# DATA ANALYSIS REPORT

**The analysis of the drivers of fuel economy in a large range of vehicles for private consumer, organisational and government use in the US**

SID: 500027851, 510522423, 480431309

The University of Sydney

QBUS2810 - Statistical Modelling for Business

4th November 2022

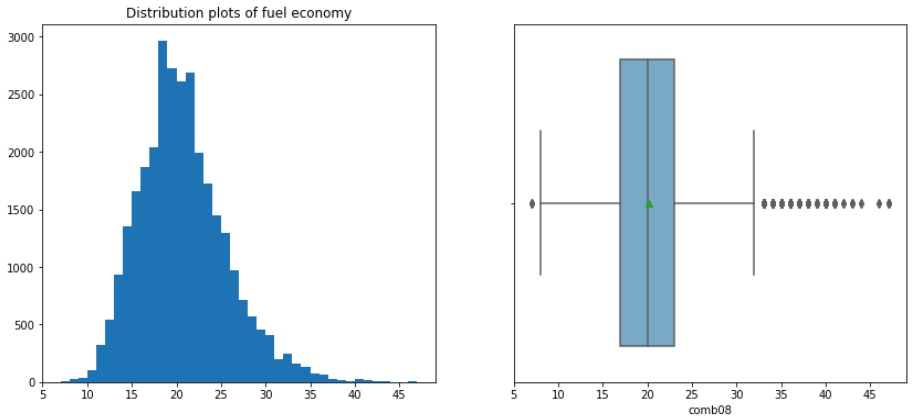
## Exploratory analysis

In this section, we do exploratory data analysis on the key variables fuel economy (comb08) and engine displacement (displ) as well as some other variables of interest.

| Measures | comb08 | displ |
|---|---|---|
| Count | 30431.00 | 30431.00 |
| Mean | 20.222 | 3.257 |
| Standard Deviation | 4.906 | 1.343 |
| Minimum | 7.00 | 0.9 |
| 1st Quantile | 17.00 | 2.2 |
| 2nd Quantile (Median) | 20.00 | 3 |
| 3rd Quantile | 23.00 | 4.2 |
| Maximum | 47.00 | 8.4 |
| Skewness | 0.667 | 0.703 |
| Excess Kurtosis | 0.893 | -0.381 |

**EDA on Fuel Economy**

In single-fuel, petrol or diesel, vehicles, the fuel economy (miles per gallon) averages to ~20.222 with a standard deviation of 4.906. The range of fuel economy is from 7.00 to 47.00. 75% of fuel economy was above 17.00, 50% was above 20.00 and 25% was above 23.00. The fuel economy distribution is quite normal although it is slightly right-skewed with a skewness of 0.667 and excess kurtosis of 0.893 as shown from the histogram and boxplot. There are a few outliers on the left and right tails of the distribution plots.

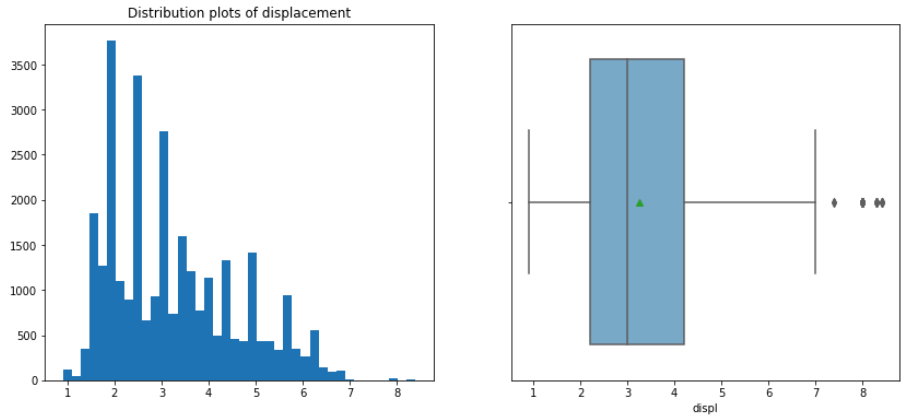



**Graph 1**: Histogram and Box plot mapped fuel economy

**EDA on Engine displacement**

The engine displacement (L) of single fuel vehicles averages ~3.257 with a standard deviation of 1.343. The range of engine displacement in this dataset is from 0.9 to 8.4. 75% of engine displacement

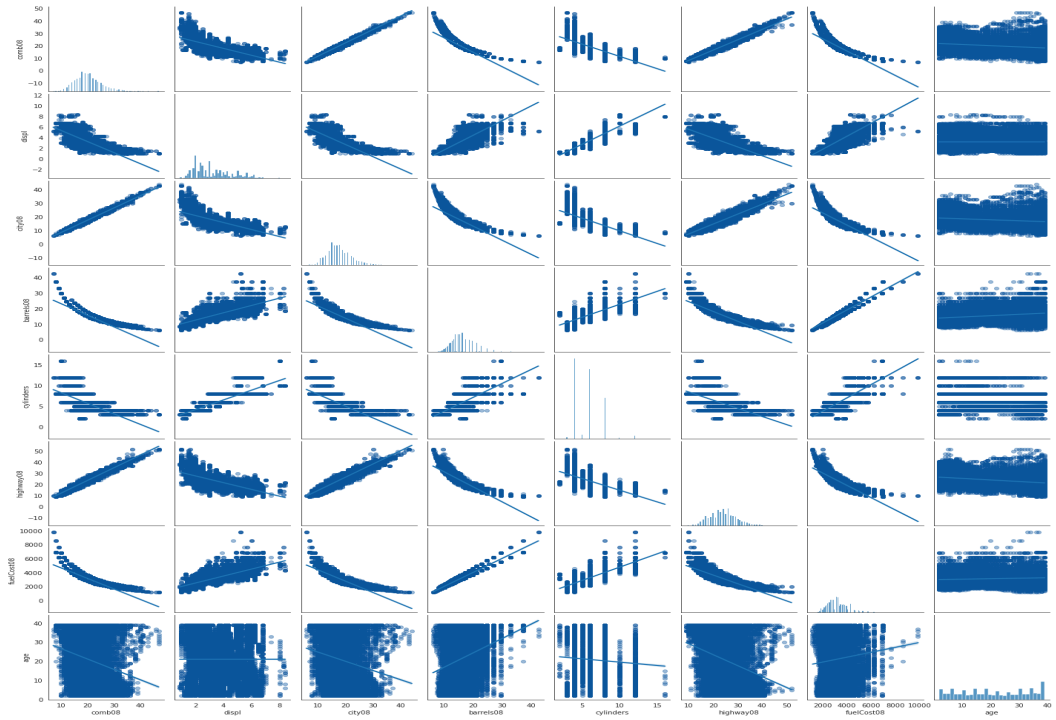
were above 2.2, 50% were above 3.0 and 25% were above 4.2. The engine displacement distribution is shown to be not normal with a skewness of 0.703 and excess kurtosis of -0.381 from the histogram and boxplot. There are a few outliers on the right tail of the distribution plots.



