The Relationship Between Engine Displacement and Fuel Economy

A report developed by students
of the University of Sydney for
The US Department of Energy

School of Business Analytics
Department of Business
The University of Sydney
Australia
May 31, 2023

Henry Ryan 500501085

Daniel Mizoguchi 520307850

Christina Wang 520430516

Lachlan Fitzpatrick 510658098

The University of Sydney have been commissioned by the US department of Energy to produce a report on the relationship between Total Engine Displacement and Fuel Economy. The data analysed is from a wide range of cars manufactured in the years 1984-2021 and is available at https://www.fueleconomy.gov/feg/ws/index.shtml. The focus is on vehicles that use either only a single fuel, being only petrol or only diesel: cars that employ electricity or gas to power them (solely or hybrid) are not considered in our analysis.

The key objectives of this study are as follows:

- Understand the relationship between fuel economy and primarily engine displacement, as well as that between fuel economy and any other useful explanatory variables.

- Develop a causal model for fuel economy, that includes engine displacement.

- Develop an optimal model for predicting fuel economy.

Contents

1 Exploratory Data Analysis **2**
- 1.1 Numerical Analysis of Primary Variables . . . 2
  - 1.1.1 Engine Displacement . . . 2
  - 1.1.2 Miles per Gallon (MPG . . . 2
- 1.2 Visual Analysis of Primary Variables . . . 3
- 1.3 Analysis of Other Variables . . . 4

2 Using SLR to Analyse Relationship between Fuel Efficiency and Engine Displacement **6**
- 2.1 Building a Simple Linear Regression Model . . . 6
- 2.2 Significance Testing of SLR Model . . . 7
- 2.3 Evaluating the Fit and Assumptions of the SLR Model . . . 7
  - 2.3.1 Fit . . . 7
  - 2.3.2 Assumptions . . . 7

3 Forming a Mutilinear Regression Model (MLR) **9**
- 3.1 Introduction to MLR and OVB . . . 9
- 3.2 Identifying Variables Relevant for Model . . . 9
  - 3.2.1 Correlation Matrix . . . 9
  - 3.2.2 Assessing OVB . . . 10
- 3.3 Formation of the Multilinear Regression Model . . . 11
- 3.4 Testing Relationship between MPG and Displacement . . . 12
- 3.5 MLR Assumptions . . . 12
- 3.6 Variance Inflation Factor and Multicollinearity . . . 14

4 Variable and Model Selection **16**
- 4.1 Variable Selection via Forward and Backward Selection . . . 16
- 4.2 Assessing Interaction Effects of Selected Variables . . . 16
- 4.3 Logarithmic Transformations of Engine Displacement and Combined MPG . . . 18
- 4.4 Spline Transformations of Engine Displacement Combined MPG . . . 20
- 4.5 Polynomial Transformations . . . 21
- 4.6 Selection of an Optimum Model from the Variable and Model Selection Exercises 22

5 Discussion of results and conclusions using training data set **25**
- 5.1 Conclusions Regarding Overall Goals . . . 25
- 5.2 Interpretation of Selected Optimal Model . . . 25

6 Generating and Assessing Forecast Predictions **27**

7 Final Conclusions and Recommendations **29**
7.1 Summary of Findings 29
7.2 Recommendations i

1

Exploratory Data Analysis

EDA performed on full data set relevant to goals of this study.

1.1 Numerical Analysis of Primary Variables

An initial exploratory analysis was performed on the engine displacement variable to assess the measures of central tendency, variance and shape. The variable analysis was conducted using the full set of data before splitting the data into 80% for training and 20% testing the predicted model.

1.1.1 Engine Displacement

The average engine displacement in litres is 3.26 with maximum value of 8.4 and minimum of 0.9. 50% of cars have a displacement greater than 3 litres and 25% greater than 4.2. From the analysis of shape measurements, the skewness and kurtosis of the data was relatively low at 0.694 and -0.399 respectively, closely resembling a normal distribution. The implications of the shape of the data set resembling a normal distribution allow us to satisfy the assumptions of the t-test and the formation of confidence intervals in Section 2 and 3. Equally as significant, the low mean, variance and kurtosis of the data set suggests that engine displacement is bound by a finite range and fourth moment. In reality, engine displacement is the measure of air volume displaced within the engine and is practically bound by a finite engine size in litres.

1.1.2 Miles per Gallon (MPG

The average miles per gallon achieved by a car is 20.15 with a maximum value of 48 miles and minimum of 7. 50% of cars can travel more than 20 miles per gallon while 25% can travel more than 23. The combined MPG variable also shows similarly low skewness and kurtosis at 0.676 and 0.943 respectively. This similarly allows us to imply that outliers are scarce in the data set and the fourth moment is bound by a finite region. From practical reason, the miles a vehicle can cover per gallon of fuel is bound by a fixed limit and justifies the performance of significance tests using the student-t distribution. From observing both of these variables, the large data count (39536) allows us to form a t-distribution that closely resembles the standard normal distribution.

	mean	median	skew	kurt
MPG (miles)	20.15	20	0.676	0.943
Displacement (litres)	3.26	3	0.694	-0.399

Table 1.1: Summary of key statistics.

1.2 Visual Analysis of Primary Variables

An initial box plot of the independent and dependent variables of study provided further insight towards the shape of the distribution.

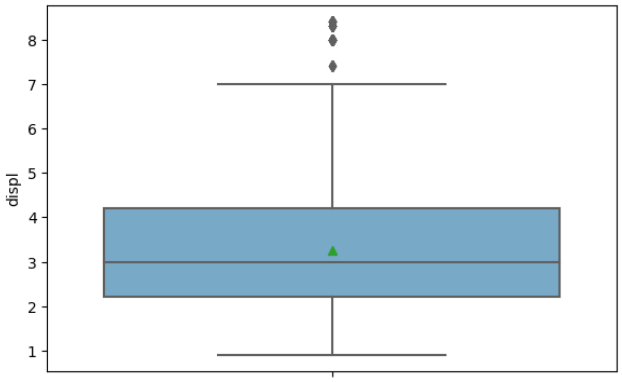

Figure 1.1: Initial Box Plot of 'displ'

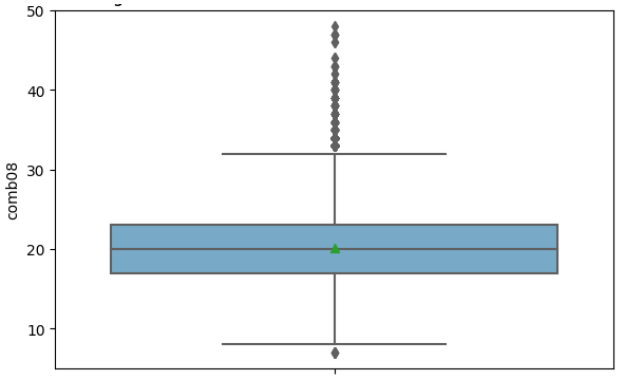

Figure 1.2: Initial Box Plot of MPG

Analysing the box plots for engine displacement and combined MPG respectively, it is apparent that the two distributions leave a tail of outliers on the right side. This was anticipated by the positive skewness of both variables but does not appear to be too significant or suggest that strong outliers are not bound by a finite range. This is similarly supported by the observation of histograms of the two variables below.

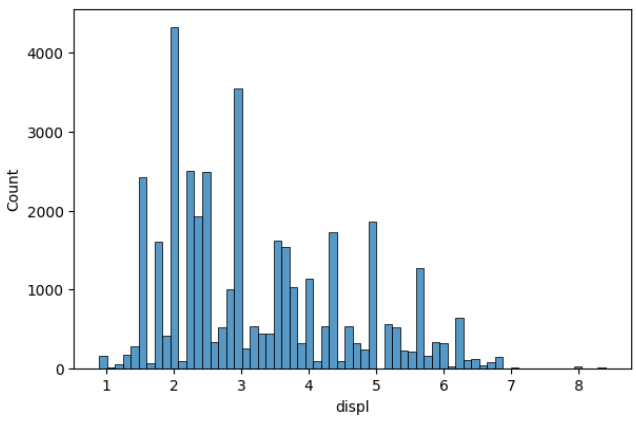

Figure 1.3: Initial Histogram of 'displ'

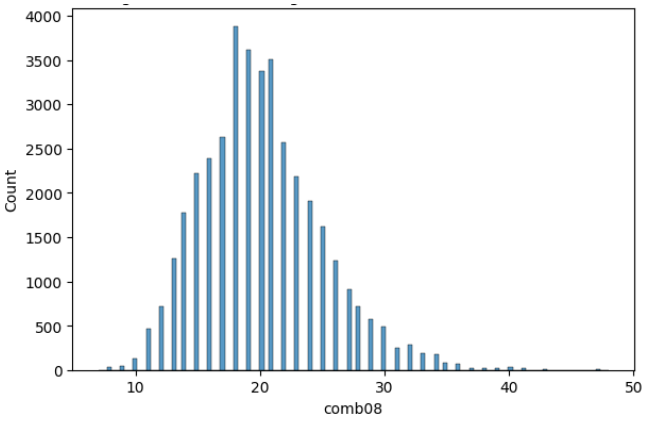

Figure 1.4: Initial Histogram of MPG

Observing the histogram of the two variables, it appears as though both variables tend towards relatively central peaks.

