DATA 603 – Final Report

GROUP 2 - SECTION L01

Rodney Sibanda - 3003

Neha Adnan

Zwaiba Khan

Rehan Chanegaon

Zhaoye LIU

DATA 603 - Statistical Modelling with Data

Dr. Thuntida Ngamkham

November 30, 2024

INTRODUCTION

Context:

Fuel efficiency can be defined as the distance a vehicle can travel per unit of fuel consumption (U.S. Department of Energy). It is typically measured in either miles per gallon (mpg) or liters per 100 kilometers (L/100KM). This topic holds significance due to the fact that fuel efficiency impacts both the environment and personal finances. For example, with inflated fuel costs and intensified climate change, it is increasingly important to choose vehicles that maximize fuel efficiency. The Canadian Climate Institute reported that in 2022, the average Canadian household spent approximately $2,000 per year on gasoline (Canadian Climate Institute, 2022). Furthermore, the U.S. Environmental Protection Agency (EPA) states that if U.S. households with multiple vehicles drove their most fuel-efficient cars, it would result in a reduction of 100 million metric tons of CO2 emissions annually (EPA, 2024). These statistics underscore the importance of understanding the factors that contribute to fuel efficiency.

Challenges:

There are several challenges that come with trying to predict fuel efficiency. Firstly, while fuel efficiency is crucial for reducing both environmental impact and costs, it is not the only factor consumers consider when purchasing a vehicle. Attributes such as power, performance, and vehicle type often conflict with fuel efficiency. For instance, larger engines with greater displacement and more cylinders provide more power but consume more fuel. For example, a V8 engine takes in more air and fuel to produce greater power but burns significantly more fuel compared to a six-cylinder engine (Toyota Canada, 2021). Similarly, all-wheel drive (AWD) vehicles, which offer better traction and higher performance in extreme weather, require additional power and are less fuel-efficient than front-wheel drive (FWD) vehicles (Lethbridge Toyota, 2023). Additionally, vehicle attributes such as engine cylinders, drivetrain type, and displacement may interact in nonlinear ways, making it difficult to isolate their individual impacts on fuel efficiency.

Objectives:

Our project aims to fit a multi-linear regression model to the MPG dataset from the ggplot2 package, which provides information on various car attributes and their fuel efficiency in miles per gallon. Our main objectives include quantifying the relationship between vehicle attributes and fuel efficiency, identifying attributes that contribute most significantly to fuel efficiency, and exploring trade-offs between performance-related features and fuel consumption. Additionally, we hope to answer the following research questions: Which car attributes have the greatest impact on fuel efficiency, how do performance-related attributes, such as engine displacement or drivetrain type, trade off against fuel efficiency, and can a regression model using car attributes be used to effectively predict fuel efficiency? Ultimately, our findings will help consumers balance performance and sustainability, enabling smarter vehicle choices.

METHODOLOGY

Data

For our project, we are using the MPG dataset, an "open" or "shared" dataset available through the ggplot2 package within the tidyverse library in R. This dataset was not created by any members of our group but is instead provided by the creators of the tidyverse for educational analytical purposes. More information on this about this ggplot2 and this data can be found here. The dataset is derived from fuel economy data provided by the Environmental Protection Agency of the United States government.

This dataset includes 235 rows and contains various attributes related to vehicle economy, such as manufacturer, model, engine displacement, and fuel efficiency in city and highway driving. Each row represents a specific car model and its respective characteristics.

The data was collected as a part of the EPA's regular fuel economy testing, which evaluates vehicles under controlled conditions to estimate fuel efficiency in city and highway scenarios. This dataset represents a subset of popular car models released from 1998 to 2008, focusing on vehicles with new releases during this period.

Using an open dataset like this ensures that our project is built on reliable, well-documented data that has been widely used and tested in the data science community. This shared dataset also allows us to focus on analysis and modeling without the time and resources required of collecting and cleaning raw data. By leveraging this open resource, we can align our methodology with best practices and concentrate on deriving meaningful insights and building predictive models.

A detailed breakdown of the dataset is provided below:

Variable	Description	Type	Unit/Category	Qualitative/Quantitative
Cty	City miles per gallon	Continuous	Miles per gallon	Qualitative

Hwy	Highway miles per gallon	Continuous	Miles per gallon	Qualitative
Displ	Engine displacement indicating the engines size	Continuous	Liters	Quantitative
Manufacturer	The manufacturer of the vehicle	Categorical	Name of Manufacturer	Qualitative
Model	The model of the vehicle	Categorical	Name of model	Qualitative
Trans	Type of transmission in the vehicle (automatic or manual)	Categorical	Example: auto(l5), (automatic, 5 gears)	Qualitative
Year	The year of manufacture	Discrete	Year(2008 or 1999)	Qualitative
Cyl	Number of engine cylinders	Discrete	Count (4,5, 6, or 8)	Qualitative
Drv	Drivetrain Type	Categorical	F,r, or 4 (f = front-wheel drive, r = rear-wheel drive, 4 = 4-wheel drive)	Qualitative
fl	Fuel Type used by the vehicle	Categorical	R,p,d,e (regular, premium, diesel, or ethanol)	Qualitative
Class	Vehicle class representing the type of car	Categorical	Compact, SUV, midsize, 2 seater, minivan, pickup subcompact	Qualitative

Approach

The Primary statistical approach of this analysis is multiple linear regression (MLR) modeling. MLR was used to identify and attempt to quantify the relationship between the response variable, combined fuel efficiency (mpg), and several predictor variables, such as engine displacement, number of cylinders, drivetrain type, fuel type, and class. This statistical method works so well because the predictors are both categorical and numerical which allowed for flexibility in modeling complex relationships.

To Refine the model, the following steps will be taken:

Data Preparation:

The data was inspected, and we found that there were two potential dependent variables, city and highway MPG. We combined this to create a single dependent variable, combined MPG. Additionally, we refactored the manufacturer variable to cover regions of production instead in order to reduce the number of categories.

Base Model Creation

An initial MLR model is built including all potential predictors, allowing us to evaluate their overall significance.

Stepwise Selection

An automated selection approach, using the ols_step_both_p function, identifies significant predictors by entering or removing variables based on their p-values, with thresholds set to p_enter = 0.05 and p_remove = 0.1. By using these parameters, we ensure that only variables that are contributing meaningfully to the model are retained.

Variable Refinement:

Non-significant variables are removed individually based on t-tests in order to get our reduced model. For categorical variables, if one level is significant, all other levels are kept. The significance level was set to 5% for the hypothesis testing that was done for the t-test. This means that terms would be kept if they fall below this significance level and are deemed significant, or consequently they are removed if they fall above the significance level and are deemed insignificant predictors for mpg.

Inclusion of Interaction Terms:

Interaction terms are introduced to attempt to create a better fitting model in predicting MPG and to capture the dependencies between predictors. These terms are tested using the partial T-test with a significance level of 5%. Conducting our hypothesis test, relevant or significant terms are retained if their p-values fall below our significance level and are disregarded if they fall above.

Testing Higher-Order Terms

Polynomial terms are evaluated to account for potential non-linear relationships and to possibly create a better fitting model.

Residual Analysis and Transformation

The following assumption are tested:

- Linearity is checked via residual vs fitted plots.
- Equal Variance is assessed using the Breusch-Pagan Test, where the hypothesis test checking for the presence of heteroscedasticity or homoscedasticity conducted at a significance level of 5%.
- Normality of residuals is evaluated with QQ plots and the Shapiro-Wilk test at significance level of 5%.
- Outliers were checked using Cook's distance, and were removed if the distance was greater than 0.5
- Multicollinearity was not checked as all variables but one are categorical.
- If there is failure to meet these assumptions, then transformations will be conducted. This includes log and Box-Cox transformations of the response variable.

Final Model Selection

The model retains significant predictors and interaction terms identified through all of the steps above.

Workflow

Data Processing

- Read and Clean Data
- Create a combined MPG variable which will be a weighted average of city and highway MPG
- Categorize manufacturers into regions for exploratory purposes

Initial Model Building

- Fit a base model with all predictors
- Use Stepwise regression to identify the most significant predictors.

Refining the model

- Test interaction terms for significance
- Eliminate non-significant interaction terms to simplify and reduce the model

Model Assumptions

- Check for linearity, equal variance, outliers, and normality of residuals
- Visualize residuals using plots and performing Shapiro-Wilk and BP statistical tests.

Transformations:

- Apply log or Box-Cox Transformation to improve model assumptions
- Reassess the model's performance and assumptions after transformation is completed

Final Model

- Choose the model with the best adjusted R_2, significant predictors, and reasonable diagnostics results.

Handling Challenges

The hardest step was during the data processing phase. The dataset contained a high number of unique values for the manufacturer and model variables. This created significant challenges due to the high correlation observed between the numerous levels of these variables, which led to multicollinearity.

Resolving this required careful deliberation. After discussions with both TA's and Professor Thunthida, the consensus was to categorize manufacturers into broader regional categories and exclude car models entirely from the base model. This approach ensured the model was more interpretable and reduced multicollinearity while maintaining the overall explanatory power.