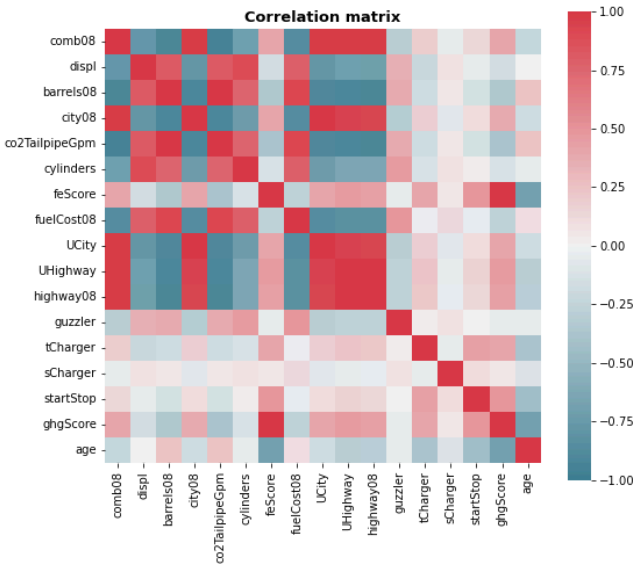

**Graph 2**: Histogram and Box plot mapped engine displacement

**EDA on other variables**

As there are many other variables that may influence fuel economy, we decided to do a grouped scatter plot and a heatmap to clearly see the correlation between the variable and fuel economy.



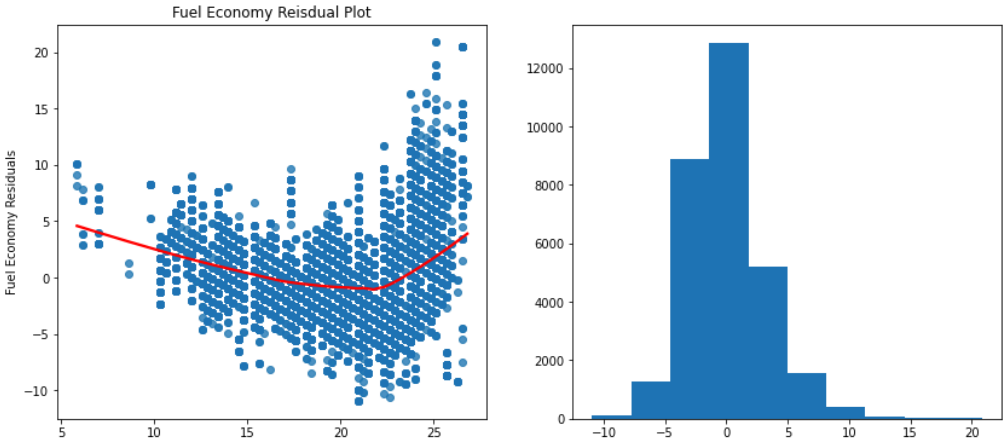

**Graph 3**: Scatterplot matrix mapped relationship between variables




**Graph 4**: Correlation heatmap

Notable variables are 'displ', 'barrel08', 'city08', 'co2TailpipeGpm', 'fuelCost08', 'UCity', 'UHighway', 'highway08' as they have a relatively high positive or negative correlation with fuel economy.

## Simple Linear Regression




**Graph 5**: Scatterplot and Histogram mapped the model's predictions

The linear model we are testing is $\widehat{Y} = \beta_0 + \beta_1 X$ where:

$Y = FuelEconomy$

$X = EngineDisplacement$

**Hypothesis:**

1. $H_0 : \beta_1 = 0$

There is no correlation between fuel economy and engine displacement meaning that fuel economy is independent of engine displacement

2. $H_1 : \beta_1 \neq 0$
Fuel economy has a dependency on engine displacement

**Assumptions of SLR:**

1. LSA 1–Linearity: From the residual plot and scatter plot from the exploratory data analysis, we can see that the fuel economy cannot be explained by a straight line using regression. Thus, this assumption is not satisfied.
2. LSA 2–Exogeneity: From the residual plot, the residuals seem to be dependent on engine displacement as its average is different from 0. From this, the assumption is not satisfied.
3. LSA 3–Independence: This dataset is randomly generated using the "grp_assnt_gendata.ipynb" file. Assuming the fact that all vehicles were sampled randomly, we can say that this assumption is satisfied.
4. LSA 4–4th moment exists: We cannot say that the extreme values or outliers are either real data or mistakes. Despite this, from the histograms and boxplots of fuel economy and engine displacement, none of the outliers seem too extreme due to normal distributions. Moreover, considering the engine displacement and fuel economy are bounded, the measures cannot be large to infinity, thus this assumption is satisfied.
5. LSA 5–Constant error variance: The errors seem to be heteroskedastic as shown by the residual plot. Therefore this assumption is not satisfied.

**Assumptions of T-test:**
The assumption that data is independent and identically distributed (i.i.d.), is satisfied following LSA 3, and the assumption that the 4th finite moment exists is unlikely to be satisfied based on LSA 5.

**Analysis of OLS Regression Results:**
We can see that the $P(t_{30431} > -206.413) = 0.000$. As the p-value is 0, we can conclude that the null hypothesis can be rejected thus the fuel economy is dependent on the engine displacement. As the engine displacement increases by 1 litre, the fuel economy decreases by $2.791$miles per gallon. Furthermore, $R^2 = 0.583$, meaning that the model can explain 58.3% of the variation in fuel economy. This number seems relatively significant. The standard error of regression is $3.12$ which is somewhat high.

We can conclude that the strength of fit is rather high as well as the fact that fuel economy is dependent on engine displacement. This dependency should be examined using other variables via MLR and their effect on fuel economy. Despite this, the test is not valid as the LSA and t-test assumptions were not satisfied enough. We can say that the fuel economy seems to be dependent on engine displacement although further analysis must be done using different models.

**<u>Standard Multiple Linear Regression</u>**
**Omitted variable bias**
In statistics, omitted variable bias (OVB) occurs when a statistical model leaves out one or more relevant variables. The bias results in the model attributing the effect of the missing variables to those that were included. In this question, we consider 2 independent variables i.e. city08, cylinders that can cause omitted variable bias in the simple linear regression since those variables have highly correlated with the engine displacement (Graph 4).

The omitted variable bias is justified based on two conditions:

1. The variable is a determinant of the fuel economy
2. There is a correlation between the variable and the main explanatory variable i.e. engine displacement.

The variable 'city08': The variable 'city08' represents driving behaviours in urban environments with low speeds in stop-and-go urban traffic, which is measured by miles per gallon (FuelEconomy. gov, n.d.).

From the above EDA of other variables, the high negative correlation of -0.78 between engine displacement (displ) and city MPG (city08) is spotted, thus the second condition is satisfied. Regarding the first condition, the question is whether the urban driving behaviours (represented by city08) are a causal factor of fuel economy. For the timing, driving behaviour needs to happen before so that the fuel consumption of the vehicle is measured; and the urban street environment and traffic are well established before how many miles a vehicle goes is determined. The variable 'city08' is argued to directly influence fuel efficiency through the speed and speed changes that drivers adopted when driving in urban streets, since speeding decreases fuel economy and vice versa (EERA, n.d.). Thus the first condition of causal effect is satisfied.

The variable 'cylinders': The variable 'cylinders' represents the number of engine cylinders.