1.3 Analysis of Other Variables

From initial research on the effects of fuel type, turbocharging and start-stop technology, we identified categorical variables that could be determinants of combined MPG and inclusions in the variable selection process in Section 4.1. As such a box plot analysis was performed on each variable to identify any assess-able changes in fuel economy metrics. Fuel type was encoded with binary indicators for Premium, Mid-grade and Diesel fuels.

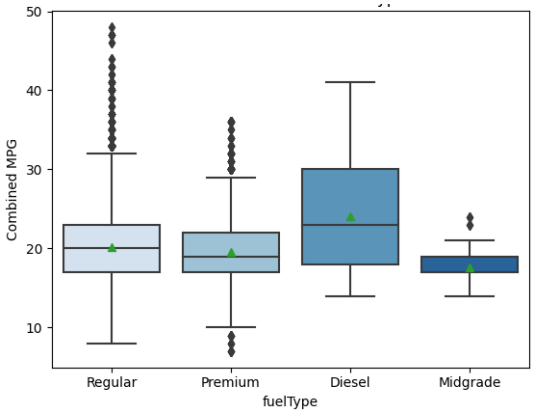

Figure 1.5: Box Plot of MPG for Different Fuel Types

An initial exploration between different fuel type effects on combined MPG showed a potential difference, namely in the effect of Diesel. Although regular, premium and mid-grade fuels show similar measures of central tendency, the differences in mean across the fuel types for a large data set motivated our decision to include the variable in the variable selection process in Section 4.1.

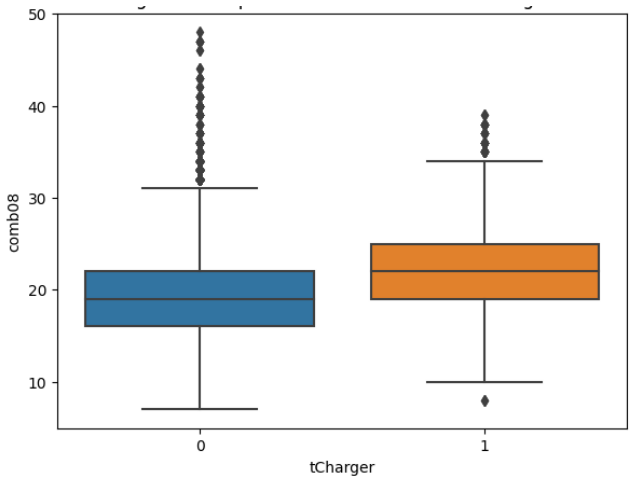

Figure 1.6: Plot of MPG for Turbocharged

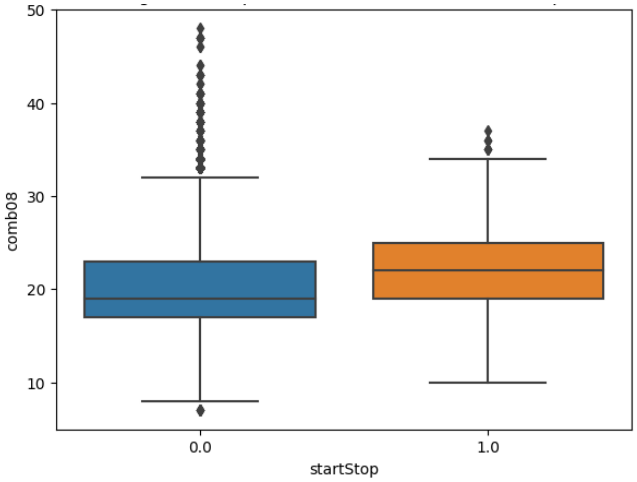

Figure 1.7: Plot of MPG for Start Stop

Initial research on turbocharging suggested that this technology forces air into engine cylinders leading to the same amount of power produced but in a smaller sized engine. Additionally, start-stop technology lowers fuel consumption while a vehicle is idle by turning off the engine. As such, these two mechanisms practically suggest an increased combined MPG. This is reflect by a positive shift in combined MPG for both vehicles with turbocharging and start-stop technology. Figure 1.6 and 1.7 hence support the inclusion of the two variables in the Section 4.1

variable selection process.

2

Using SLR to Analyse Relationship between Fuel Efficiency and Engine Displacement

2.1 Building a Simple Linear Regression Model

In practice, engine displacement is associated with lower combined MPG as the larger engine size must consume a greater volume of fuel to maintain engine operation. We would hence, expect a decreasing regression slope in a simple linear model. An initial simple linear model was conducted using the heteroskedastic-robust covariance type. This was validated by the Figure 2.1 plot of residuals against engine displacement that suggested that residuals were more significant when displacement was between 1 and 2 litres.

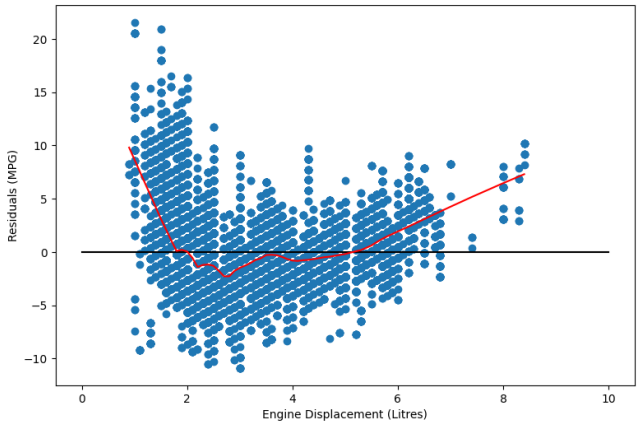

Figure 2.1: Scatter plot of residuals (MPG) against Engine Displacement

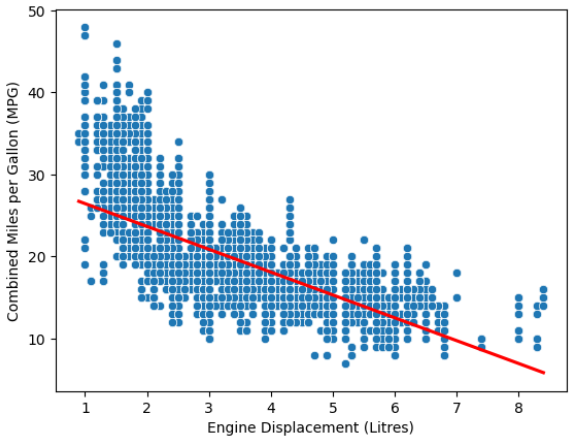

Figure 2.2: Regression plot of Combined MPG against Engine Displacement

$$\widehat{MPG} = 29.23 - 2.79(\text{Engine Displacement})$$

An initial plot of the simple linear relationship in Figure 2.2, suggested that there was a general downward trend in combined MPG with increasing engine displacement. The observed

$\hat{\beta}_1 = -2.79$, suggests that a unit increase in engine displacement in litres corresponds to a 2.79 average decrease in combined MPG assuming all else held constant. All though not practically valid, $\hat{\beta}_0$ suggests that there is a 29.23 average combined MPG for vehicles with an engine displacement of 0. In reality this is an extrapolation of data beyond the practical range of engine displacement as an engine cannot displace 0 litres.

2.2 Significance Testing of SLR Model

The hypotheses for the test of significance of the SLR model from Section 2.1 are as follows:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$
$$\alpha = 0.05$$

A two-sided test was conducted to test the significance of any linear relationship, regardless of sign. The conducted student-t hypothesis test had a test-statistic of -202.25 and a p-value of 0. This implied that the probability of achieving a test statistic equal or more significant to the test-statistic is extremely low. As the p-value is less than 0.05 (the stated alpha), we can reject the null hypothesis that the true beta 1 is equal to 0 in favour of the alternate hypothesis. We can hence further state that there is a relationship between engine displacement and combined MPG at the 0.05 significance level.

2.3 Evaluating the Fit and Assumptions of the SLR Model

2.3.1 Fit

The R^2 value of this regression is 0.585, suggesting that 58.5% of the total variance can be explained by the variance in the regression model. This suggests that there is a relatively strong fit between the independent and dependent variable. Furthermore, the relatively low standard error, 3.15, supports the strength of the model fit. Beyond this, Section 2.2 asserts that there is a significant relationship between the two variables.