Contributions

The project team was organized with specific roles to ensure efficient collaboration and progress. Neha, as the Meeting Chair, led the meetings, ensuring the agenda was fulfilled and discussed. Rodney, the Meeting Scheduler, coordinated meeting times and locations, and facilitated the setup of shared files for online collaboration. Zhaoye, the Project Manager, set internal deadlines and tracked task completion for coding and writing. Rehan, the Code Manager, oversaw code quality to ensure compliance with the team's unified standards. Zwaiba was responsible for Meeting Minutes and helped mediate and resolve internal conflicts while documenting meeting outcomes.

Regarding specific academic tasks, Rodney contributed to writing and presenting the methodology section in both the repor0t and presentation. Neha authored the introduction and results sections, while Hunter and Rehan wrote and presented the results section. Zwaiba wrote and presented the conclusion.

RESULTS

Data Preparation:

According to the description of our dataset and our goal, the dependent variables were fuel efficiency in city driving (cty) and fuel efficiency in highway driving (hwy). The independent variables consisted of manufacturer, model, engine displacement (displ), year, cylinder count (cyl), transmission type (trans), drive type (drv), fuel efficiency (fl), and vehicle class (class).

Among the independent variables, engine displacement was numeric, while the rest were categorical. The manufacturer and model variables contained many categories—15 and 35, respectively, and had a hierarchical relationship. In this context, the hierarchical relationship means that the model variable is dependent on the manufacturer variable. A manufacturer (such as Ford or Toyota) produces multiple models (like Ford Focus or Toyota Corolla), so each model belongs to one specific manufacturer, but each manufacturer may have many models. Due to this and the fact that the model variable had many categories, we excluded it and restructured the manufacturer variable. Specifically, we grouped manufacturers by region, creating three categories: European, American, and Asian. This also resulted in fewer categories as compared to the original variable, making our model easier to interpret and test. This left us with the following independent variables: region, engine displacement, year, cylinder count, transmission type, drive type, fuel efficiency, and vehicle class.

Additionally, we ran into the challenge of two dependent variables in the dataset. We decided to use these two variables to create a single dependent variable that could be tested using multiple linear regression (MLR). Specifically, combined fuel efficiency is a common metric that is reported when trying to determine the overall efficiency of a vehicle. The formula for calculating this involves calculating a weighted average, where the highway fuel efficiency is given a 45% weight, and the city fuel efficiency was assigned a 55% weight. We applied this formula to our dataset, which resulted in the new variable named combined.

Base Model Creation:

To commence our analysis, we first created a base linear regression model that included all potential independent variables, following the formula:

$$\text{Combined} = \beta_0 + \beta_1 \cdot \text{displ} + \beta_2 \cdot \text{cyl} + \beta_3 \cdot \text{trans} + \beta_4 \cdot \text{drv} + \beta_5 \cdot \text{fl} + \beta_6 \cdot \text{class} + \beta_7 \cdot \text{year} + \beta 8 \cdot \text{region}$$

Stepwise Selection:

After creating the base model, we used the OLS best subset functions to gain an overall sense of which variables might be significant predictors. The OLS output indicated that the most significant predictors were displ, cyl, drv, fl, class, and year, while region and transmission were not significant and were therefore excluded from the model (Figure 1).

Variable Refinement:

Due to the fact that the best subset method does not handle categorical variables that appear as numerical in the dataset, like year and cyl appropriately, we retested the model using individual t-tests and the summary function. This allowed us to properly assess the significance of the categorical variables, such as year and cyl, and ensure that their effects were correctly captured.

In this case the null and alternative hypotheses tested at an alpha of 0.05 were as follows:

$H_0 : B_k = 0$ where k = displ, cyl, drv, fl, class, year, region, trans.

$H_A : B_k \neq 0$ where k = displ, cyl, drv, fl, class, year, region, trans.

The t-test results showed that the following results (Figure 2):

- Engine displacement (displ): p-value of 8.63e-05, t-statistic of -4.006, meaning we reject the null hypothesis, indicating that this predictor is significant.
- Cylinder types (cyl): The categories cyl 6 and cyl 8 were significant, with p-values of 2.51e-06 and 0.008889, respectively. However, cyl 5 had a p-value of 0.304932, indicating that it was not significant. However, due to other levels of this categorical variable being significant, we retained it.
- Drive type (drv): Both drv f and drv r were significant, with p-values of 7.56e-07 and 0.066925, respectively. While the p-value for drv r is slightly above 0.05. However, due to other levels of this categorical variable being significant, we retained it.
- Fuel type (fl): The categories fld, fle, and flp were significant, with p-values of 0.002387, 2.48e-05, and 0.039916, respectively. The category flr had a p-value of 0.063594. Because all other levels of this categorical variable are significant, we kept flr.
- Vehicle class (class): All levels of class were significant, with p-values less than 0.05. For example, classcompact had a p-value of 0.001573, and classpickup had a very significant p-value of 2.52e-10.
- Year of manufacture (year): The year 2008 category had a p-value of 7.10e-06, meaning we rejected the null
- Region: The Asian region had a p-value of 0.820561, and region European had a p-value of 0.094277. Since no levels of this variable were significant, we failed to reject the null and concluded that this variable was not significant.
- Transmission type (trans): No levels of the 9 levels of variable were found to be significant. This means that we failed to reject the null, and did not include this variable in our first order model.

Upon discovering which variables were significant and which were not, we refitted our model to only include significant variables and re-ran the summary function. The variables included were displ, cyl, drv, fl, class and year. We retested our null and alternative hypotheses (at an alpha of 0.05) as stated below to ensure all variables were still significant:

$H_0 : B_k = 0$ where k = displ, cyl, drv, fl, class, year.

$H_A : B_k \neq 0$ where k = displ, cyl, drv, fl, class, year.

From the output (Figure 3), we observed that all terms in the model were statistically significant, as the p-values for each are below 0.05, indicating that we reject the null hypothesis that the coefficients are equal to zero. However, there are two exceptions: the r level of fuel (with a p-value of 0.084142) and the 5 level of cylinder (with a p-value of 0.062633). Despite these p-

values, all other levels of both the fuel and cylinder variables are significant. Therefore, we will retain these variables in the model.

This made our first-order order model as follows:

$$\hat{combined} = 15.7793 - 0.5743_{displ} - 0.7893_{cyl_5} - 1.1993_{cyl_6} - 1.4512_{cyl_8} + 1.0645_{drvf} + 0.4868_{drvr} + 2.6555_{fld} - 3.6000_{fle} - 1.8439_{flp} - 1.3974_{flr} - 1.7561_{classcompact} - 1.8105_{classmidsize} - 3.1158_{classminivan} - 3.3563_{classpickup} - 1.7460_{classsubcompact} - 3.1880_{classsuv} + 0.6358_{year2008}$$

This model has an adjusted R squared of 0.9 and an RSE of 0.7851 (Figure 3). This means that 90% of the variance in the combined mpg can be explained by the current model. Additionally, the RSE of 0.7851 indicates that the model's predictions of combined mpg vary by about 0.7851 miles per gallon compared to actual values.

Inclusion of Interaction:

Following the development of the first order model, we proceeded to test all possible interaction terms by raising the model to the second power. We then tested the significance of these terms using the t-test. Where the null and alternative hypotheses for the terms were as follows and tested at an alpha of 0.05:

$$H_0 : B_{inter} = 0 \text{ for each interaction term}$$
$$H_A : B_{inter} \neq 0 \text{ for each interaction term}$$

From the output we encountered issue with interactions between categorical variables, where the majority of levels were either insignificant (p-value > 0.05) or had missing data (NA). Since categorical-categorical interactions lack interpretability and do not make much sense in the context of the model, and for the reasons above, we did not end up retaining any categorical-categorical interaction. Specifically the following variables were excluded for reasons outlined:

- Drivetrain and class: 10 out of 12 levels were NA, and although two levels were significant, the overwhelming presence of NA levels made this interaction unsuitable for inclusion.
- Fuel and year: All levels were either insignificant (p-values less than 0.05) or NA, so we excluded this interaction.
- Class and year: None of the levels were significant, leading to the exclusion of this interaction.
- Drivetrain and year: Similarly, none of the levels were significant, so we removed this interaction as well.
- Fuel and class: 18 levels out of 24 were NA, the remaining 5 were all insignificant.
- Drivetrain and class: Two out of twelve levels were significant, but the remaining ten were NA.

- Cylinders and drivetrain: Two out of six levels were significant, one was insignificant, and three were NA.
- Drivetrain and fuel: One out of 9 levels were significant, two were insignificant, and five were NA.
- Cylinders and year: One out of 3 levels was significant, one was NA, and one was insignificant.

For interactions involving a numerical variable and a categorical variable, we followed a different methodology. Even if only one level of the interaction term was significant, we retained the entire interaction term. This approach was adopted because numerical-categorical interactions are interpretable and essential for understanding how a numerical variable (e.g., displacement) interacts with categorical levels.