The correlation between 'cylinders' and 'displ' is found to be equal to 0.90 which is pretty high, hence the second condition is satisfied. Clearly, the engine cylinders are installed in the vehicle far before the fuel economy is measured, thus the timing factor is satisfied. The causal relationship between 'cylinders' and 'comb08' is supported by Wang et al (2015) findings that engine capacity which is proxied by the number of cylinders importantly influences a vehicle's average rate of fuel consumption. Larger engine capacity strongly correlates with higher fuel use which is the same as lesser fuel economy. Therefore, it is possible that the number of cylinders is causing some OVB in the SLR model.

**Standard MLR model**

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | comb08 | R-squared: | 0.976 |
| Model: | OLS | Adj. R-squared: | 0.976 |
| Method: | Least Squares | F-statistic: | 4.139e+05 |
| Date: | Fri, 04 Nov 2022 | Prob (F-statistic): | 0.00 |
| Time: | 13:10:19 | Log-Likelihood: | -34776. |
| No. Observations: | 30431 | AIC: | 6.956e+04 |
| Df Residuals: | 30427 | BIC: | 6.959e+04 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 0.1998 | 0.045 | 4.446 | 0.000 | 0.112 | 0.288 |
| displ | -0.1539 | 0.008 | -18.884 | 0.000 | -0.170 | -0.138 |
| city08 | 1.0947 | 0.002 | 704.520 | 0.000 | 1.092 | 1.098 |
| cylinders | 0.1650 | 0.006 | 28.602 | 0.000 | 0.154 | 0.176 |

| | | | |
|---|---|---|---|
| Omnibus: | 270.314 | Durbin-Watson: | 1.994 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 277.361 |
| Skew: | -0.233 | Prob(JB): | 5.91e-61 |
| Kurtosis: | 3.050 | Cond. No. | 201. |

**Graph 6**: OLS Regression Results of MLR

Interpretation:
The estimated regression model is:

$$Predicted\ fuel\ economy\ =\ 0.1998\ +\ -\ 0.1539 \times displ\ +\ 1.0947 \times city08\ +\ 0.1650 \times cylinders$$

The effects of the variables on fuel economy can be shown as holding all other variables constant:

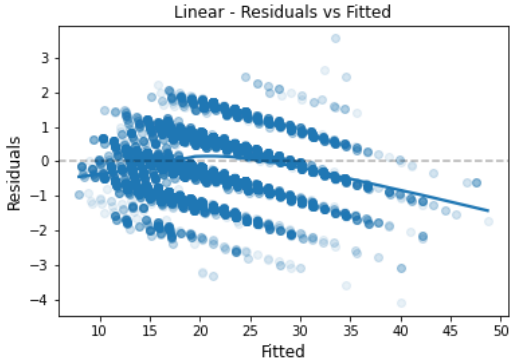
- As the engine displacement increased by 1 litre, the combined MPG decreased by 0.1539 miles per gallon on average.
- As the city MPG for fuel Type1 increased by 1 mile per gallon, the combined MPG increased by 1.0947 miles per gallon on average.
- As the engine cylinders increased by 1 cylinder, the combined MPG increased by 0.1650 miles per gallon on average.

Since there are no vehicles with *Engine displacement = Cylinders = 0* in the data, we don't interpret the intercept parameter.

Noticeably, the coefficient estimates of engine displacement appear to change from -2.7910 to -0.1539 i.e. engine displacement has a stronger negative marginal effect than its partial effect on fuel economy. Thus, 'cylinders' and 'city08' evidently have caused omitted variable bias in the SLR model.

**Assumptions of MLR LSA:**

1. LSA 1–Linearity: From the residual plot, we can see that the fuel economy can not be explained by a straight line. Thus, this assumption is not satisfied.
2. LSA 2–Exogeneity: From the residual plot, the residuals seem to be independent of engine displacement to a certain degree. However, after fitted values are larger than 30, the residuals appear to be hugely negative. From this, the assumption is not satisfied i.e. $E(e|X) \neq 0$.
3. LSA 3–Independence: This data is randomly generated. Assuming the fact that all vehicles were sampled randomly, we can say that this assumption is satisfied.
4. LSA 4–4th moment exists: We cannot say that the extreme values or outliers are either real data or mistakes. Despite this, from the histograms and boxplots of fuel economy and engine displacement, none of the outliers seem too extreme due to normal distributions. Moreover, all three measures of engine displacement, city MPG and cylinders are bounded. Thus, this assumption is satisfied.
5. LSA 5–No perfect collinearity: From the EDA section, we did not find any perfect correlation between the 3 explanatory variables. Hence this assumption holds.
6. LSA 6–Constant error variance: The errors seem to be slightly heteroskedastic as shown by the residual plot. Therefore this assumption is loosely satisfied.



**Graph 7**: Residual plot of standard MLR model

From the R^2 = 0.976, this model can explain 97.6% of the data within the training dataset. As the assumptions of MLR are somewhat unsatisfactory, we cannot say that the model fits the training data well. However, this is much stronger than the SLR model with some evidence of omitted variable bias (OVB) as the $R^2$ value jumped from 0.583 to 0.976.

**Multicollinearity**

Multicollinearity occurs when independent variables in a regression model are correlated. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results. The stronger the correlation, the more difficult it is to change one variable without changing another.

Three variables ‘'city08', 'displ', 'cylinders' have VIF of 2.55, 6.33, 5.44 consecutively. There is observed multicollinearity between 'displ' and 'cylinders' since both variables have VIF above the idle number which is 5. It is suggested that these two variables be excluded to minimise problems with multicollinearity. However, the average VIF of 3 variables is 4.77, slightly below 5. Thus, collinearity might not be a serious issue in this model.

## Variables and Model Selection

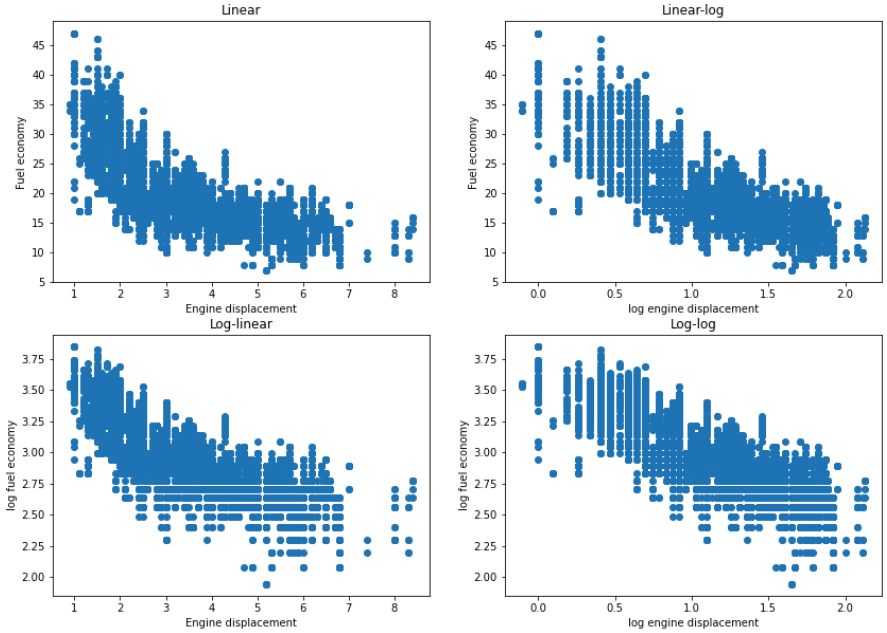
**Transforming nonlinearity and nonlinear regression**

In exploratory analysis, the scatter diagram plotting the relationship between fuel economy and engine displacement apparently displays the non-linear pattern (Graph 3). It suggests the application of techniques to account for this curvature behaviour of the dataset. Hence, the few first models are tried with the goal of capturing the nonlinear behaviour of fuel economy and engine displacement.