2.3.2 Assumptions

In order to assess whether this fit is strictly linear, it is important to evaluate the assumptions of the SLR model. The assumptions of the model are listed as follows:

1. Linearity: $Y = \beta_0 + \beta_1 X + \epsilon$ is the true population model.
2. Exogeneity: $E(\epsilon|X) = 0$
3. Independence: Y, X are *i.i.d.*
4. 4th moment exists: $E(Y^4), E(X^4)$ are finite
5. Constant error variance: $Var(\epsilon|X) = \sigma^2$

LSA 1 assumes that a linear line makes sense with our data. Whilst it is clear that the relationship between MPG and engine displacement is always negative, the "lowless" locally smoothed regression curve fit in Figure 2.3, suggests that this relationship becomes less negative around the 3 litre mark as the curve flattens out indicating that the true population model may not be linear. This is similarly supported by the local smoothing line in Figure 2.1 that suggests residuals do not lie evenly above and below the zero axis. Hence, we can no validate LSA 1 and the linear fit model.

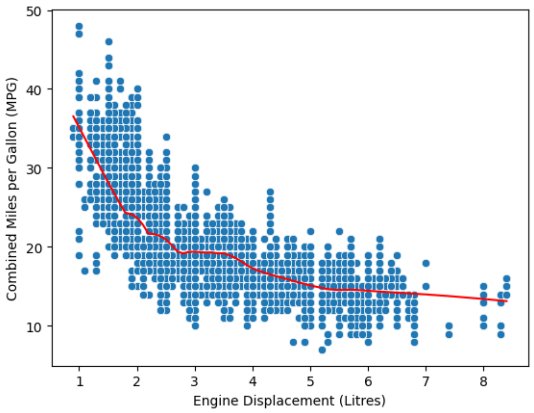

Figure 2.3: Scatter Plot of Engine Displacemnt vs MPG

As the red locally smoothed curve fit in Figure 2.1 shows, $E(\epsilon|X)$ is only approximately 0 between 2 litres and 6 litres and is positive before and after this interval. This may be caused by omitted variable bias or a non-linear fit between the variables but clearly suggests that LSA 2 is violated.

LSA 3 assumes that the data is independant and identically distributed. We cannot assume this assumption to be true as we are unaware of the sampling methodology used to create this data set and hence we must rely on satisfying the other assumptions of the SLR model.

LSA 4 assumes that both X and Y have finite fourth moments i.e outliers are rare. As highlighted in Section 1.1, both variable units (distance and volume) are bounded from below by 0 and can be argued that they are bounded from above as you cannot practically travel infinite miles per gallon, this also applies to having infinite engine displacement. Additionally, the very low kurtosis of both variables implies outliers are very rare. Thus, it is safe to assume LSA 4.

LSA 5 assumes constant error variance. It can be seen in Figure 2.1 when deciding on the covariance type for the SLR model that this does not hold. Most significantly, residuals are large when engine displacement is between 1 and 2 litres and fluctuate to the right of the regression. As such we can not validate the LSA 5.

Thus, we cannot trust the predictions of this model.

3

Forming a Mutilinear Regression Model (MLR)

3.1 Introduction to MLR and OVB

In simple linear regressions, LSA 2 explains that it is crucial for all variables to be exogenous to ensure consistency, efficiency and unbiased parameter estimates of the model. A multiple regression model describes each slope coefficient as the average difference in Y when the respective X differs by a unit, holding all other independent variables constant. The exogeneity assumption infers that the independent variables in the model, are not influenced other variables in the model, including the dependent variable Y. Violations of this assumption can lead to biased estimates and false inferences.

Omitted variable bias (OVB) occurs when an important variable is excluded from the regression model, hence, the effect of some variable X_j on Y is biased and the assumption of exogeneity fails. Consequently, the bias may cause an overestimation or underestimation of the parameters, which is reflected in the corresponding slope coefficient j.

In order to evaluate the presence of OVB in the model, we need to consider:

- Whether the omitted variable is a determinant variable and part of the true ϵ, this is analysed through data evaluation of various statistical inference processes. A determinant may be characterised by a variable that precedes the dependent which aids in the prediction of the model.
- Whether the variable is correlated with a modeled regressor, X_j.

3.2 Identifying Variables Relevant for Model

3.2.1 Correlation Matrix

A correlation matrix is created to assess the correlation between excluded variables and engine displacement.

The matrix gives us information on the correlation (positive or negative) between two variables. It is also important to identify variables that may be linear combinations of one another that

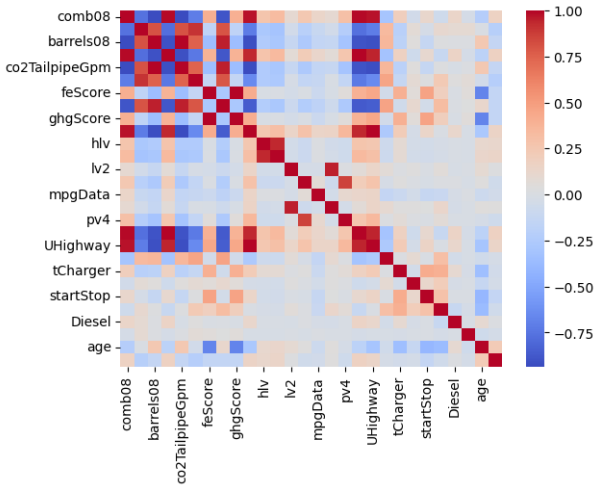

Figure 3.1: Correlation Matrix between all variables

may further limit the ability to uniquely solve the OLS model. This can be assessed by checking the definitions of variables and their respective correlations.

3.2.2 Assessing OVB

'Barrels08'

'Barrels08' is the annual petroleum consumption in barrels for fuelType1. As shown above, this is very strongly positively correlated with engine displacement at 0.81. This is expected as engines with higher displacement will consume more fuel. Additionally, it shows a stronger negative correlation with MPG than 'displ' with -0.93 and -0.73 respectively. One concern with the strong correlation between 'comb08' and 'barrels08' is the similarity in the definitions of the two. Although the correlation with engine displacement may infer OVB, the variable does not precede 'comb08' and hence is not a determinant.

'City08' and 'highway08'

These two variables are the MPG achieved in the city and on the highway. Whilst they are both highly correlated with 'comb08' and 'displ', 'comb08' represents the combined MPG meaning it is a linear combination of 'city08' and 'highway08'. Thus, including these variables would lead to perfect co-linearity and prevent us from isolating the effect of each variable X_j separately. The same applies to 'ucity' and 'uhighway' as they are just the un-adjusted values for 'city08' and 'highgway08'.

'Cylinders'

This variable shows the number of cylinders in the engine. Again this is strongly correlated with engine displacement, 0.90, as the number of cylinders in a car are a factor in determining the total displacement of the engine. The addition of this variable may be a source of multi-collinearity and hence, should be re-assessed after refitting the new model. It is also strongly negatively correlated with 'comb08' at -0.71 and it makes sense as a determinant MPG. Thus, it may be causing OVB.

'Guzzler', 'ghgScore' and 'feScore'

These categorical variables assess the fuel economy of a vehicle and determine a score or Gas Guzzler Tax. Although these variables are correlated with vehicle engine displacement, 0.34 and -0.16, they are not determinants of fuel efficiency but rather a product of fuel efficiency. As such they do not satisfy the criteria for OVB and should not be included in the model.

'co2TailpipeGpm'

Similar to 'Barrels08', this variable represents the tailpipe CO2 in grams/mile for fuelType1. It is strongly correlated with engine displacement at 0.81, thus satisfying one condition of OVB. Cars producing large amounts of CO2 per mile inherently consume more fuel per mile, hence, this is not an explicit determinant of 'comb08' but rather a joint effect of fuel inefficient vehicles. Thus this variable is does not satisfy the conditions of a variable causing OVB.

'fuelType

Fuel type is broken into three variables, 'Premium','Mid-grade' and 'Diesel' as per Section 1.3. Although the correlation between each variable and engine displacement is relatively weak (0.11, 0.09, 0.09 respectively), the coefficient of 'displ' may be jointly measuring the effect of fuel type. As different fuel types may influence the combined MPG metric differently, these variables may be causing OVB.

'tCharger'

As discussed in Section 2.3, turbo charging technology may be a positive determinant of combined MPG. Furthermore, the 'tCharger' variable is negatively correlated with 'displ' (-0.20). As such it is valid to include the categorical variable in the MLR model.

'Manual'

Similar to turbocharging, a manual transmission is negatively correlated with 'displ' (-0.21) and positively correlated with 'comb08' (0.17). Although this correlation is relatively low, it should be included in the model to assess the presence of OVB in the SLR model.

Summary

Variables to include in MLR: 'Cylinders', 'Premium', 'Midgrade', 'Diesel', 'tCharger' and 'Manual'. It is important to note that the strong correlation between 'Cyclinders' and 'displ' should be assessed for multicollinearity later.