Specifically, as seen in Figure 4:

- Displ and class: Two out of six levels were significant (p-value < 0.05), specifically for compact and mid-size categories, while the remaining levels (minivan, pickup, subcompact, and SUV) were not significant.
- Displ and fuel: One out of four levels was significant, with fld being significant, while the levels fle, flp, and flr had p-values greater than 0.05.
- Displ and cyl: One level (5 cylinders) was NA, but 6 and 8 cylinders were significant.
- Displ and drv: The F level was significant, while the R level was not.
- Displ and year: Only one level was tested, and it was not significant.

Therefore, the final interaction terms found to be significant included displ:cyl, displ:fl, displ:drv and displ:class.

Once again, we had to refit the model to only include significant interaction terms. The interactions were once again tested at an alpha of 0.05, with the following null and alternative hypotheses:

$H_0 : B_{inter} = 0$ for each interaction term (displ:fl, displ:drv, displ:cyl, displ:class)

$H_A : B_{inter} \neq 0$ for each interaction term (displ:fl, displ:drv, displ:cyl, displ:class)

From the output (Figure 4) we found that

- All levels of displ:cyl are significant with p values less than 0.05
- All values of displ:drv are significant
- 2 out of 4 levels of displ:fl are significant (fld and fle) with flp being non-significant and flr being NA
- Only class midsize is significant for displ:class, while compact, minivan, pickup, subcompact, and SUV are all non-significant

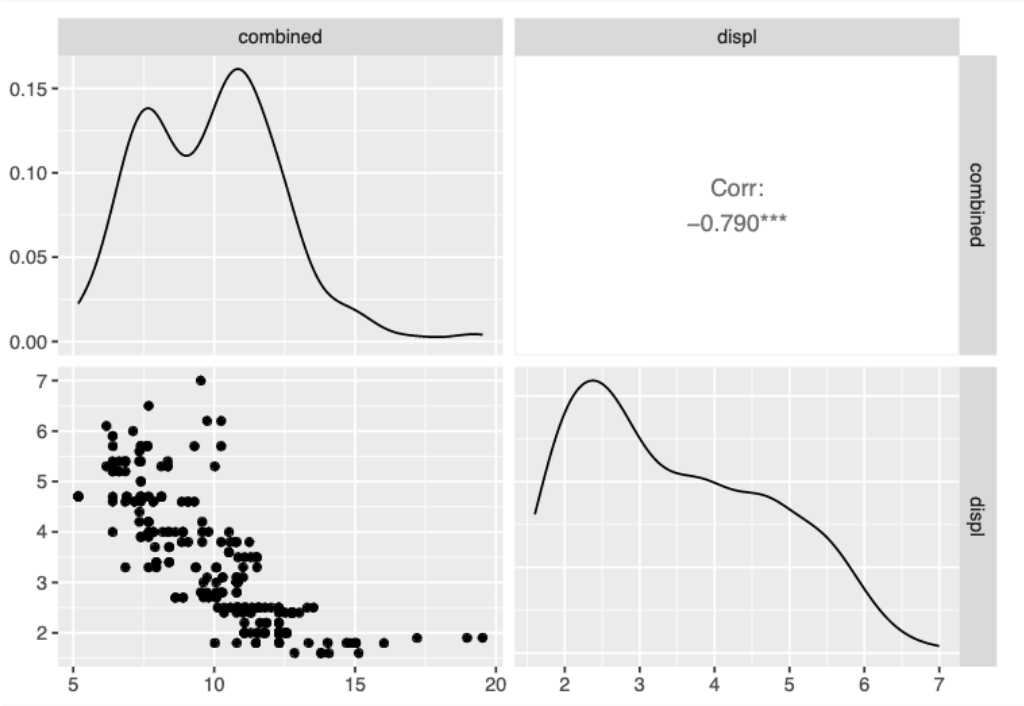
Based on these results we will retain all of these interactions terms, as each one had at least one level significant. This makes our final interaction model as follows:

$$\hat{combined} = 15.7793 - 0.5743_{displ} - 0.7893_{cyl_5} - 1.1993_{cyl_6} - 1.4512_{cyl_8} + 1.0645_{drvf} + 0.4868_{drvr} + 2.6555_{fld} - 3.6000_{fle} - 1.8439_{flp} - 1.3974_{flr} - 1.7561_{classcompact} - 1.8105_{classmidsize} - 3.1158_{classminivan} - 3.3563_{classpickup} - 1.7460_{classsubcompact} - 3.1880_{classsuv} + 0.6358_{year2008} + 1.0815_{displ*classcompact} + 1.5135_{displ*classmidsize} + 1.2339_{displ*classminivan} + 0.1248_{displ*classpickup} + 0.2292_{displ*classsubcompact} + 0.0595_{displ*classsuv} + 0.9647_{displ*cyl_5} + 1.5214_{displ*cyl_6} + 2.4410_{displ*cyl_8} - 0.8638_{displ*drvf} - 0.8775_{displ*drvr} - 1.0701_{displ*fld} + 0.8622_{displ*fle} + 0.1268_{displ*flp} + 0.9647_{displ*cyl_5} + 1.5214_{displ*cyl_6} + 2.4410_{displ*cyl_8} - 0.8638_{displ*drvf} - 0.8775_{displ*drvr} - 1.0701_{displ*fld} + 0.8622_{displ*fle} + 0.1268_{displ*flp}$$

This model has an adjusted R squared 0.9304 and an RSE of 0.6551 (Figure 4). This value indicates that 93.04% of the variation in the dependent variable, combined fuel efficiency, is explained by the independent variables included in the model, and the fit has improved from adding interaction terms. Similarly, the lower RSE also indicates a better fit than the prior model, as a lower value indicates that predictions are closer to the actual data points.

Testing Higher-Order Terms

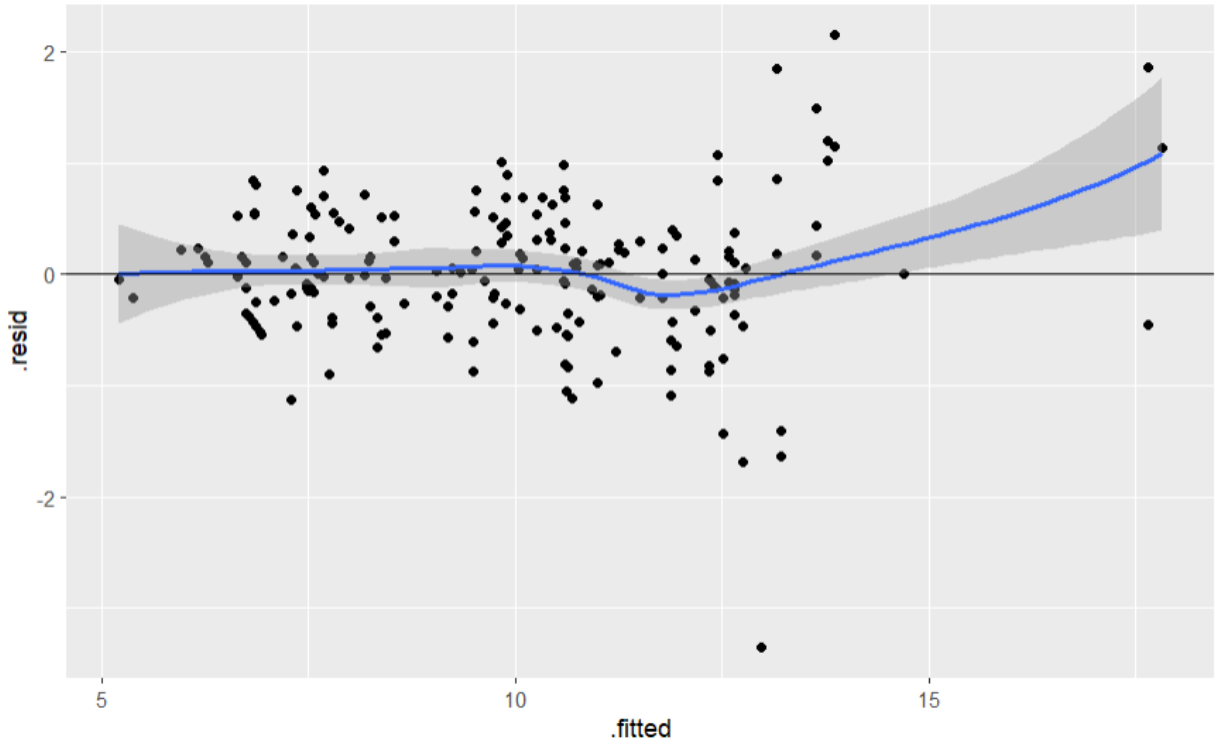
We then moved on to testing higher order terms. Since we only had one numerical variable, which was displacement, we created a plot exploring the relationship and correlation between it and the dependent variable. As seen in the figure below, the correlation between displacement and the dependent variable was found to be -0.790, suggesting a negative linear relationship. However, the relationship between displacement and the dependent variable appears to exhibit a higher-order effect.