Assess further, preliminary examination of the data also suggested that the relationship of fuelCost08, barrels08 with comb08 seems curved and nonlinear (Graph 3). However, since the goal of analysis is interested in the effect of engine displacement, thus it is set to be the main explanatory variable in evaluating predictive models for fuel economy. Therefore, to reduce the complication, the specifications for non-linearity will mainly conducted on displ variable. Log transformation and polynomial regression are the main methods used to capture the non-linear behaviour of engine displacement.

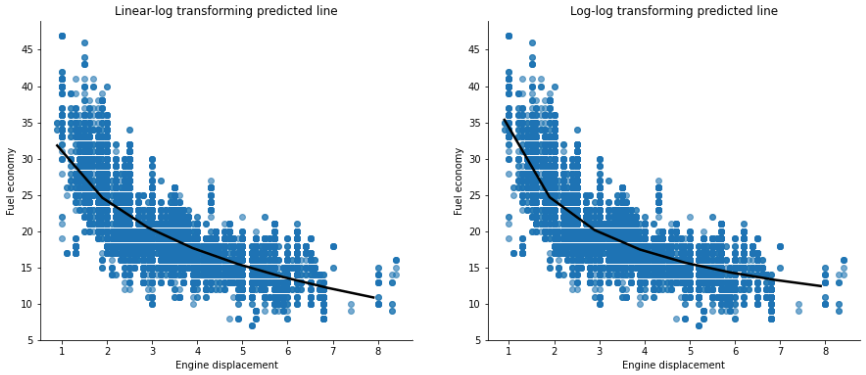
Log transforming fuel economy and engine displacement

Initially, the 2 by 2 matrix of scatterplots plotted: comb08 and displ; log_comb08 and displ; comb08 and log_displ; log_comb08 and log_displ was created to examine their suitability in mapping the non-linearity.




**Graph 8**: Scatter plot mapped the relationship between log transforming variables

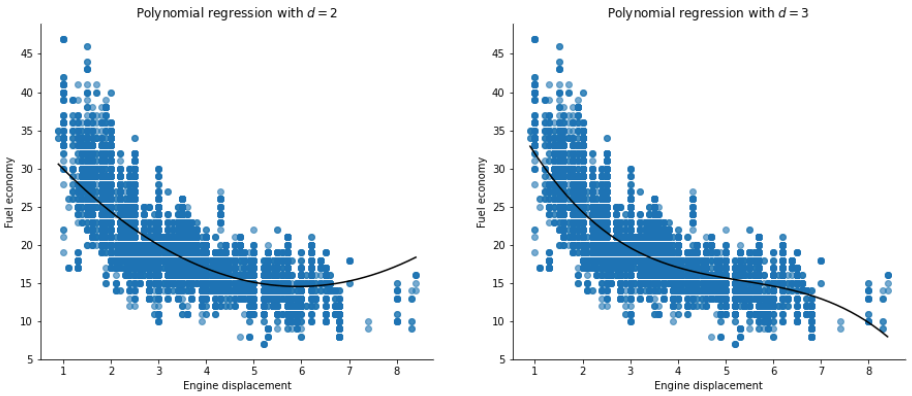
Clearly, the scatterplots of log_comb08 and log_displ with comb08 and log_displ produce a closely linear pattern, while the linear-linear and log-linear scatter plots still appear to have a curved relationship. Hence, only two models log-log and linear-log are chosen to account for non-linearity.




**Graph 9**: Log transformation

<u>Polynomial regression</u>

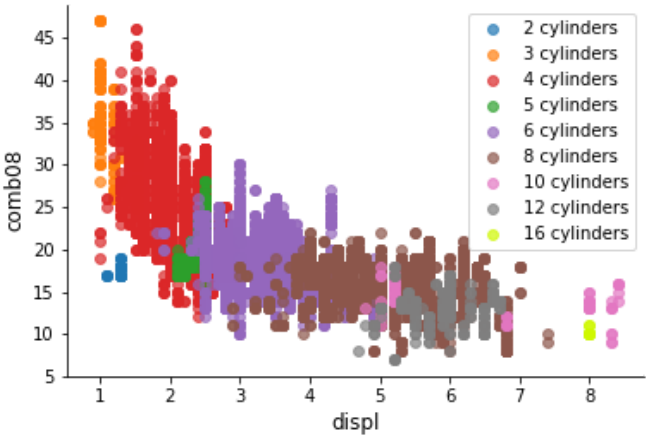
Apart from log transformation, quadratic and cubic regression are also used to clarify curvature behaviour. The polynomials of higher degrees are avoided to alleviate the problem of overfitting.




**Graph 10**: Polynomial regression

**Interaction effect**

The consideration to include engine cylinders largely formed on two justifications. Firstly, in the initial data examination, there are high correlations between engine cylinders and fuel economy, and engine displacement, thus it might have been causing some OVB in the simple model (Graph 4). Both cylinders and engine displacement seem to have large main effects on the outcome of fuel economy. Secondly, it suspects that there could be an interaction between engine displacement and cylinder, since the number of cylinders is known as the important factor in the measurement of engine displacement i.e. high-displacement engines often have higher cylinders (Leanse, 2015). It is suspected that the effect of engine displacement on fuel economy could be higher if the car engine has a smaller number of cylinders.



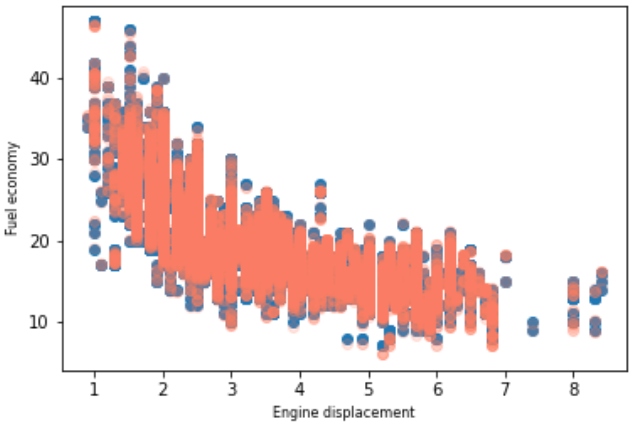

**Graph 11:** Interaction effects between displ and cylinders

**Multiple Linear Regression models**

Exploratory examination of data via correlation heatmap shows more concerns about potential omitted bias variables since there are many high correlations between predictors and responses, or between predictors themself. Moreover, several multiple linear regression models were built with the goal of understanding the importance of each of the independent variables to the relationship with a dependent variable–fuel economy, thus, to find the optimal model for predicting fuel economy.