3.3 Formation of the Multilinear Regression Model

To minimize bias on the slope coefficient, a multi-linear regression model including the above variables, together with displacement is needed. We will then receive an improved estimate of the effect of total engine displacement on MPG.

$$\begin{aligned}\widehat{MPG} =& 29.79 - 2.59(\text{Engine Displacement}) - 0.27(\text{cylinders}) + 0.61(\text{Premium}) \\ &+ 4.18(\text{Midgrade}) + 6.23(\text{Diesel}) - 0.08(\text{tCharger}) - 0.04(\text{Manual})\end{aligned}$$

	coef	$P > \|z\|$
int	29.79	0.00
displ	-2.59	0.00
cylinders	-0.27	0.00
Premium	0.61	0.00
Midgrade	4.18	0.00
Diesel	6.23	0.00
tCharger	-0.08	0.15
Manual	-0.04	0.36

Table 3.1: Initial MLR Model

This model creates an R^2_{adj} of 0.635, suggesting that 63.5% of the changes in MPG can be explained by the inputs. This is a relatively high value, and is also a notable increase from the R^2_{adj} value of 0.585 (58.5%) presented in the simple linear regression model, which only includes the displacement variable. The standard error of residuals (SER) was also calculated for the multilinear regression model, to be 2.98, which has significantly decreased from 3.15 in the simple linear model.

3.4 Testing Relationship between MPG and Displacement

Our null and alternative hypothesis are as follows:

$$H_0 : \beta_{displ} = 0$$
$$H_1 : \beta_{displ} \neq 0$$
$$\alpha = 0.05$$

As the $P(t_{3.16e+04} > |-79.193|) = 0.000 < 0.05$ we can reject the null hypothesis that $\beta_{displ} = 0$ and conclude that there is a significant relationship between the displacement of an engine in litres and the miles achieved per gallon (Assuming the following LSA's are met).

Note: β_{displ} has increased slightly from -2.786 to -2.587. This change may indicate that OVB has been removed through the addition of the variables discussed above. However the two times increase in standard error from 0.014 to 0.033 may indicate that multi-colinearity may still be present. We will hence test for the presence of multi-colinearity in Section 3.6 using the variable inflation factors of each variable to justify removing any variables from the above model.

3.5 MLR Assumptions

Linearity

When a linear regression model is suitable for a data set, the residuals should be randomly distributed around the zero line when plotted against fitted values. Figure 3.2 shows that this is not the case as a distinct pattern is formed. This indicates that a linear model might not be the best fit for our data and we should assume that assumption 1 is not met.

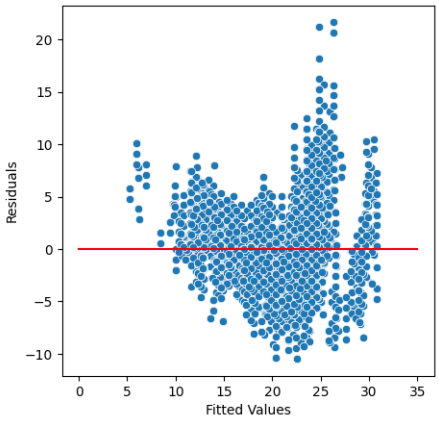

Figure 3.2: Plot of Residuals vs Fitted Values

Exogeneity

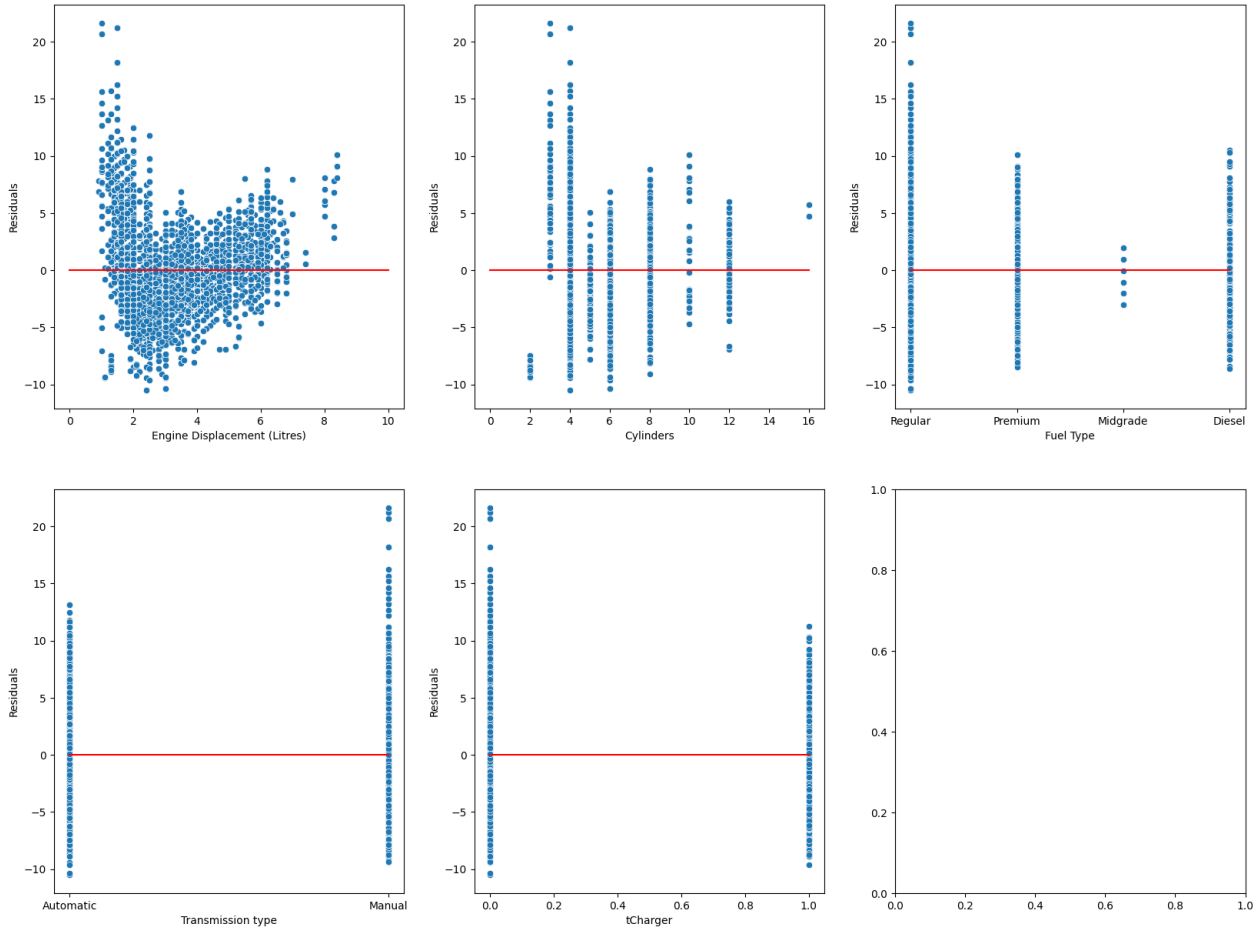

Figure 3.3: Plot of Residuals vs Independent Variables

Exogeneity is assessed using Figure 3.3 to determine if the mean error term is roughly 0 for each of the variables included in the MLR. Most significantly, 'displ' shows an obvious pattern of positive residuals when engine displacement is between 1 and 2 litres. This pattern is similarly reflected in 'cylinders', suggesting that Cylinders may be causing multicollinearity with engine displacement. In order for assumption 2 to hold, the expected value of epsilon should be 0 for each value of X. This is clearly not the case so we conclude assumption 2 is not met. It is

important to note that the remaining predictors show a relatively even distribution of positive and negative residuals suggesting that they may have a linear effect on fuel economy.

Independence

As discussed eairlier we can assume data to be i.i.d. based on the random sampling of our train data set. We require more information on the method of sampling each vehicle in the data set to validate this assumption.

Finite Fourth Moments

Assumption 4 relies on all variables having a finite fourth moment. As discussed in Section 1.1, engine displacement and combined MPG are bound by finite limits with relatively low kurtosis. Additionally, the remaining categorical variables are inherently bound by a fixed limit. Thus we can conclude that the response variable and all variables, X_j, have a finite fourth moment.

No perfect collinearity

Assumption 5 was discussed earlier in regards to selection of variables to include in MLR. The high correlation between cylinders and engine displacement, as well as their similar definitions may suggest perfect collinearity between variables. Excluding these two variables, the discussion in Section 3.3 suggests that there is relatively low collinearity between included variables.

Constant error variance

Referring back to Figure 3.3, residuals vary significantly along the X axis. One example of this is the small variance in residuals for midgrade fuel compared to other types with large residual variances. This suggests that LSA 6 is violated and further validates our choice of a heteroskedasticity robust model in Section 3.3.

Conclusion

Moreover, as LSA assumptions 1 and 2 are clearly violated, we can not justify a linear regression model using the current independent and dependent variables. This further motivated our assessment of interaction terms and transformed models in Section 4.