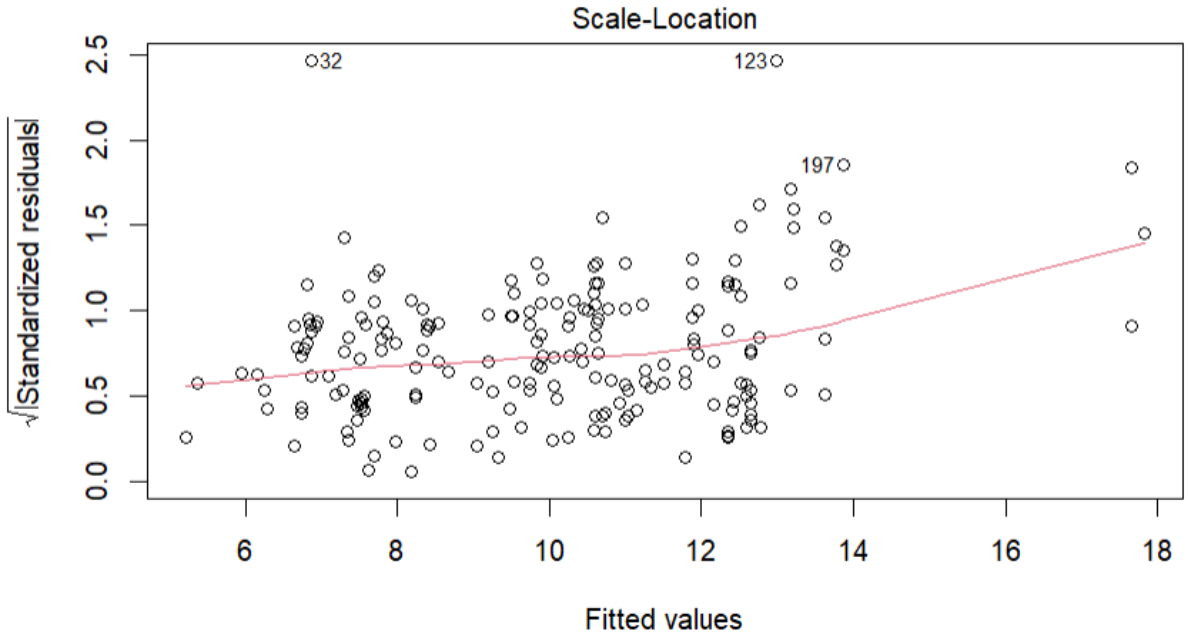
Based on this, we added a quadratic term to the equation and reran the summary function (Figure 5) and used the t-test to evaluate its significance. The p-value of the added term was found to be 0.242237 and the t- value was 1.173, indicating that it was not statistically significant. As a result, we returned to the interaction model presented earlier for further analysis. This was determined to be our final model that we would use, we then proceeded to test assumptions.

Residual Analysis and Transformation:

First, we checked the assumption of linearity using a residual vs fitted plot. As seen in the figure below, we found that the data did not appear to be linear, as the plot shows a slight curvature, suggesting some non-linearity in the model. This indicates that the relationship between the predictors and the response may not be fully captured by the current linear model.



Next, we proceeded to test the assumption of homoscedasticity, using both the scale location plot below, and the Brusch Pagan Test (Figure 6). As seen in the plot below, there does appear to be a slight pattern where we observe an increasing spread of residuals as fitted values increase, indicating potential heteroscedasticity.

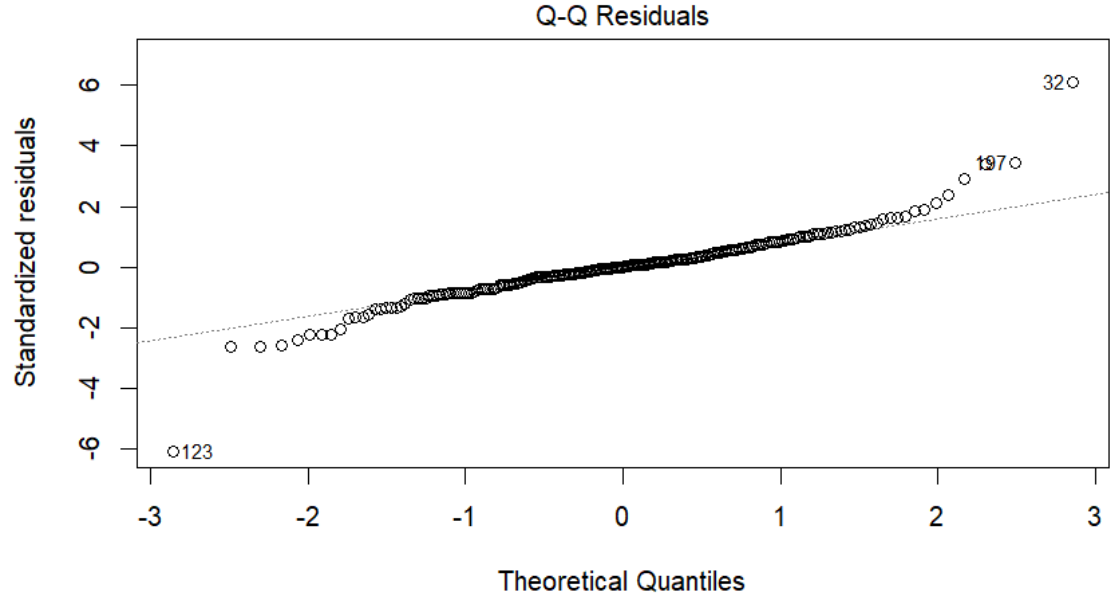


This was confirmed through the Breusch-Pagan test on the model. In this case the null and alternative hypothesis tested are as follows:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2$$
$$H_A : \sigma_1^2 \neq \sigma_2^2 \neq \cdots \neq \sigma_n^2$$

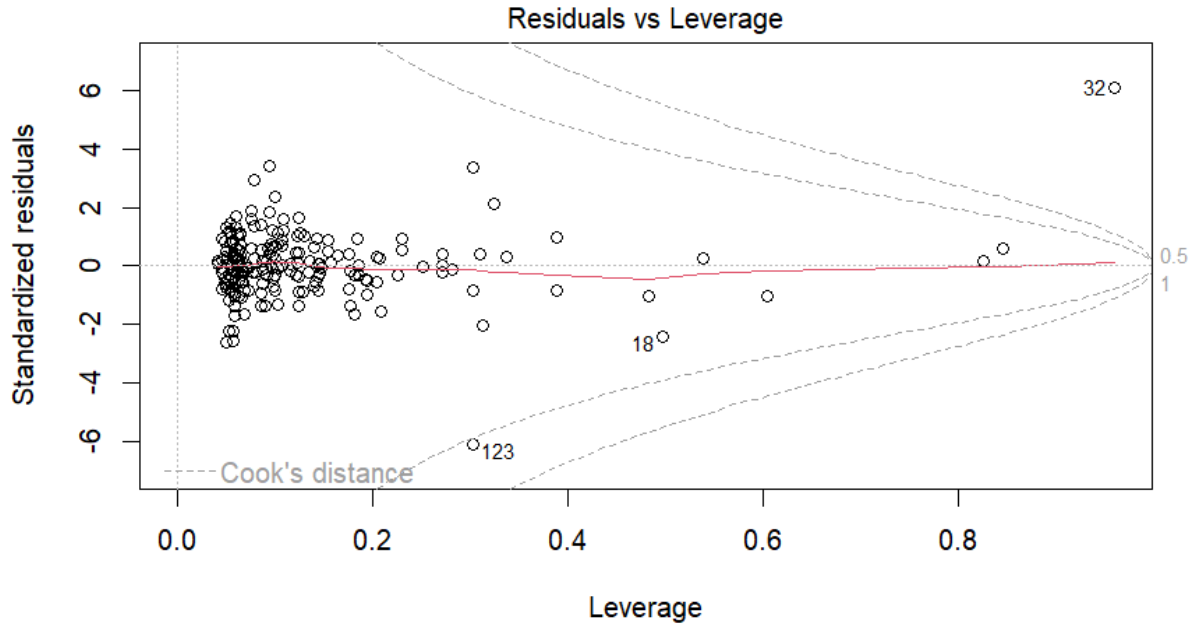
As seen in Figure 6, the test statistic of the Breusch-Pagan test is 87.414 and the p-value is 0.00000009055, which is much lower than the significance level of 0.05, so we reject the null hypothesis that the data has equal variance. We conclude that heteroscedasticity exists in this model.

Next, we tested the assumption of Normality, using a qq plot, along with the Shapiro Wilk Test. As seen in the plot below, there is heavy deviation at tails indicating non normality.