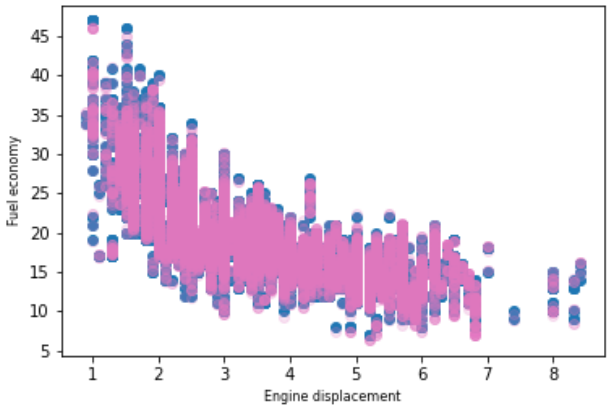
Before going deeply into choosing and reasoning input variables according to domain knowledge, we initially use the forward selection technique (Appendix 6) to choose a subset of predictors in which to construct a multiple linear regression model with the best in-sample fit–R-squared adjusted is used as the criterion. (Note that, we tried the forward selection with log_displ in replace of displacement, however, the model did not produce better results, this 'displ' is used as the favour in the selection exercise).

The full model with 17 regressors including 'UCity', 'highway08', 'barrels08', 'city08', 'co2TailpipeGpm', 'displ', 'VClass', 'tCharger', 'age', 'UHighway', 'trany', 'sCharger', 'startStop', 'feScore', 'ghgScore', 'drive', 'fuelType' is chosen by forward selection algorithm as it produces the highest R-squared adjusted.




**Graph 12**: 17 regressors model's predictions (orange dot) vs original data

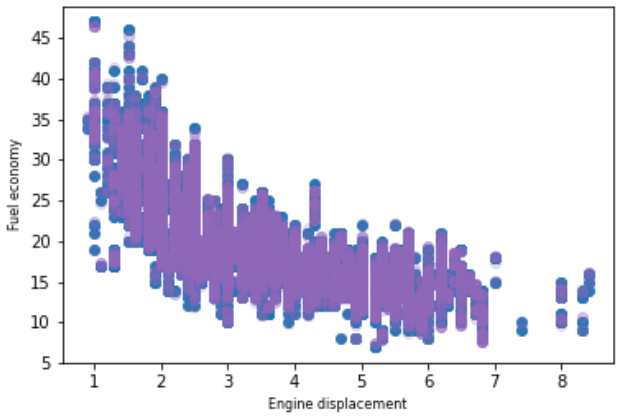
However, in the explanatory analysis, it is spotted perfect collinearity between 'UCity' and 'city08', 'highway08' and 'UHighway', 'barrels08' and 'co2TailpipeGpm', 'feScore'and 'ghgScore' (Graph 4). Thus, we run the forward selection exercise again with the elimination of variables with perfect collinearity such as 'UCity', 'UHighway', 'co2TailpipeGpm' and 'feScore'. It resulted in the model with 14 regressors as an alternative to the full model (Appendix 7).




**Graph 13**: 14 regressors model's predictions (pink dot) vs original data

In addition, we also found a simpler predictive model for fuel economy, which is motivated by the principle of parsimony. This time, to choose regressors for the MLR model predicting fuel economy,

we followed domain knowledge and previous findings. According to Wang et al. (2014) findings, two categories that have a direct influence on fuel economy are driving behaviours (e.g. speed, speed change, trip length) and vehicle fuel technology (e.g. engine capacity, vehicle age, fuel type). To proxy driver behaviours, we chose two explanatory variables 'city08' which represents urban driving with low speeds in stop-and-go urban traffic (FuelEconomy.gov, n.d.) and 'highway08' which represents rural and highway driving with free-flow traffic at highway speeds (FuelEconomy.gov, n.d.). For vehicle characteristics, number of engine cylinders and engine displacement are used as surrogates for engine capacity. Thus, the simpler model is fitted with 6 regressors including numeric variables 'displ', 'city08', 'highway08', 'cylinders', 'age' and categorical variable 'fuelType'. We also employed some model specifications that are justified above such as the log transformation of engine displacement which accounts for the non-linear patterns and interaction terms between 'cylinders' and 'displ'.



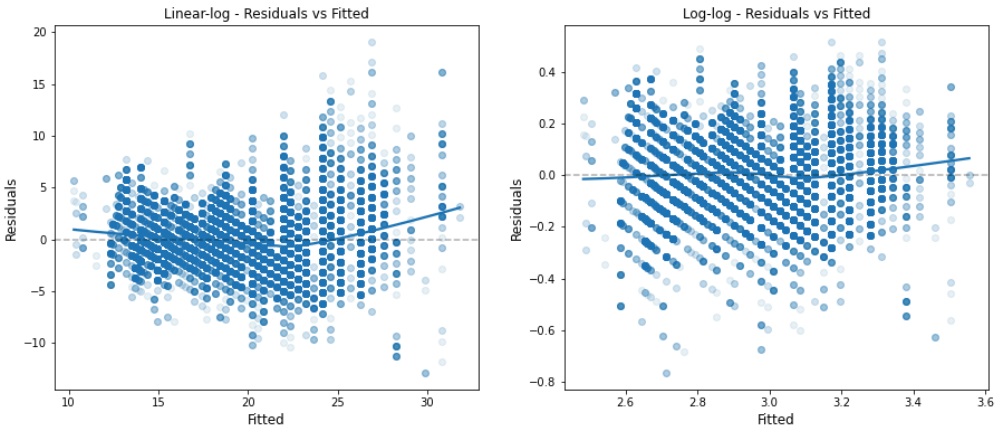

**Graph 14**: 6 regressors model's predictions (purple dot) vs original data

## A summary of the comparison of the strength of the model

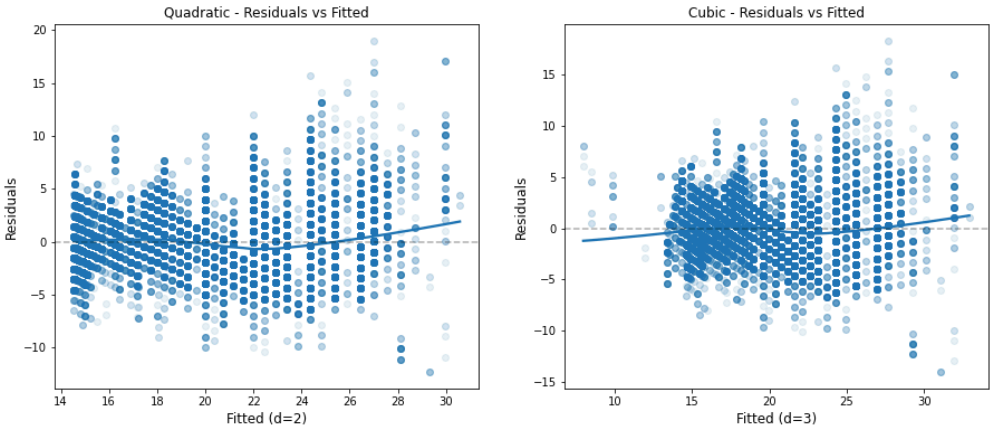
| Model | SER | R-squared | R-squared adjusted | Average VIF |
|---|---|---|---|---|
| SLR | 3.12 | 58.3% | 58.3% | - |
| Linear-Log | 2.92 | 64.6% | 64.6% | - |
| Log-Log (Duan) | 2.86 | 65.96% | 65.96% | - |
| Quadratic regression | 2.93 | 64.4% | 64.4% | - |
| Cubic regression | 2.88 | 65.4% | 65.4% | - |
| Model with interaction term | 2.998 | 62.7% | 62.7% | - |
| MLR with 17 regressors | 0.2919 | 99.6% | 99.6% | 336.87 |
| MLR with 14 regressors | 0.3541 | 99.5% | 99.5% | 5.62 |