3.6 Variance Inflation Factor and Multicollinearity

	VIF
displ	5.93
cylinders	6.03
Premium	1.01
Midgrade	1.41
Diesel	1.08
tCharger	1.35
Manual	1.06

Table 3.2: Variance Inflation Faactors of β_i

The variance inflation factors for each of the variables selected in the MLR model above can be assessed by forming a model with each variable X_j against all other predictors. The calculated VIF for engine displacement and cylinders were both above 5 (5.93 and 6.03) respectively, and hence signalled potential multicolinearity between the two variables. Multicollinearity is problematic here as it results in a large variance in the predicted coefficients of engine displacement and cylinders. This would hence make it difficult to assess the significance of the effect of each variable on combined MPG. As such, the cylinders variable was dropped in the variable selection process in Section 4.1. All other variables appeared to have low VIF (all below 2) and hence can be considered in the model in Section 5 and 6.

4

Variable and Model Selection

4.1 Variable Selection via Forward and Backward Selection

The variable selection process involved performing forward and backward selection on all meaningful variables that provided sufficient data entries. Using a standard maximum threshold of 5% missing data from the entire dataset, the decision was made to drop the variables 'hlv', 'hpv', 'lv2', 'lv4', 'pv2' and 'pv4' from the variable selection process due to a large proportion of zero entries and inconsistency across volume metrics. As mentioned in Section 3.3, fuel consumption metrics such as 'co2TailpipeGpm', 'city08' and 'highway08' were removed as a result of their direct correlation and lack of precedence over the response variable, 'comb08', that would overshadow the effects of more meaningful variables. In order to arrive at a meaningful predictive model, the variables included in the model selection were those that strictly preceded combined MPG. The forward and backward selection was performed according to the variable influence on R^2_{adj} with 'displ' nominated for inclusion in both selections by default. Categorical variables with multiple categories were also monitored to ensure that the selection did not violate the principle of marginality by selecting a subset of categories.

The forward and backward selection models both included the same variables leaving a final model of:
$comb08 \sim displ + age + Diesel + tCharger + Manual + Premium + Midgrade + startStop + sCharger + 1$
with an R^2_{adj} of 0.704. Once more, the variable inflation factor of each selected variable was below 5 with an average of 1.26 (below 3) implying low multicollinearity between variables.

4.2 Assessing Interaction Effects of Selected Variables

Interaction effects and the inter-dependencies of variables from the selection in Section 4.1 can be assessed by performing an f-test on variables that have potential to influence one another. One interaction of interest is that between engine displacement and fuel type, inferring that the relationship between combined MPG and engine displacement may be different for Regular, Premium, Diesel and Midgrade fuels. To assess the interaction effects between the two main effects, the variables displPrem, displMid and displDies will introduced into the model to measure their overall significance.

	coef	$P > \|z\|$
int	30.19	0.00
displ	-3.19	0.00
Premium	-2.56	0.00
Diesel	6.10	0.00
Midgrade	-2.5	0.015
displPrem	0.89	0.00
displMid	1.36	0.00
displDies	0.11	0.037

Table 4.1: MLR Model Interactions

Interpreting the above table, expected change in combined MPG with a litre increase in engine displacement is the sum of -3.19 and the interaction coefficient for the corresponding fuel type, where regular fuel is the standard. Using the OLS model with the proposed new variables, an f-test was conducted to assess the significance of at least one interaction effect under the following hypotheses:

$$
\begin{aligned}
&H_0 : \beta_5 = \beta_6 = \beta_7 = 0 \\
&H_1 : \beta_5, \beta_6, \beta_7 \neq 0 \\
&\alpha = 0.05
\end{aligned}
$$

Here L should be a 3 by 8 as we wish to select 3 variables (interaction effects) from a total of 8 (including the intercept term) for overall significance.

$$
L = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}
$$

The test statistic is $F_{stat} = 334.56$ which follows an $F_{3,n-8} = F_{3,3.16\times10^{04}}$distribution under the null hypothesis. As the p-value, $P(F_{3,3.16\times10^{04}} > 334.56) = 6.88 \times 10^{-2140}$, we can reject the null and conclude that at the very least 1 interaction effect between fuel type and engine displacement on combined MPG is significantly different to 0. By the principle of marginality in accepting this alternate hypothesis and including the interaction effects, we must also include each main effect in the interaction model. The implications of this overall significance test may infer that the effect of engine displacement on combined MPG is partially effected by the fuel type associated with a given vehicle. The introduction of the interaction effect variables into the model increased the adjusted R-Squared value of the model from 0.633 to 0.645. It is important to note however that the addition of the interaction effects cause the main effects of premium and midgrade fuels to flip signs from positive to negative. Although this may be valid, the data must be doubted as premium fuels have strong carbon-reducing detergents that may positively influence fuel economy metrics. The robust coefficient for the diesel dummy variable motivated analysing a reduced model in Section 4.5 where fuel type was encoded more generally for petrol against diesel.

Another interaction effect of interest is the joint effect of a manual transmission and engine displacement on combined MPG. This relationship was of interest due to the potential link between a manual transmission and the effectiveness of pushing higher volumes of air and fuel. This interaction effect was assessed by isolating the model with the variables engine displacement, manual transmission and their interaction (displManual).

	coef	$P > \|z\|$
int	28.62	0.00
displ	-2.61	0.00
Manual	1.88	0.00
dispManual	-0.60	0.00

Table 4.2: MLR Model Interactions

The coefficient of displManual, -0.604, implies that the expected decrease in combined MPG with a litre increase in engine displacement for vehicles with manual transmission is $-2.612 - 0.604 = -3.216$. The significance of the interaction effect can be assessed by performing a one variable t-test under the following null and alternate hypotheses:

$$H_0 : \beta_3 = 0$$
$$H_1 : \beta_3 \neq 0$$
$$\alpha = 0.05$$

As above, the calculated test statistic for the significance of displManual is -14.622. Using this test statistic we calculate a p-value, $P(t_{3.16\times10^4} > |-14.622|)$, of 0.000. Using a standard 0.05 significance level, we can reject the null hypothesis in favour of the alternate that there is a non-zero coefficient for the interaction between engine displacement and manual transmission on combined MPG. Furthermore, the addition of this interaction effect increases the adjusted r-squared from 0.585 to 0.590 beyond the main effects of engine displacement and transmission type.

From the assessment of two different interaction effects including the main effect of engine displacement, we can assert that there is a non-zero coefficient for these effects at the 0.05 significance level. As such, they will be included in the model selection exercise in Section 4.5.

4.3 Logarithmic Transformations of Engine Displacement and Combined MPG

As the relationship between Engine Displacement and MPG has been shown to be non-linear in previous sections through residual plots and visual analysis, we can apply log transformations to these variables to find a relationship that satisfies the MLR assumptions.

It is clear in Figure 4.1 that the Linear-log and Log-log models fit the data most accurately, to be sure which one is better further analysis is needed.

Plotting lowless locally smoothed line through plots of the respective residuals vs fitted values in Figure 4.2 shows a pattern still present in the log-linear model and a more random distribution