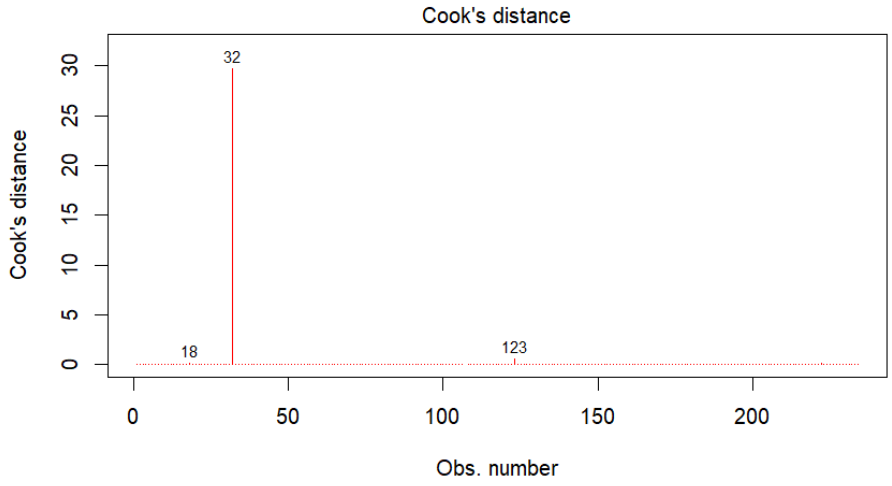


We can confirm this with a Shapiro-wilk test. In this case our null is that the data is normally distributed while our alternative is that it is not. As seen in Figure 7, the test statistic is 0.94836, and the p-value is found to be 2.147e-07, which is much less than the significance level. Therefore, we reject the null that the data is normally distributed and conclude that the assumption of normality is violated.

Lastly, we checked for any outliers, using a Residuals vs Leverage plot as shown below, and looking for any points that had a distance larger than 0.5. As seen in the plot we can see three observations that appear to be outliers, 32, 18, and 123.



Additionally, as seen in the plot below, we find the same points to be outliers again.

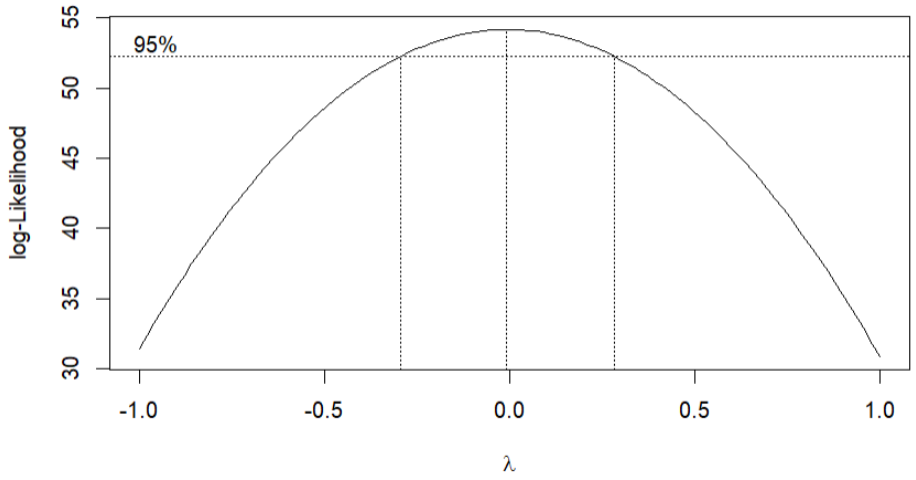


Therefore, we created a separate data frame without these points, and reran our interaction model, to ensure that all variables were still significant (Figure 8). We found that they were, so we proceeded to apply transformations to try to meet assumptions.

Note that we did not test for multicollinearity, as all our variables were categorical but one.

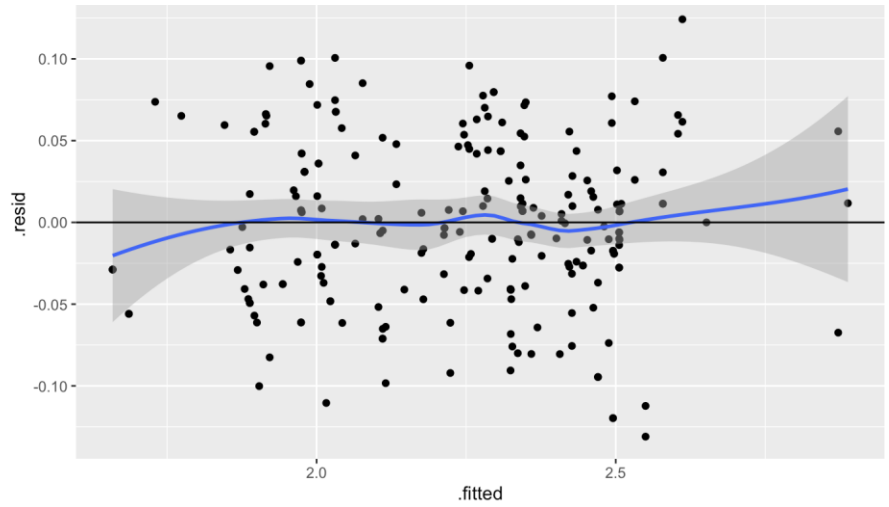
Transformation:

Due to the assumption of homoscedasticity not being met, we decided to proceed with a Box-Cox transformation. First, we created the plot below and determined what the best lambda was. From this figure, it appears to be around –0.1, but we confirmed and did indeed find it to be -0.01010101.



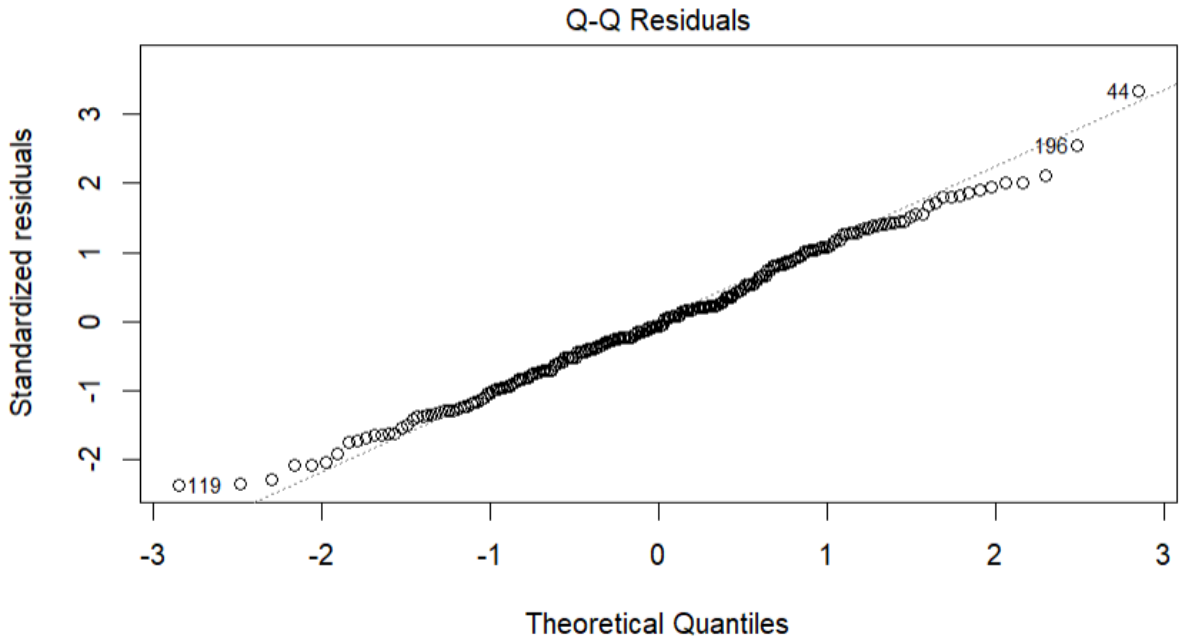
We then transformed the combined variable using this lambda and proceeded to refit the model. As seen in Figure 12, all predictors were found to be significant, we then continued to retest all assumptions.

First, we tested the assumption of linearity:



As seen in the plot above, the assumption of linearity does appear to still be violated though it has slightly improved. There does not appear to be as many fanning out patterns as was observed in the original model, but it is slightly noticeable.

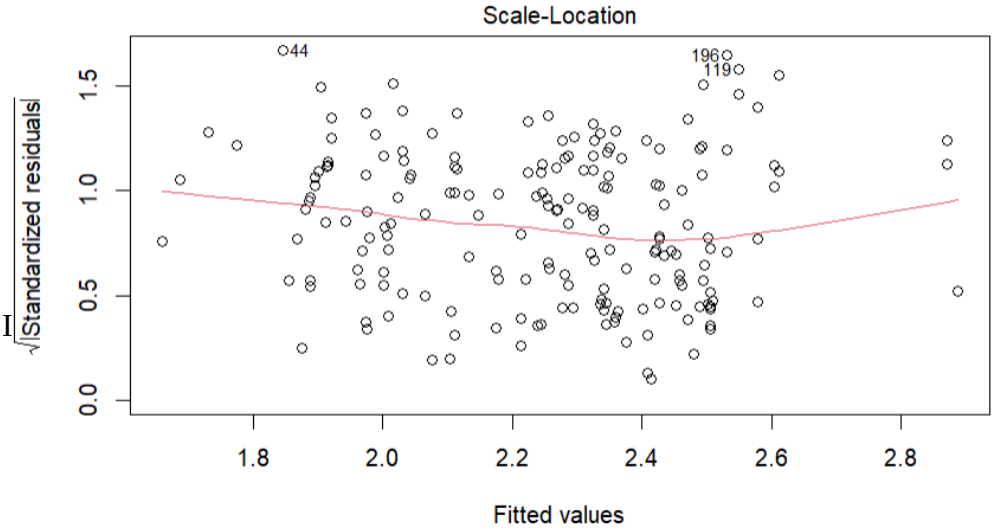
Next, we tested the assumption of normality. In this case we created a QQ plot and used the Shapiro-Wilk test once again. As seen in the plot below, normality seems to have improved.