| | | | | |
|---|---|---|---|---|
| MLR with 6 regressors (including interaction term and linear-log transformation) | 0.3606 | 99.5% | 99.5% | 6.73 |
| MLR with 6 regressors (including interaction term and log-log transformation) | 0.8561 | 97% | 97% | 6.73 |
| MLR with 2 regressors (city08 & highway08) | 0.3623 | 99.5% | 99.5% | 8.20 |

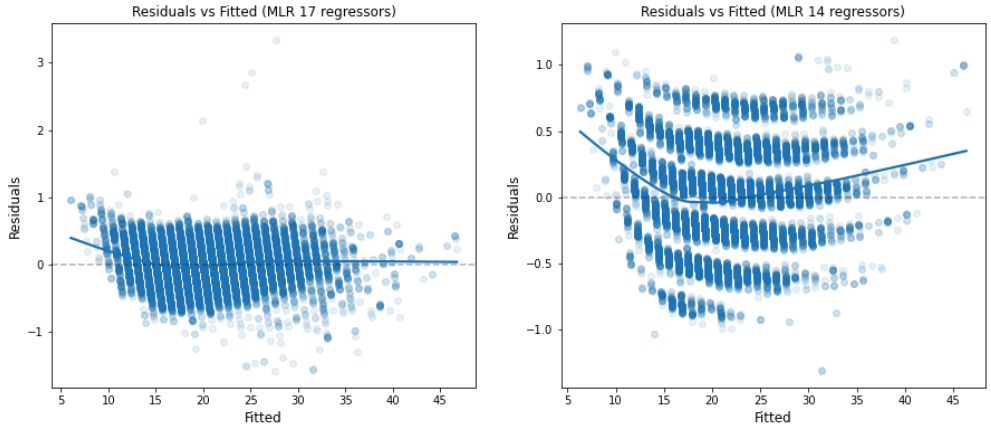
**Residual plots**



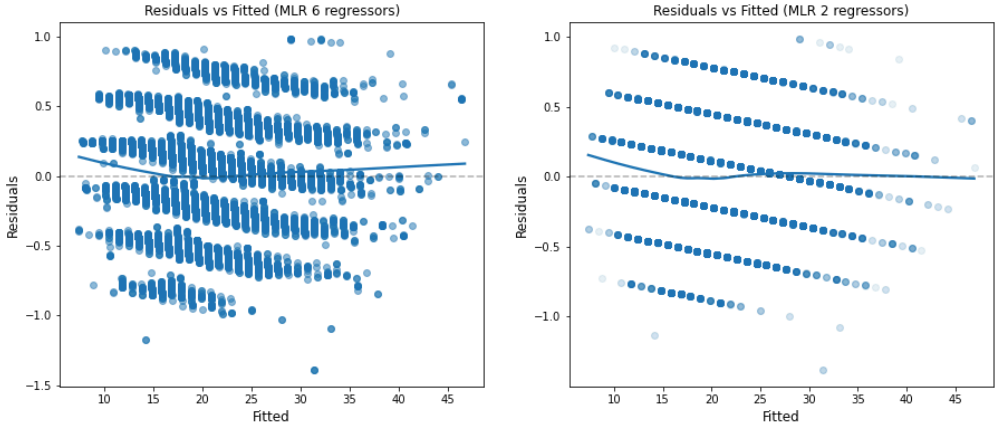

**Graph 15:** Residual plots of linear-log and log-log model




**Graph 16:** Residual plots of polynomial regressions




**Graph 17:** Residual plots of MLR with 17 regressors and 14 regressors




**Graph 18:** Residual plots of MLR with 6 regressors and 2 regressors

**Discussion**

Speaking of the strength of fit, all the models with specifications accounting for interaction and transformations/nonlinear effects for the regressors did better than standard SLR models by reducing SER at most 0.26 miles per gallon and increasing R-squared adjusted at most 7.6%. Among the models fitted to capture nonlinear patterns between engine displacement and fuel economy, clearly the log-log model is the best fit by the strength of fit measures. Assess residual plots (Graph 15 & 16), the log-log model is shown to capture the most non-linear behaviour than the other three models although there are curvature patterns when fitted values are large, but the model's goodness of fit is still acceptable. Nevertheless, when we used both logs transforming for fuel economy and engine displacement in the MLR model with 6 regressors, the model strength of fit seems to significantly reduce.

Taking all models into consideration, the full model with 17 regressors chosen by the forward selection algorithm shows the best fit. In terms of goodness of fit, the model majorly captures the non-linearity in engine displacement although log transformation of the variable is not used (Graph 17). For strength of fit, the full model has the highest R-squared adjusted of 0.996 and the lowest standard error of 0.2919. However, as said earlier, there are perfect linear relationships among some predictors in the model, thus the estimation of least-square coefficients went with large standard error and impreciseness. This problem could be seen in the estimated coefficients of engine displacement.

Although the coefficient standard error reduces from 0.014 to 0.003, the negative relationship between engine displacement and fuel economy in the SLR model is reversed to the positive relationship with beta estimated to be 0.0331.

Clearly, the 17, 14 and 6 regressors regression models are all very similar in strength of fit, the 17 model is only better than the other two models by 0.1% in R-squared adjusted and 0.06 to 0.07 in SER. The variable selection exercise is limited as it potentially generates a model with an overfitting problem. In a human sense, when we looked deeply at the forward selection exercise (Appendix 7), from the model fitted 2 regressors (i.e. 'city08' and 'highway08'), the addition of new variables only improves the R-squared adjusted by a small percentage of 0.01% to 0.07%. Thus by the principle of parsimony, the simpler model with two predictors 'city08' and 'highway08' might be good enough which produce a similar strength of fit (99.5% in R-squared adjusted and SER of 0.3623) and goodness of fit (Graph 18) to the full model.

In summary, we chose the MLR model with 6 regressors including interaction term and log transforming 'displ' variable to be the optimal model, as the simpler model still produces a similar strength of fit as the full models with 17 regressors while also might alleviate potential problems of high variance inflation factor. Moreover, the model's goodness of fit appears to be acceptable, even though some nonlinear pattern is apparent for fitted values below 15, where the residuals seem to be positive on average (Graph 18). In rank 2 is the model with 2 independent variables 'city08' and 'highway08'.