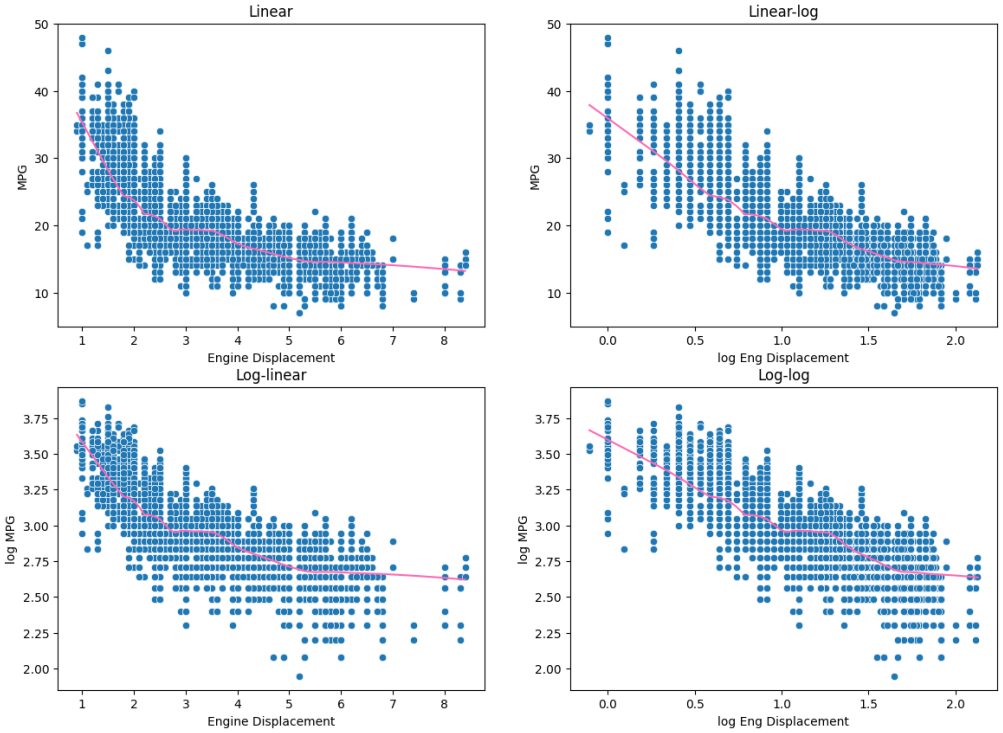

Figure 4.1: Options for Log Transformations

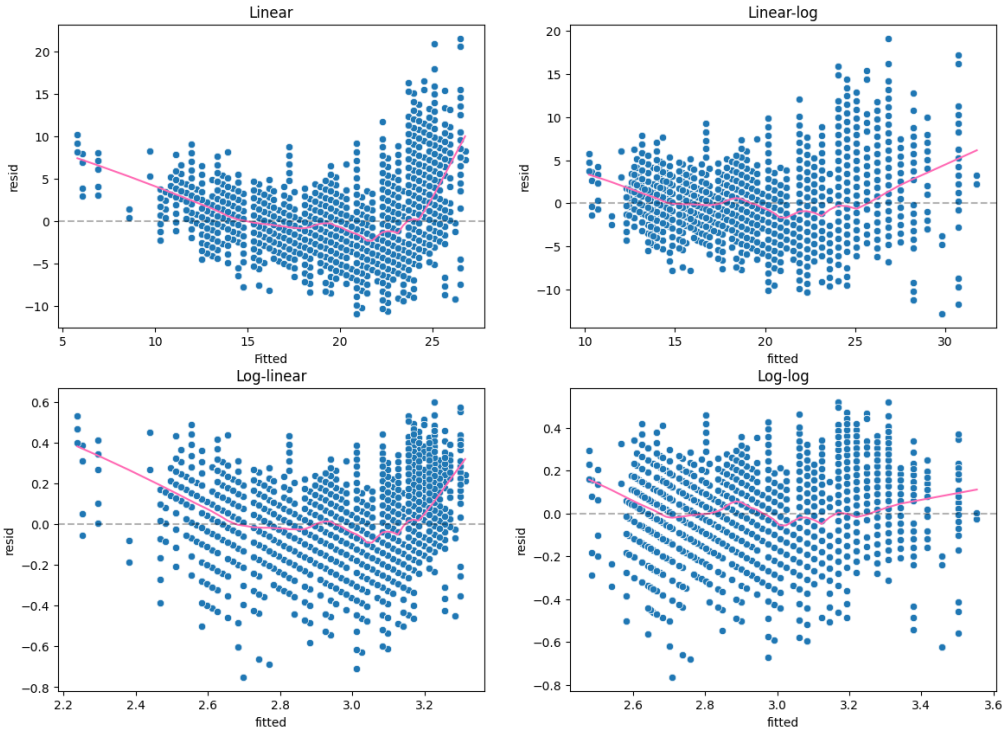

Figure 4.2: Residuals vs Fitted Values

around 0 of residuals in both the log-log and linear-log models. This indicates these models may be a better fit to the data as the first assumption of MLR can now be argued.

Table 4.3 Shows the log-log model to be a better fit to the data in all three corrections of R^2. However, as these are close it we should take into account the interpretability of each model.

	$R^2_{uncorrected}$	R^2_{Normal}	R^2_{Duan}
Lin-log	0.6128	0.6158	0.6158
log-log	0.6601	0.6619	0.6619

Table 4.3: Adjusted R^2 Comparison

In the case of a log-log model, a 1% change in X will result in a β_1% change in Y. Whereas with a linear-log model a 1% change in X is associated with a $0.01\beta_1$ expected change in Y. Which in the context of engine displacement and MPG will be easier to use as increasing engine displacement by a percentage will result in an expected increase in MPG in units Miles. This is discussed further in later sections.

4.4 Spline Transformations of Engine Displacement Combined MPG

Another way to account for the non-linearity of our data is to add knots to our regression line. These knots allow the line to change gradient at different points in the data.

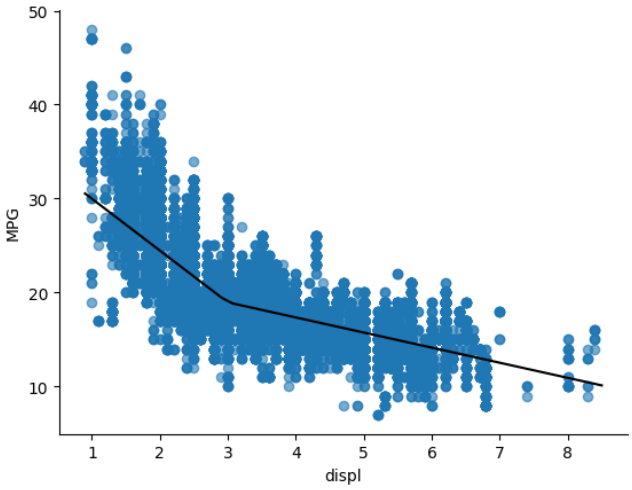

Figure 4.3: Regression with 1 Knot

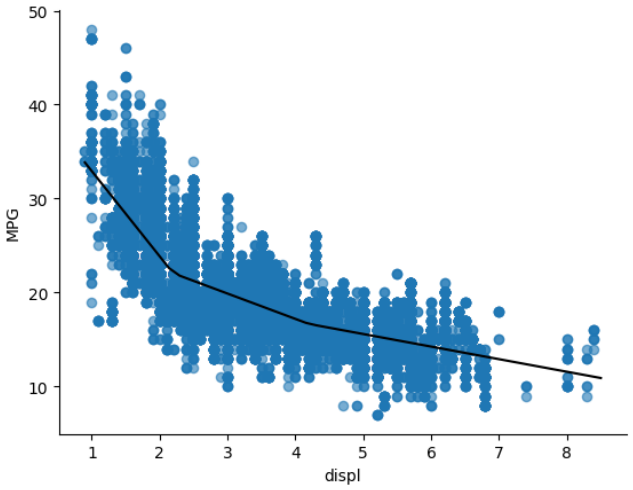

Figure 4.4: Regression with 2 Knots

	R^2
1 Knot	0.649
2 Knots	0.664

Table 4.4: R^2 Comparison of Spline Models

As Shown in Table 4.4, the inclusion of a 2nd knot increased the R^2. We c an also see two changes in gradient seem to fit the model better as shown in figure 4.3 and 4.4.

4.5 Polynomial Transformations

Following general best practice, polynomial transformations of the two main variables were investigated up until the fourth degree. Visual analysis of the fitted regression line to the data points in figures 4.5-4.7 suggest the second and third degree parametric regression of displacement fits the data best.

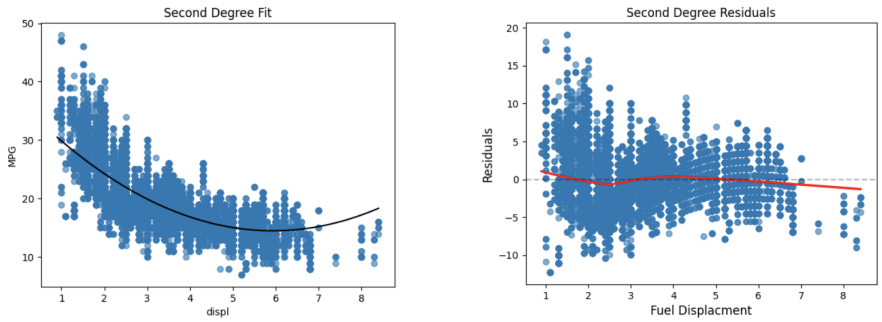

Figure 4.5: 2nd Degree Transformation

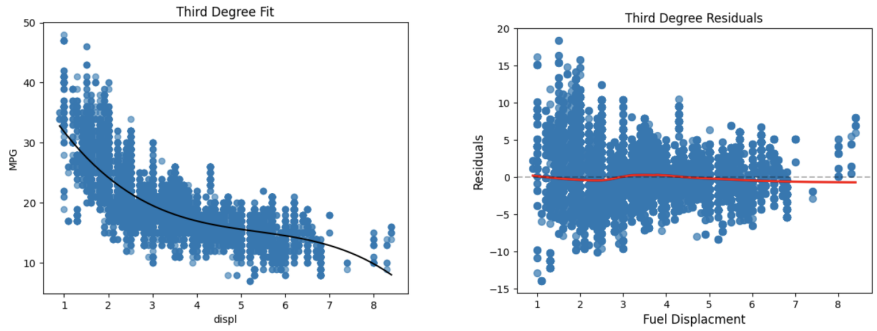

Figure 4.6: 3rd Degree Transformation

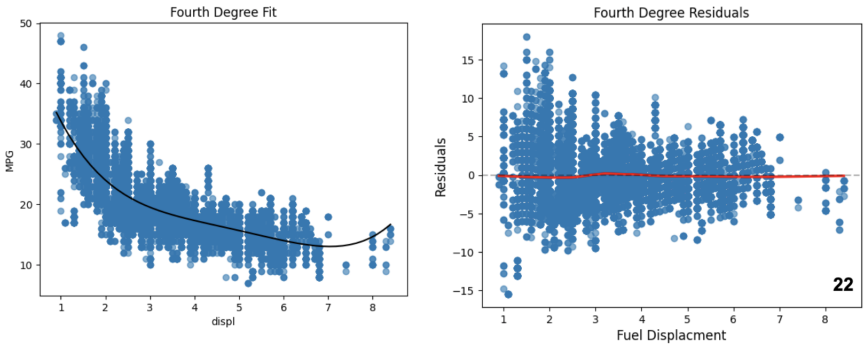

Figure 4.7: 4th Degree Transformation

The fourth degree model demonstrates potential overfitting with tail effects dragging the right hand side of the regression line up to fit the cluster of higher observations. Despite differences in overall shape the third and fourth degree models demonstrate low levels of combined residuals with both overestimating the model slightly before 3 units of fuel displacement. The second degree model shows more erratic distribution of residuals and a less clear fit to the data. Concernedly, all residual plots demonstrate variance decreasing as fuel displacement increases suggesting violation of homoskedacicity, detrimental to each models applicability.