While there are still minor deviations at the tails the points appear to align closer to the line indicating the data is now normally distributed. We can confirm with the Shapiro Wilk test. Once again, the null is that the data is normally distributed while the alternative is that it is not. As seen in Figure 9, the test statistic is 0.99423, and the p- value is 0.518, which is much greater

than the alpha of 0.05, this means that we fail to reject the null and can conclude the assumption of normality is met.

Lastly, we tested the assumption of homoscedasticity using a scale location plot, and the Breusch-Pagan test.



As seen in the plot above, the points appear to be randomly scattered, and no fan shape is observed indicating that homoscedasticity is met. We can confirm with our Breusch-Pagan test. Once again, the null and alternative are:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2$$
$$H_A : \sigma_1^2 \neq \sigma_2^2 \neq \cdots \neq \sigma_n^2$$

As seen in Figure 10, the test statistic is 39.475 and the p-value is 0.06967, based on this we fail to reject the null that the data has equal variance. This means that we can conclude our data is homoscedastic.

After our Box Cox transformation, all our assumptions but linearity are met. However, we did want to assess the degree of overfitting in the final model. Overfitting would occur if the model were good at predicting the training data and performs poorly on new unseen data (Chugani 2024). To do this, we performed a 10-fold cross validation, using the train function from the caret package. The technique involves dividing the data into 10 subsets, from there the model is trained on 9 of them. Essentially, the model learns the relationship between the predictors and the target variable using the training folds. Once training has occurred, the model is used to create predictions for the testing subset, predicted values are compared to actual values, and an R-squared is calculated. This process is repeated 10 times and then all R squared values are

averaged to provide a final value. The R squared from this process is then compared to the R squared obtained in the final model. If the difference between the two values is small, this indicates that the model is not overfitted, as it performs equally well on both the training data and unseen test data. As seen in Figure 11, the R-squared value from cross-validation was **0.9423**, while the R-squared value calculated from the training data was **0.9577**. As there is not a large gap between these two values, we see that the model performs consistently across both the training set and testing folds. This indicates that the model generalizes well to new data and does not exhibit overfitting.

Based on this we can now proceed with reporting our final model and interpreting it.

Final Model Overview

Our final model is the one that has had the Box Cox transformation applied and is as follows:

$$\hat{combined} = 3.2110 - 0.4107_{displ} - 0.1112_{cyl_5} - 0.3046_{cyl_6} - 0.7978_{cyl_8} + 0.2706_{drvf} + 0.4125_{drvr} + 0.3268_{fld} - 0.8067_{fle} - 0.1179_{flp} - 0.0476_{flr} - 0.5740_{classcompact} - 0.6992_{classmidsize} - 0.6612_{classminivan} - 0.4134_{classpickup} - 0.4213_{classsubcompact} - 0.3441_{classsuv} + 0.0795_{year2008} + 0.1218_{displ \cdot classcompact} + 0.1757_{displ \cdot classmidsize} + 0.1273_{displ \cdot classminivan} - 0.0004_{displ \cdot classpickup} + 0.0328_{displ \cdot classsubcompact} - 0.0094_{displ \cdot classsuv} + 0.0499_{displ \cdot cyl} - 0.0928_{displ \cdot drvf} - 0.0696_{displ \cdot drvr} + 0.0945_{displ \cdot fle} + 0.0005_{displ \cdot flp}$$

The model has an adjusted R-squared of 0.9518, meaning that the model explains 95.18% of the variance in the transformed dependent variable (Figure 12). Additionally, the RSE is 0.05439 meaning that on average, predictions differ from actual values by 0.05439 units. This indicates that our model is a good fit, both metrics are improved from both our first order model, and our interaction model.

The key takeaways and interpretations of our model are as follows: For every one-liter increase in engine displacement, combined mileage decreases by 0.4107 mpg, indicating that larger engines tend to have lower fuel efficiency. Similarly, vehicles with more cylinders generally have lower fuel efficiency compared to those with 4 cylinders (the reference group). Specifically, vehicles with 5, 6, and 8 cylinders are expected to have fuel efficiency 0.1112, 0.3046, and 0.7978 mpg lower, respectively, holding other factors constant.

Regarding drive type, vehicles with front-wheel drive (FWD) and rear-wheel drive (RWD) are expected to have fuel efficiencies 0.2706 and 0.4125 mpg higher, respectively, than vehicles with four-wheel drive (4WD), holding all else constant. This suggests that FWD and RWD vehicles are generally more fuel-efficient than 4WD vehicles.

Fuel type also influences fuel efficiency. Diesel vehicles are expected to have fuel efficiency 0.3268 mpg higher than gasoline vehicles, while ethanol and propane vehicles are expected to have fuel efficiencies 0.8067 and 0.1179 mpg lower than gasoline vehicles, respectively. This

indicates that diesel vehicles are more fuel-efficient, while ethanol and propane vehicles are less efficient.

When considering vehicle class, all vehicle types, except 2-seaters, are expected to have lower fuel efficiency. Specifically, compact cars, midsize cars, minivans, pickup trucks, subcompact cars, and SUVs are expected to have fuel efficiency reductions of 0.5740, 0.6992, 0.6612, 0.4134, 0.4213, and 0.3441 mpg, respectively, compared to 2-seaters. This suggests that larger vehicles generally have lower fuel efficiency. Lastly, vehicles manufactured in 2008 are expected to have fuel efficiency 0.0795 mpg higher than those from 1998, indicating a slight improvement in fuel efficiency over time.

In summary, larger engines (more cylinders and displacement), larger vehicles (vehicle class), and all-wheel drive (AWD) are associated with lower fuel efficiency, while front-wheel drive (FWD), rear-wheel drive (RWD), and diesel fuel tend to improve fuel efficiency.

CONCLUSION AND DISCUSSION

We met the objectives of our project by fitting a multiple linear regression model to the dataset, exploring the relationships between car attributes and fuel efficiency, incorporating interaction terms to capture data complexity, and testing assumptions. The final model provides a reliable tool for predicting fuel efficiency based on vehicle attributes, helping consumers make informed decisions when purchasing a car. This can promote more sustainable vehicle purchases, reduce household spending on fuel, and ultimately improve financial well-being.

By creating a multiple linear regression model with significant predictors and a high adjusted R-squared, we achieved our objective of analyzing the relationship between attributes and fuel efficiency. Additionally, we answered our research question on whether a regression model could effectively predict fuel efficiency using attributes, as evidenced by the high adjusted R-squared value. While we could not identify which attributes contribute the most due to the dataset's reliance on categorical variables, which prevented us from exploring correlations, we did identify how performance-related attributes trade off against fuel efficiency. Specifically, larger engines (measured by displacement) and more cylinders were associated with lower fuel efficiency, as these components require more power to operate. Additionally, larger vehicles, such as SUVs and minivans, tend to have lower fuel efficiency due to their size. The model also reveals that performance-related features, such as engine displacement and drivetrain type, trade off against fuel efficiency. For example, rear-wheel and all-wheel drive vehicles, which generally require more power, were found to be less fuel-efficient. This trade-off between performance and fuel efficiency is an important consideration for consumers, helping them balance performance needs with environmental sustainability.

Despite the model’s effectiveness, there are opportunities for improvement in both the dataset and the statistical approach. Increasing the sample size beyond 235 observations would improve predictive accuracy, as a larger sample reduces bias and provides more precise and realistic

predictions. Moreover, the dataset's limited number of numerical variables posed challenges in testing for correlations. Expanding the dataset with more continuous variables would strengthen the model's reliability. Additionally, using nonlinear regression could provide greater flexibility to model curved data, potentially leading to more accurate coefficient predictions and better fuel efficiency forecasts

REFERENCES

Harland, S., & Smith, T. (2021). New analysis finds most Canadian households will save money in switch to electricity.