## The diagnostic analysis of the optimal model

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | comb08 | R-squared: | 0.995 |
| Model: | OLS | Adj. R-squared: | 0.995 |
| Method: | Least Squares | F-statistic: | 6.226e+05 |
| Date: | Thu, 03 Nov 2022 | Prob (F-statistic): | 0.00 |
| Time: | 20:04:11 | Log-Likelihood: | -12133. |
| No. Observations: | 30431 | AIC: | 2.429e+04 |
| Df Residuals: | 30421 | BIC: | 2.437e+04 |
| Df Model: | 9 | | |
| Covariance Type: | nonrobust | | |

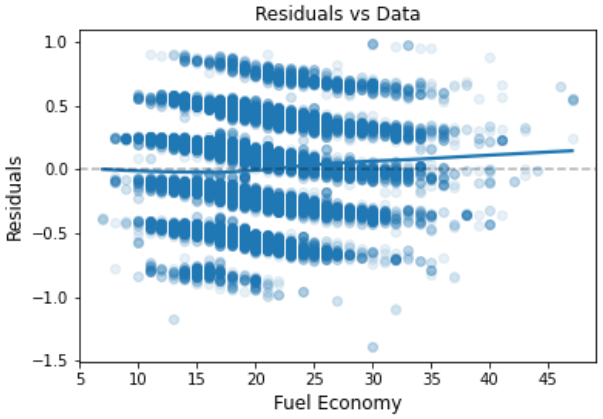
| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 0.6175 | 0.048 | 12.748 | 0.000 | 0.523 | 0.712 |
| C(fuelType)[T.Midgrade] | -0.1668 | 0.041 | -4.084 | 0.000 | -0.247 | -0.087 |
| C(fuelType)[T.Premium] | -0.0509 | 0.014 | -3.534 | 0.000 | -0.079 | -0.023 |
| C(fuelType)[T.Regular] | -0.0272 | 0.013 | -2.024 | 0.043 | -0.053 | -0.001 |
| log_displ | -0.2621 | 0.021 | -12.197 | 0.000 | -0.304 | -0.220 |
| city08 | 0.6598 | 0.002 | 377.604 | 0.000 | 0.656 | 0.663 |
| highway08 | 0.3405 | 0.001 | 301.595 | 0.000 | 0.338 | 0.343 |
| cylinders | -0.0496 | 0.007 | -7.452 | 0.000 | -0.063 | -0.037 |
| cylinders:log_displ | 0.0288 | 0.004 | 7.890 | 0.000 | 0.022 | 0.036 |
| age | -0.0020 | 0.000 | -9.336 | 0.000 | -0.002 | -0.002 |

| | | | |
|---|---|---|---|
| Omnibus: | 774.584 | Durbin-Watson: | 2.002 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 398.230 |
| Skew: | -0.056 | Prob(JB): | 3.35e-87 |
| Kurtosis: | 2.451 | Cond. No. | 1.00e+03 |

**Graph 19:** OLS Regression results of MLR with 6 regressors

From the comparison table of the strength of models, we can see that the optimal model is the MLR model with 6 regressors including the interaction term as well as the log-log transformation. This model's coefficient of determination ($R^2$) is $0.995$, its Standard Error of Regression is $0.3606$ and its average variance inflation factor is $6.73$. We can see that the optimal model can explain a much higher proportion of average fuel economy with a relatively low error rate. As the VIF is greater than 5, we can say that the model indicates high multicollinearity between the fuel economy and engine displacement as well as the other 6 regressors.

The potential nonlinear effects in the model are spotted in the plot mapped model's residuals and fuel economy, where the average residuals for fuel economy above 25 MPG are positive (Graph 20). The optimal model seems to underpredict the combined MPG when its value is larger than 25 MPG.




**Graph 20:** Scatterplot mapped residuals of the optimal model and fuel economy

The ANOVA test suggests that all variables are significant ($P < 0.05$) in determining the fuel economy except for the variable, "cylinders". This means that the model is quite significant in predicting the fuel economy using the variables within the model.

| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| C(fuelType) | 3.0 | 17107.669566 | 5702.556522 | 4.386347e+04 | 0.000000e+00 |
| log_displ | 1.0 | 486157.122632 | 486157.122632 | 3.739470e+06 | 0.000000e+00 |
| city08 | 1.0 | 212676.492824 | 212676.492824 | 1.635886e+06 | 0.000000e+00 |
| highway08 | 1.0 | 12533.594370 | 12533.594370 | 9.640711e+04 | 0.000000e+00 |
| cylinders | 1.0 | 0.087514 | 0.087514 | 6.731519e-01 | 4.119613e-01 |
| cylinders:log_displ | 1.0 | 6.857940 | 6.857940 | 5.275057e+01 | 3.877970e-13 |
| age | 1.0 | 11.331151 | 11.331151 | 8.715804e+01 | 1.067801e-20 |
| Residual | 30421.0 | 3954.941403 | 0.130007 | NaN | NaN |

**Graph 21:** ANOVA result of MLR with 6 regressors

In addition to the above, the assumptions of MLR fit much better where the regression line is much more linear and the residuals seem to be much more homoskedastic. This allows the model to have a much stronger fit compared to others.

## Result Discussion and Conclusion

Considering the goals of the study, all the models with 14, 6 or 2 predictors show the relationship between fuel economy and the potential set of useful explanatory variables and give a pretty good explanation for the variations in fuel economy since all three models have a really high R-squared adjusted of 99.5%. Moreover, in the light of previous articles studying the factors that influence real-world fuel efficiency, the predictive model of fuel economy fitted 6 independent variables including 'displ', 'cylinders', 'city08', 'highway08', 'fuelType', and 'age' becomes the appropriate casual model for fuel economy regarding the third goal of the study.

However, all of the models have variance inflation factors above idle number (5), signalling the multi-collinearity which might produce biased coefficient estimates. Thus, it might be harder to precisely understand the relationship between fuel economy and primarily engine displacement regarding the first goal of the study. For instance, in the 17 and 14 variables model, the sign of coefficient estimates of engine displacement is reversed.

The partial effect of displacement on fuel economy will be interpreted, using the optimal model with 6 predictors without the interaction term for simplicity (since the interaction term is formed by two numeric variables and the interpretation of it is outside of the subject scope). For a brief interpretation, the positive coefficient estimate (i.e. 0.0048) of the interaction term would imply that the additional numbers of engine cylinders would increase the effectiveness of 'displ' on 'comb08'. Thus the negative partial effect of engine displacements on fuel economy would be lower, holding other variables constant.

