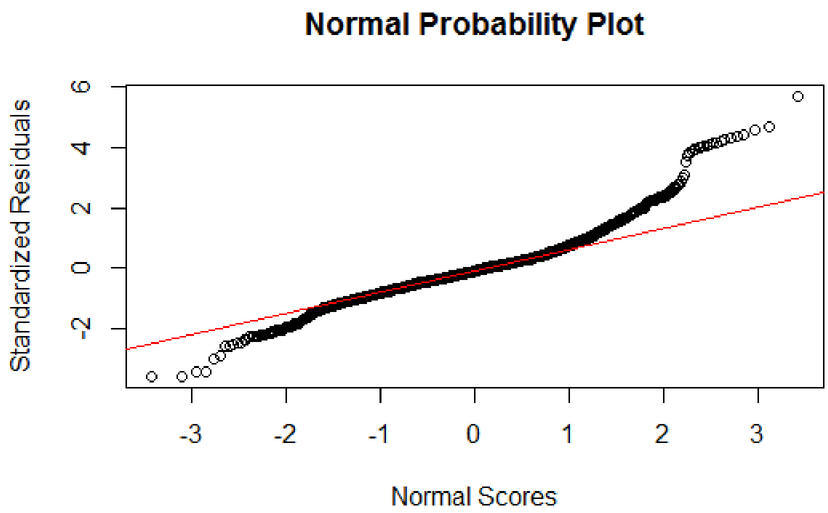


5.3 Standardized Residual



Looks like there is a standard distribution over the mean, but we can see a pattern which needs further investigation.

5.4 QQ plot



We see QQ plot getting deviated at the edges, which means violation of our assumption. We need to be aware of this by interpreting the model results.

5.5 Shapiro Wilk Test

W = 0.9329, p-value < 2.2e-16

The low p value above indicated that we are rejecting the null hypothesis that the distribution of residuals is normal and it's a violation of our assumption. Further testing is needed to check these assumptions.