	SER	R^2	R^2_{adj}
$displ^2$	2.914	0.645	0.645
$displ^3$	2.872	0.656	0.656
$displ^4$	2.848	0.621	0.621

Table 4.5: Comparison of Polynomial Models Fit

Each model offers an equally strong fit, accounting for approximately 63% of the variation in data however the aforementioned heteroskedacicity and the known inefficiency of parametric models in predicting data outside of the given range mean that these models and adjustments will likely not feature in our optimal model.

4.6 Selection of an Optimum Model from the Variable and Model Selection Exercises

An optimum model was determined by identifying a regression with the largest adjusted r-squared whilst also limiting the number of variables to ensure the model is parsimonious in its prediction of combined MPG. As the primary function of the model is its predictability of a test set in Section 6.2, the model selection process favoured accuracy of fit over interpretability of variables. Drawing from Section 4.1 and 4.2, all selected variables and significant interaction effects were included in a full model before performing a variable reduction to reach a final optimum model and close alternate for generating forecast predictions. Section 4.3 and 4.4 further motivated the transformations of 'comb08' and 'displ' in the final models. As identified in Section 4.3, the log-log model provided the strongest fit in accordance with adjusted r-squared (using the Duan and normal bias correctors). This motivated the decision to assess a full model with log transformations performed on the response variable and engine displacement.

Full Model with Interactions: $log\_MPG \sim log\_displ + age + startStop + sCharger + tCharger + Manual + displManual + Diesel + Premium + Midgrade + displDies + displPrem + displMid$

The above log-log model was reduced by first assessing the effect on adjusted r-squared of removing the interaction effects before removing main effects in the model. After removing the effects of variables that influenced adjusted r-squared by no more than 0.002, the reduced model was achieved. As explained in Section 4.2, fuel type was reduced to a dummy variable 'Diesel' that identified whether a vehicle used diesel fuel or petrol (default). Beyond the nominated engine displacement, the variables that had the strongest explanatory power for the total variation of the response variable were vehicle age, fuel (diesel vs petrol) and the presence of turbocharging technology.

Reduced Model: $log\_MPG \sim \ log\_displ + age + tCharger + Diesel$

Although Section 4.3 and 4.4 suggested the log-log to be the strongest fit, the one and two-knot spline models were also assessed with the addition of the variables from the new reduced model (age, tCharger, Diesel).

Spline Models:

$comb08 \sim displ + Step + age + tCharger + Diesel$
$comb08 \sim displ + Step1 + Step2 + age + tCharger + Diesel$

For each model, SER, r-squared and adjusted r-squared were calculated and tabulated below. The entries for log models were adjusted further using the Duan bias corrector.

Model	SER	R^2	R^2_{adj}
Log-Log All Variables	2.274	0.785	0.785
Log-Log All Variables + interactions	2.247	0.790	0.790
Log-Log Reduced	2.291	0.782	0.782
Spline (single) Reduced	2.358	0.769	0.769
Spline (double) Reduced	2.322	0.776	0.776

Table 4.6: Model Fit Comparison

The strongest fitting model was the log-log transformed model with all main and interaction effects included. As the model was fit with 13 variables, the principle of parsimony may suggest we select a model with less variables but a similarly accurate fit. Favouring the reduced log-log model may also lead to better predictability of the test data as a consequence of overfit in the full model. As an alternative, the two-knot linear spline fit is also similarly parsimonious and performs relatively well in the SER and adjusted r-squared metrics.

	coef	$P > \|z\|$
int	3.6819	0.00
log_displ	-0.508	0.00
age	1.88	0.00
tCharger	-0.0739	0.00
Diesel	0.3144	0.00

Table 4.7: Optimal MLR Model

Collinearity in the final optimum model has been handled through variable selection in Section 3.7 and 4.1. Using the threshold of 3 and 5 for individual and mean variance inflation factor we maintain that each variable, X_j, brings in information above the information from the other variables. One concern however when comparing the final coefficients and Figure 1.6 (Section 1.3) is the change in sign of the tCharger variable from positive to negative. As such the variable inference should be doubted to an extent due to multicollinearity with other variables.

The optimum model can be diagnosed by performing an overall significance test with the following null and alternate hypotheses.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$
$$H_1 : \beta_1, \beta_2, \beta_2, \beta_4 \neq 0$$
$$\alpha = 0.05$$

The test statistic is $F_{stat} = 21440$ which follows an $F_{4,n-5} = F_{4,3.16\times10^4}$distribution under the null hypothesis. As the p-value,$P(F_{4,3.16\times10^4} > 21440) = 0$, we can reject the null and conclude that at least one variable significantly predicts the response variable.

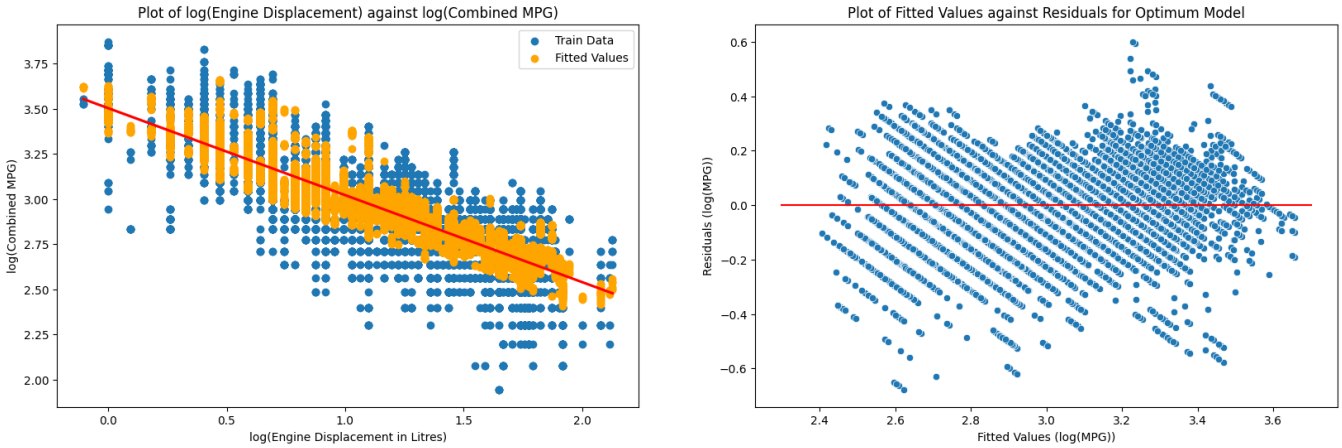

Figure 4.8: Optimal Model

The plot of the optimum model with axis log('displ') and log('comb08') provides a relatively strong linear fit (LSA 1). Analysis of the plot of fitted values and residuals also shows that the distribution of errors roughly fits the LSA of exogeneity and linearity. Similarly, the relatively constant band of variance in the residuals with fitted values allows us to satisfy the LSA of homoskedasticity. As such, the optimum model and the introduction of a nonlinear transformation corrects the linear models fitted in Section 2.1 and 3.4. One limitation of the model is still the uncertainty in the independence of the response variable and predictors that requires more information on the sampling of vehicles.