The estimated regression model without interaction terms is:

$$Predicted\ fuel\ economy = 0.3779 + -0.1358 \times displ_{log} + 0.6630 \times city08 + 0.3392 \times highway08 - 0.0015 \times cylinders - 0.0019 \times age$$ for vehicles using Diesel

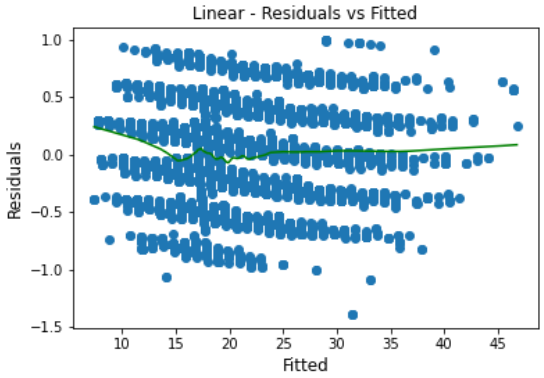
$$Predicted\ fuel\ economy = 0.2267 + -0.1358 \times displ_{log} + 0.6630 \times city08 + 0.3392 \times highway08 - 0.0015 \times cylinders - 0.0019 \times age$$ for vehicles using Midgrade

$$Predicted\ fuel\ economy = 0.3257 + -0.1358 \times displ_{log} + 0.6630 \times city08 + 0.3392 \times highway08 - 0.0015 \times cylinders - 0.0019 \times age$$ for vehicles using Premium

$$Predicted\ fuel\ economy = 0.3504 + -0.1358 \times displ_{log} + 0.6630 \times city08 + 0.3392 \times highway08 - 0.0015 \times cylinders - 0.0019 \times age$$ for vehicles using Regular

The partial effect of engine displacement on fuel economy, holding other variables constant, is estimated to be negative, which is consistent with the marginal effect in the SLR model. The partial effect of engine displacement on fuel economy can be summarised as keeping the number of engine cylinders, age of the vehicle, city MPG and highway MPG constant, one per cent increase in engine displacement is associated with an average decrease in the fuel economy of 0.001358 miles per gallon, with the 95% confidence interval being $(-0.00164, -0.00108)$ miles per gallon.

The accuracy of coefficient estimates and the confidence interval depend on MLR assumptions of the model holding. Regarding the residual plot, the mean of the residual seems to be closely equal to zero for all fitted values above 15 (Graph 22). Thus the LSA1 and LSA2: $E(e|X) = 0$ forcedly hold. Additionally, the residual plot also did not show any pattern of heteroskedasticity (Graph 22), thus LSA 6 holds. For LSA 3 and 4, all vehicles were assumed to be sampled randomly, the fuel economy and other vehicle measures are bounded, and these measures cannot be large to infinity. Hence, LSA 3 and 4 hold. Lastly, as all of the perfectly correlated variables in the dataset have been identified and removed in the above exercises, thus the last assumption (LSA 5) of no perfect collinearity is also satisfied.




**Graph 22:** Residual plot of MLR with 6 regressors without an interaction term

## Out-of-sample forecast

The three models with 2, 6 and 13 predictors (modified from 14 predictors without categorical variable 'VClass' since we encountered the error of converting data to categorical i.e. observation with value 'Special Purpose Vehicle' does not match any of the expected levels in the test dataset) are chosen to further generate forecast predictions in the test dataset.

| Model | RMSE/RMSFE | MAD/MAFE | Forecast $R^2$ |
|---|---|---|---|
| MLR with 2 regressors (city08 & highway08) | 0.356 | 0.296 | 99.5% |
| MLR with 6 regressors (including interaction term and linear-log transformation) | 0.355 | 0.294 | 99.5% |
| MLR with 13 regressors (without categorical variable VClass) | 0.349 | 0.286 | 99.5% |

**Discussion**

Regarding out-of-sample prediction, all 3 models that were built by using the training dataset do a pretty good job of predicting new data in the testing dataset. Considering forecast measures, the 13 regressors provide the lowest number of RMSE (0.349) and MAD (0.286), but these are only slightly higher than the numbers produced by the other two models. For forecast R-squared, all three 2, 6 and 13 regressors models show the results of 99.5%, which is similar to the above adjusted R-squared we got when doing model estimation in the training set. Overall, all 3 models have satisfactory out-of-sample forecast performance.

Considering all of the analysis, we still choose the model with 6 predictors including interaction terms and log transforming engine displacement as the optimal model with the goal of developing a causal model for predicting fuel economy. Regarding the residual plot, this model produces better goodness of fit than the 14 regressors model, even though it still shows some non-linearity but overall it is acceptable (Graphs 17 & 18). In terms of strength of fit, the model has a slightly higher strength of fit

(SER: 0.3606; R-squared adjusted: 99.5%) than a model with 2 regressors (SER: 0.3623; R-squared adjusted: 99.5%). The model also provides a pretty good prediction with the unseen data. Therefore, the model with 6 predictors is selected as it produces both adequate fits from the residual plots and a high strength of fit. However, there is a trade-off between getting a good fit and high multicollinearity (VIF of 6.73 is found in a diagnostic analysis of this model). Hence, the model is chosen mainly based on the quality of in-sample fit with the acknowledgement of multicollinearity as the model's limitations.

## Final report

This report aims to showcase the effect that engine displacement has on fuel economy in one-fuel-type vehicles. We have created a predictive model to understand and predict the fuel economy based on different variables.

By not using variables that are obviously linearly dependent on another variable, as well as removing variables that were unrelated to the objective variables (fuel economy and displacement), we have decreased our predictive model's complexity and increased its validity and accuracy. This also removes different sources of multicollinearity.

As expected, fuel economy is negatively correlated with engine displacement. Fuel economy describes how many miles a vehicle can go per gallon of fuel and engine displacement measures the cylinder volume (in litres) swept by all pistons of an engine. From this, we can see that the higher an engine's displacement, the more power it generates, meaning fuel usage is higher. Similarly, the lower the displacement, the less fuel it consumes.

The selection of variables relating to fuel economy on highways and cities, fuel cost, age of the vehicle and number of cylinders within the engine is very significant. These multiple regressors conclude that 99.5% of the vehicle data can be explained by these variables.

In accordance with our optimal model, we identified two main groups that can have a large influence on fuel economy which are driving behaviours (represented by 2 variables ‘city08’ and ‘highway08’) and engine technology (shown by 3 variables ‘displ’, ‘cylinders’ and ‘fuelType). Regarding driving behaviours, we recommend the Department of Energy invest more in fixing road conditions allowing for smoother driving which will in turn increase the fuel economy of the vehicles driving on it. Regarding engine technology, granting subsidies to support the movement of vehicle production toward engines with a lower number of cylinders, the total engine size, or displacement may be beneficial in terms of improving overall the fuel economy.

For future studies, we suggest the Department of Energy attempt to do a deeper analysis of factors related to engine technology that can influence fuel economy since it might be less challenging to improve vehicle technology than driver behaviours.

## Bibliography

EERE. (n.d.). T*echniques for Drivers to Conserve Fuel.* https://afdc.energy.gov/conserve/behavior_techniques.html#:~:text=Slow%20Down%20and%20Drive%20Conservatively&text=For%20light%2Dduty%20vehicles%2C%20for,by%207%25%E2%80%9314%25.

FuelEconomy.gov (n.d.). *Detailed Test Information.* https://www.fueleconomy.gov/feg/fe_test_schedules.shtml

Leanse, A. (2015). *What Is Engine Displacement?.* Your mechanic. https://www.yourmechanic.com/article/what-is-engine-displacement

Wang, H., McGlinchy, I., Badger, S., & Wheaton, S. (2015). *Real –world fuel efficiency of light vehicles in New Zealand*, Paper presented at the Australasian Transport Research Forum, 30 September – 2 October 2015, Sydney, Australia.

Wang, X., Chao, L., Kostyniuk, L., Shen, Q., & Bao, S. (2014). The influence of street environments on fuel efficiency: insights from naturalistic driving. *International Journal of Environmental Science and Technology, 11*(8), pp. 2291–2306.