Canadian Climate Institute. https://climateinstitute.ca/new-analysisfinds-most-canadian-households-will-save-money-in-switch-toelectricity/#:~:text=Consider%20how%20much%20you%20spend,year%20for%20the%20sa me%20mileage

Chugani, V. (2024, June 21). *A comprehensive guide to K-fold cross validation*. DataCamp. https://www.datacamp.com/tutorial/k-fold-cross-validation

Lethbridge Toyota. (2023, February 22). The difference between FWD and AWD. Lethbridge Toyota. https://www.lethbridgetoyota.com/the-difference-between-fwd-and-awd/

Toyota Canada. (2021). V6 vs V8 engine. Toyota Canada. https://www.toyota.ca/toyota/en/connect/3760/v6-vs-v8-engine

tidyverse. (2024). ggplot2/data-raw/mpg.csv at e594b49fdd5e4d95bf1031edaf6c7ccfc0cdedb0 · tidyverse/ggplot2. Retrieved November 9, 2024, from GitHub website: https://github.com/tidyverse/ggplot2/blob/e594b49fdd5e4d95bf1031edaf6c7ccfc0cdedb0/data-raw/mpg.csv

U.S. Environmental Protection Agency. (2021). What if we drove our most efficient car?. U.S. Environmental Protection Agency. Retrieved November 10, 2024, from https://www.epa.gov/greenvehicles/what-if-we-drove-our-most-efficient-car

APPENDIX

Figure 1

```
## Stepwise Selection Method
## ------------------------
##
## Candidate Terms:
##
## 1. displ
## 2. cyl
## 3. trans
## 4. drv
## 5. fl
## 6. class
## 7. year
## 8. region
##
##
## Step   => 0
## Model  => combined ~ 1
## R2     => 0
##
## Initiating stepwise selection...
##
## Step      => 1
## Selected  => cyl
## Model     => combined ~ cyl
## R2        => 0.626
##
## Step      => 2
## Selected  => drv
## Model     => combined ~ cyl + drv
## R2        => 0.752
##
## Step      => 3
## Selected  => fl
## Model     => combined ~ cyl + drv + fl
## R2        => 0.831
```



```
##
## Step      => 4
## Selected  => year
## Model     => combined ~ cyl + drv + fl + year
## R2        => 0.844
##
## Step      => 5
## Selected  => class
## Model     => combined ~ cyl + drv + fl + year + class
## R2        => 0.896
##
## Step      => 6
## Selected  => displ
## Model     => combined ~ cyl + drv + fl + year + class + displ
## R2        => 0.901
##
##
## No more variables to be added or removed.
```

Figure 2

```
##
## Call:
## lm(formula = combined ~ displ + factor(cyl) + drv + fl + class +
##     factor(year) + factor(region) + trans, data = mpg_cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5551 -0.3624 -0.0111  0.4230  2.8945
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              16.27108    1.22687  13.262  < 2e-16 ***
## displ                    -0.68041    0.16983  -4.006 8.63e-05 ***
## factor(cyl)5             -0.48128    0.46795  -1.028 0.304932
## factor(cyl)6             -1.05520    0.21783  -4.844 2.51e-06 ***
## factor(cyl)8             -1.10442    0.41809  -2.642 0.008889 **
## drvf                      1.06705    0.20906   5.104 7.56e-07 ***
## drvr                      0.46250    0.25109   1.842 0.066925 .
## fld                       2.74071    0.89109   3.076 0.002387 **
## fle                      -3.72772    0.86388  -4.315 2.48e-05 ***
## flp                      -1.72988    0.83659  -2.068 0.039916 *
## flr                      -1.52260    0.81635  -1.865 0.063594 .
## classcompact             -1.64111    0.51226  -3.204 0.001573 **
## classmidsize             -1.78504    0.49766  -3.587 0.000418 ***
```



```
## classminivan             -3.02291    0.57913  -5.220 4.38e-07 ***
## classpickup              -3.32639    0.49972  -6.656 2.52e-10 ***
## classsubcompact          -1.76262    0.49560  -3.557 0.000467 ***
## classsuv                 -3.12872    0.47017  -6.655 2.54e-10 ***
## factor(year)2008          0.66864    0.14506   4.609 7.10e-06 ***
## factor(region)Asian      -0.04426    0.19488  -0.227 0.820561
## factor(region)European   -0.47354    0.28169  -1.681 0.094277 .
## transauto(l3)            -0.18254    0.72385  -0.252 0.801158
## transauto(l4)            -0.23424    0.41027  -0.571 0.568658
## transauto(l5)            -0.27516    0.40532  -0.679 0.497990
## transauto(l6)            -0.33462    0.52203  -0.641 0.522243
## transauto(s4)            -0.30348    0.60269  -0.504 0.615117
## transauto(s5)            -0.12875    0.58987  -0.218 0.827444
## transauto(s6)            -0.12099    0.42713  -0.283 0.777254
## transmanual(m5)           0.08055    0.40700   0.198 0.843314
## transmanual(m6)          -0.37185    0.41103  -0.905 0.366694
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7837 on 205 degrees of freedom
## Multiple R-squared:  0.9123, Adjusted R-squared:  0.9004
## F-statistic: 76.19 on 28 and 205 DF,  p-value: < 2.2e-16
```

Figure 3

```
##
## Call:
## lm(formula = combined ~ (displ + factor(cyl) + drv + fl + class +
##     factor(year)), data = mpg_cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3355 -0.3526 -0.0453  0.4360  2.8627
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       15.7793     1.0371  15.215  < 2e-16 ***
## displ             -0.5743     0.1487  -3.862 0.000149 ***
## factor(cyl)5      -0.7893     0.4217  -1.871 0.062633 .
## factor(cyl)6      -1.1993     0.2057  -5.831 1.99e-08 ***
## factor(cyl)8      -1.4512     0.3902  -3.719 0.000255 ***
## drvf               1.0645     0.1994   5.339 2.36e-07 ***
## drvr               0.4868     0.2377   2.048 0.041758 *
## fld                2.6555     0.8725   3.044 0.002628 **
## fle               -3.6000     0.8548  -4.211 3.72e-05 ***
## flp               -1.8439     0.8166  -2.258 0.024940 *
## flr               -1.3974     0.8054  -1.735 0.084142 .
## classcompact      -1.7561     0.4772  -3.680 0.000295 ***
## classmidsize      -1.8105     0.4750  -3.811 0.000180 ***
## classminivan      -3.1158     0.5265  -5.917 1.27e-08 ***
## classpickup       -3.3563     0.4802  -6.989 3.38e-11 ***
## classsubcompact   -1.7460     0.4778  -3.654 0.000324 ***
## classsuv          -3.1880     0.4461  -7.147 1.35e-11 ***
## factor(year)2008   0.6358     0.1103   5.764 2.82e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7851 on 216 degrees of freedom
## Multiple R-squared:  0.9073, Adjusted R-squared:    0.9
## F-statistic: 124.3 on 17 and 216 DF,  p-value: < 2.2e-16
```

Figure 4

```
##
## Call:
## lm(formula = combined ~ displ + factor(cyl) + drv + fl + class +
##     factor(year) + displ:class + displ:cyl + displ:drv + displ:fl,
##     data = mpg_cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3557 -0.3180  0.0028  0.3357  2.1553
##
```



```
## Coefficients: (1 not defined because of singularities)
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           19.05099    4.43275   4.298 2.67e-05 ***
## displ                 -3.98695    1.17762  -3.386 0.000852 ***
## factor(cyl)5          -1.29194    0.41431  -3.118 0.002082 **
## factor(cyl)6          -3.35958    0.55637  -6.038 7.25e-09 ***
## factor(cyl)8          -8.36282    1.62590  -5.144 6.31e-07 ***
## drvf                   3.11391    0.72065   4.321 2.43e-05 ***
## drvr                   5.20115    1.61329   3.224 0.001472 **
## fld                    6.00084    0.90290   6.646 2.69e-10 ***
## fle                   -7.11239    2.10819  -3.374 0.000888 ***
## flp                   -1.94480    0.75683  -2.570 0.010892 *
## flr                   -0.92660    0.68534  -1.352 0.177863
## classcompact          -4.40437    4.35047  -1.012 0.312551
## classmidsize          -5.99757    4.32932  -1.385 0.167462
## classminivan          -6.63726    4.57071  -1.452 0.148002
## classpickup           -3.90235    4.29844  -0.908 0.365028
## classsubcompact       -3.11408    4.34788  -0.716 0.474669
## classsuv              -3.39855    4.23652  -0.802 0.423369
## factor(year)2008       0.70006    0.09643   7.260 7.99e-12 ***
## displ:classcompact     0.89824    0.78934   1.138 0.256468
## displ:classmidsize     1.53760    0.76746   2.003 0.046448 *
## displ:classminivan     1.33982    0.88316   1.517 0.130797
## displ:classpickup      0.08966    0.72612   0.123 0.901850
## displ:classsubcompact  0.12790    0.76323   0.168 0.867085
## displ:classsuv         0.01708    0.70518   0.024 0.980703
## displ:cyl              0.50850    0.12572   4.045 7.43e-05 ***
## displ:drvf            -0.98097    0.26952  -3.640 0.000346 ***
## displ:drvr            -0.92670    0.30984  -2.991 0.003124 **
## displ:fld             -1.08546    0.18281  -5.938 1.23e-08 ***
## displ:fle              0.85993    0.42370   2.030 0.043698 *
## displ:flp              0.10712    0.10844   0.988 0.324398
## displ:flr                   NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6607 on 204 degrees of freedom
## Multiple R-squared:  0.938,  Adjusted R-squared:  0.9292
## F-statistic: 106.4 on 29 and 204 DF,  p-value: < 2.2e-16
```