Optimal Model

$$\widehat{Log(MPG)} = 3.6819 - 0.508Log(displ) + 1.88age - 0.0739tCharger + 0.3144Diesel$$

5

Discussion of results and conclusions using training data set

5.1 Conclusions Regarding Overall Goals

In light of the optimum model drawn from this exploration using the training data set, their exists a negative linear relationship between fuel economy and engine displacement on the logarithmic scale. This is such that a 1% increase in engine displacement translates to a constant percentage decrease in the expected combined MPG metric. This is validated by the assessing the assumptions drawn from Section 4.3 and 4.5 where log-log models of engine displacement and fuel economy provided the best fit with exogeneity in the residuals. An alternative view, using the reduced two-knot spline model, is that engine displacement and combined MPG have a negative linear relationship that undergoes 2 gradient changes at the first and third quartiles that dampen the negative relationship between the two variables (as discussed in Section 5.2). Beyond engine displacement, the variables with the strongest explanatory power were age, turbocharging and use of diesel fuel. By restricting the variables selected in Section 3.3 and 4.1, the optimum model employs variables that strictly precede and have dependence on the response variable. By minimising the effect of confounding factors through multicollinearity and VIF tests, we were able to develop a strong causal model for fuel economy with the selected variables. The strongest limitation on the causal relationship is the lack of knowledge of vehicle sampling for the data set tested. A truly randomized experiment would allow us to more confidently eliminate confounding factors in the final model. In developing the overall optimum model, the risks of overfit have been managed by limiting the number of variables and the parsimony of transformations/non-linear effects. Additionally, assessing an alternate model's forecasting accuracy is apt in finding a strongest overall predictive model.

5.2 Interpretation of Selected Optimal Model

Using the optimum model selected in Section 4.5, the prediction of change in combined MPG with a 1% increase in engine displacement (in litres) is -0.51% assuming ceterus paribus. The unit increase in age of the vehicle is associated with a -0.67% predicted change in combined MPG. Finally, the dummy variables infer that turbocharging and diesel fuel (over petrol) causes a -7.39% and 31.44% predicted change in combined MPG respectively. As for the alternative model, in the first quartile of 'displ', a unit increase in engine displacement (in litres) translated

to a -9.32 predicted unit change in combined MPG. Between the first and third quartile the predicted unit change in combined MPG with a unit increase in engine displacement was -2.81. Finally in, in the final quartile of 'displ', a unit increase in engine displacement translated to a -1.58 predicted unit change in combined MPG. The direction of change in combined MPG with the remaining variables in the alternate model is similarly reflected in the optimum model.

6

Generating and Assessing Forecast Predictions

	RMSFE	MAFE	Forecast R^2	p	SER	R^2_{adj}
Log-log Reduced	0.1126	0.0862	77.9	4	2.291	78.2
Log-log Full	0.1108	0.0847	78.6	13	2.247	79.0
2 Knot Spline	2.2974	1.7284	77.5	6	2.322	77.6

Table 6.1: Accuracy of forecast predictions for relevant models

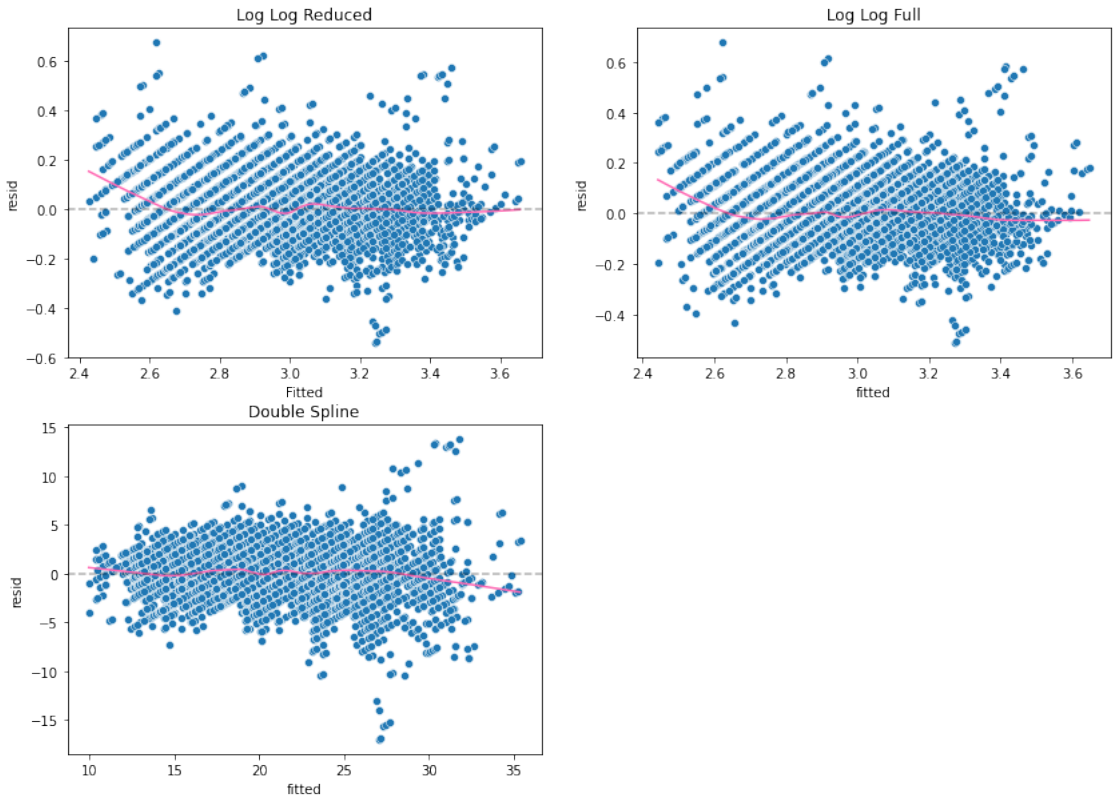

Figure 6.1: Residuals vs Fitted Values using test data set

Root Mean Squared Error (RMSE), measures the difference between the values predicted by the model, and the values observed. Therefore, a low RMSE is indicative of higher performing models. Shown in Table 6.1, the best performing in all metrics is the log-log full model. Although with 13 predictors, compared to the log-log reduced model's 4 predictors it is not significantly better than the reduced model. Thus we apply the principle of parsimony; the simplest model among equally well fitting models is preferable.

When residuals of each model are plotted against the fitted values shown in figure 6.1, we can see a strong fit from each model. The Log-log Full and reduced are almost indistinguishable, although it should be noted that there are 'start' effects present in both i.e. the model is less reliable when predicting very small changes in MPG. Whereas, 2 knot model becomes a poorer fit when predicting very high instances of MPG.

Thus, our conclusions regarding the overall goals of this study do not change based on forecast predictions. The optimal model is still considered to be the Log-log Reduced with an alternative being the 2 knot spline as this may be easier to interpret depending on the use case.

7

Final Conclusions and Recommendations

7.1 Summary of Findings

Following the directive of the Department of Energy Office of Energy Efficiency, our team has investigated the relationship between a vehicles fuel efficiency and the total displacement of its engine. Further we have demonstrated the way in which other variables such as a vehicles age, the presence of a turbo charger and diesel fuel sources may mediate this relationship.

Firstly, following the departments instructions we investigated a simple linear fit between a the fuel efficiency of a car as measured by miles per gallon (MPG) and the total displacement of fuel in its engine measured in liters (displ). The linear relationship between these two variables found to be such that for every litre of total engine displacement there is a decrease of 2.79 miles per gallon in fuel efficiency.

Continued analysis of this model however revealed that only 58.5% of the variance in MPG can be explained by fuel displacement. Furthermore several of the key mathematical underpinnings of the linear relationship described were violated such that we cannot confidently label this relationship as significant. This lead to our team investigating more complex models predicting MPG.

Next, we considered a multiple linear regression model whereby we included extra explanatory factors in our model. The number of cylinders a car has, the quality/type of the fuel, the presence of a turbo charger and whether a car is manual or not, were all considered relevant. This mixed model then increased our explanation of MPG to 63.5 percent, however once again violations in mathematical underpinnings of the relationship sought us to seek further adjustments to the model.

Specifically, a number of the relationships between variables in this model are non-linear meaning we must adjust the data sets to facilitate a stronger linear relationship between variables. Sparing technicalities we investigated logarithmic adjustments, polynomial changes, splines (or piecewise linear functions) as well as a mix of them all. Similarly we examined the interaction of various elements within the model, for instance how the number of cylinders a car has and

its overall fuel displacement are inherently linked and thus must be isolated from one another.

After making all these adjustments we came to our final Optimal model:

$$\widehat{Log(MPG)} = 3.6819 - 0.508Log(displ) + 1.88age - 0.0739tCharger + 0.3144Diesel$$

Interpretation of this model is discussed in detail in section 5.2 but in essence the model suggests that the most relevant factors effecting the overall MPG/fuel efficiency of a car are:

- Engine Fuel Displacement (as given)
- The age of the vehicle
- Diesel Fuel

Further, the presence of a turbo charger was shown to have a significant negative impact on fuel efficiency, such that it is included in our final model.

7.2 Recommendations

In light of the findings of this report, it is clear that if the US Department of Energy would like to increase the fuel efficiency of cars on US roads they should incentivize cars with lower engine displacement and penalize cars with larger displacement. It may also consider investing in research and development of more efficient engines. As may be expected the age of a car has a strong influence on its efficiency, thus a possible consideration could be to impose a tax on the sale of cars past a certain age to encourage the use of newer more efficient cars.

Further studies are required to determine the ideal range of engine displacement to achieve the desired efficiency if they are looking to regulate displacement. These future studies may also take into account the impact of hybrid engines as ours did not.