Figure 5

```
##
## Call:
## lm(formula = combined ~ displ + factor(cyl) + I(displ^2) + drv +
##     fl + class + factor(year) + displ:cyl + displ:drv + displ:fl,
##     data = mpg_cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4893 -0.3450 -0.0158  0.3780  2.2211
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
```



```
## (Intercept)       17.87704    1.54012  11.608  < 2e-16 ***
## displ             -4.44296    0.88316  -5.031 1.05e-06 ***
## factor(cyl)5      -2.28057    0.79307  -2.876 0.004450 **
## factor(cyl)6      -5.07568    1.59107  -3.190 0.001641 **
## factor(cyl)8     -12.88211    4.17485  -3.086 0.002306 **
## I(displ^2)        -0.22028    0.18784  -1.173 0.242237
## drvf               0.57856    0.55235   1.047 0.296107
## drvr               4.19678    1.55086   2.706 0.007370 **
## fld                5.75181    0.96241   5.976 9.72e-09 ***
## fle               -7.55657    2.21762  -3.408 0.000786 ***
## flp               -2.59698    0.78721  -3.299 0.001141 **
## flr               -1.18820    0.72016  -1.650 0.100463
## classcompact      -1.61476    0.58574  -2.757 0.006354 **
## classmidsize      -1.43022    0.59164  -2.417 0.016491 *
## classminivan      -2.58173    0.64762  -3.987 9.27e-05 ***
## classpickup       -3.14884    0.57175  -5.507 1.06e-07 ***
## classsubcompact   -1.90685    0.58808  -3.242 0.001379 **
## classsuv          -2.96548    0.53384  -5.555 8.39e-08 ***
## factor(year)2008   0.72012    0.10102   7.128 1.63e-11 ***
## displ:cyl          0.84539    0.29883   2.829 0.005125 **
## displ:drvf         0.03076    0.19582   0.157 0.875329
## displ:drvr        -0.76494    0.30256  -2.528 0.012203 *
## displ:fld         -1.03304    0.20616  -5.011 1.15e-06 ***
## displ:fle          0.88779    0.44622   1.990 0.047940 *
## displ:flp          0.26244    0.10871   2.414 0.016635 *
## displ:flr               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6976 on 209 degrees of freedom
## Multiple R-squared:  0.9292, Adjusted R-squared:  0.921
## F-statistic: 114.2 on 24 and 209 DF,  p-value: < 2.2e-16
```

Figure 6

```
##
##  studentized Breusch-Pagan test
##
## data:  inter_model_2
## BP = 87.414, df = 29, p-value = 9.055e-08
```

Figure 7

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(inter_model_2)
## W = 0.94836, p-value = 2.174e-07
```

Figure 8

```
##
## Call:
## lm(formula = combined ~ displ + factor(cyl) + drv + fl + class +
##     factor(year) + displ:class + displ:cyl + displ:drv + displ:fl,
##     data = mpg_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63711 -0.29208 -0.02782  0.32045  2.03281
##
## Coefficients: (2 not defined because of singularities)
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           19.69639    3.98364   4.944 1.60e-06 ***
## displ                 -4.32201    1.07607  -4.016 8.33e-05 ***
## factor(cyl)5          -1.12779    0.38103  -2.960 0.003446 **
## factor(cyl)6          -3.18207    0.51082  -6.229 2.67e-09 ***
## factor(cyl)8          -8.17226    1.50183  -5.442 1.52e-07 ***
## drvf                   3.76806    0.77012   4.893 2.03e-06 ***
## drvr                   4.93265    1.45142   3.399 0.000816 ***
```



```
## fld                    4.91399    0.69322   7.089 2.22e-11 ***
## fle                   -7.37975    1.88926  -3.906 0.000128 ***
## flp                   -2.05954    0.67913  -3.033 0.002742 **
## flr                   -0.87396    0.61442  -1.422 0.156448
## classcompact          -5.10565    3.90053  -1.309 0.192034
## classmidsize          -7.08835    3.89736  -1.819 0.070430 .
## classminivan          -7.59159    4.10681  -1.849 0.065987 .
## classpickup           -4.17148    3.85168  -1.083 0.280088
## classsubcompact       -3.61748    3.89603  -0.929 0.354254
## classsuv              -3.11897    3.79580  -0.822 0.412222
## factor(year)2008       0.80280    0.08779   9.144  < 2e-16 ***
## displ:classcompact     1.14834    0.71218   1.612 0.108428
## displ:classmidsize     1.97660    0.71128   2.779 0.005969 **
## displ:classminivan     1.73158    0.80644   2.147 0.032970 *
## displ:classpickup      0.14603    0.65076   0.224 0.822677
## displ:classsubcompact  0.22928    0.68392   0.335 0.737787
## displ:classsuv        -0.04291    0.63186  -0.068 0.945926
## displ:cyl              0.52682    0.11596   4.543 9.53e-06 ***
## displ:drvf            -1.31998    0.30166  -4.376 1.94e-05 ***
## displ:drvr            -0.85914    0.27858  -3.084 0.002328 **
## displ:fld                  NA         NA      NA       NA
## displ:fle              0.91185    0.37966   2.402 0.017222 *
## displ:flp              0.14452    0.09734   1.485 0.139168
## displ:flr                  NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5919 on 202 degrees of freedom
## Multiple R-squared:  0.9505, Adjusted R-squared:  0.9437
## F-statistic: 138.6 on 28 and 202 DF,  p-value: < 2.2e-16
```

Figure 9

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(bcmodel)
## W = 0.99423, p-value = 0.5218
```

Figure 10

```
## data:  bcmodel
## BP = 39.745, df = 28, p-value = 0.06967
```

Figure 11

```
231 samples
  6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 210, 209, 207, 208, 208, 207, ...
Resampling results:

  RMSE        Rsquared   MAE
  0.06136919  0.9422629  0.047861

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Figure 12:

```
## Call:
## lm(formula = (combined^(-0.01010101) - 1)/-0.01010101 ~ displ +
##     factor(cyl) + drv + fl + factor(year) + class + displ:cyl +
##     displ:drv + displ:fl + displ:class, data = mpg_2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.131026 -0.033528 -0.002506  0.038465  0.141077
##
## Coefficients: (2 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.2110129  0.3660735   8.771 7.36e-16 ***
## displ                  -0.4106544  0.0988843  -4.153 4.84e-05 ***
## factor(cyl)5           -0.1112343  0.0350142  -3.177 0.001722 **
## factor(cyl)6           -0.3046398  0.0469410  -6.490 6.51e-10 ***
## factor(cyl)8           -0.7977891  0.1380096  -5.781 2.79e-08 ***
## drvf                    0.2706251  0.0707697   3.824 0.000175 ***
## drvr                    0.4124841  0.1333767   3.093 0.002264 **
## fld                     0.3268121  0.0637029   5.130 6.77e-07 ***
## fle                    -0.8067016  0.1736117  -4.647 6.08e-06 ***
## flp                    -0.1178510  0.0624077  -1.888 0.060405 .
## flr                    -0.0476110  0.0564617  -0.843 0.400090
## factor(year)2008        0.0795397  0.0080678   9.859  < 2e-16 ***
## classcompact           -0.5740417  0.3584361  -1.602 0.110825
## classmidsize           -0.6991927  0.3581453  -1.952 0.052289 .
## classminivan           -0.6612070  0.3773925  -1.752 0.081284 .
## classpickup            -0.4134232  0.3539475  -1.168 0.244169
## classsubcompact        -0.4213127  0.3580229  -1.177 0.240670
```



```
## classsuv               -0.3440513  0.3488119  -0.986 0.325140
## displ:cyl               0.0498515  0.0106564   4.678 5.30e-06 ***
## displ:drvf             -0.0927888  0.0277208  -3.347 0.000973 ***
## displ:drvr             -0.0695774  0.0255994  -2.718 0.007140 **
## displ:fld                     NA         NA      NA       NA
## displ:fle               0.0945103  0.0348886   2.709 0.007330 **
## displ:flp               0.0004842  0.0089446   0.054 0.956879
## displ:flr                     NA         NA      NA       NA
## displ:classcompact      0.1217622  0.0654450   1.861 0.064264 .
## displ:classmidsize      0.1756686  0.0653629   2.688 0.007797 **
## displ:classminivan      0.1272544  0.0741075   1.717 0.087483 .
## displ:classpickup      -0.0004399  0.0598011  -0.007 0.994138
## displ:classsubcompact   0.0328277  0.0628479   0.522 0.602009
## displ:classsuv         -0.0094364  0.0580647  -0.163 0.871062
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05439 on 202 degrees of freedom
## Multiple R-squared:  0.9577, Adjusted R-squared:  0.9518
## F-statistic: 163.2 on 28 and 202 DF,  p-value: < 2.2e-16
```