Climate vs Cars

Aparna Gupta

University of Arizona, DATA - 375

Author Note

Aparna Gupta – guptaa@email.arizona.edu

Contents

Climate vs Cars 4

Introduction 4

Methods 5

a. Make 5

c. Vehicle Class 5

e. Transmission Type 5

g. Fuel Consumption 5

h. CO2 Emissions 5

Engine size vs CO2 emissions 6

Fuel Type vs CO2 emissions 7

Engine Size vs Fuel Type 7

Results and Discussion of analysis 8

Basic Analysis 11

Larger Model 11

(CO2 emission ~ engine size + cylinders + fuel consumption + co2 rating + smog rating + make + class + fuel type) 11

QQ-Plot interpretation 11

Residual-Plot interpretation 12

Interpretation of Coefficients 12

Hypothesis Test 13

Reduced Model 14

(CO2 emission ~ fuel consumption + co2 rating + class + fuel type) 14

QQ-Plot interpretation 14

Residual Plot Interpretation 15

Interpretation of Coefficients 15

Hypothesis Test 16

Comparing larger and smaller model 16

Advanced Analysis 17

Hypothesis Testing 1 17

Hypothesis Testing 2 19

Conclusions 20

References 21

Climate vs Cars

Introduction

With reports of severe climate changes from around the world, it is high time to keep a check on our carbon footprint and sustainability practices. Everyday car use can be one of the greatest contributors to this. However, we can't stop everyday car use, but we can identify the factors responsible for higher CO2 emissions and minimize them. In this project, we are planning on finding the different factors that can minimize carbon emissions when regulated. Worldwide, Passenger cars produced approximately three billion metric tons of carbon dioxide in 2020 alone and the rising cases of global warming due to heat rise is one of the direct consequences. Other consequences that arise due to CO2 emissions from cars are air and noise pollution. The greenhouse gases make the environment unhealthy to sustain and rising heat levels also affect people's bodies and their lives which are sometimes dependent on the weather. We intend to model and see the relationship between multiple predictor variables, which predicts and analyzes what are the optimal conditions like fuel type and engine size that led to minimizing the CO2 emissions. In our initial analysis, we modeled fuel type and engine size with the CO2 emissions and realized that there is a bias, which deviates from a linear behavior since larger engine cars usually sport cars only use premium gasoline.

Methods

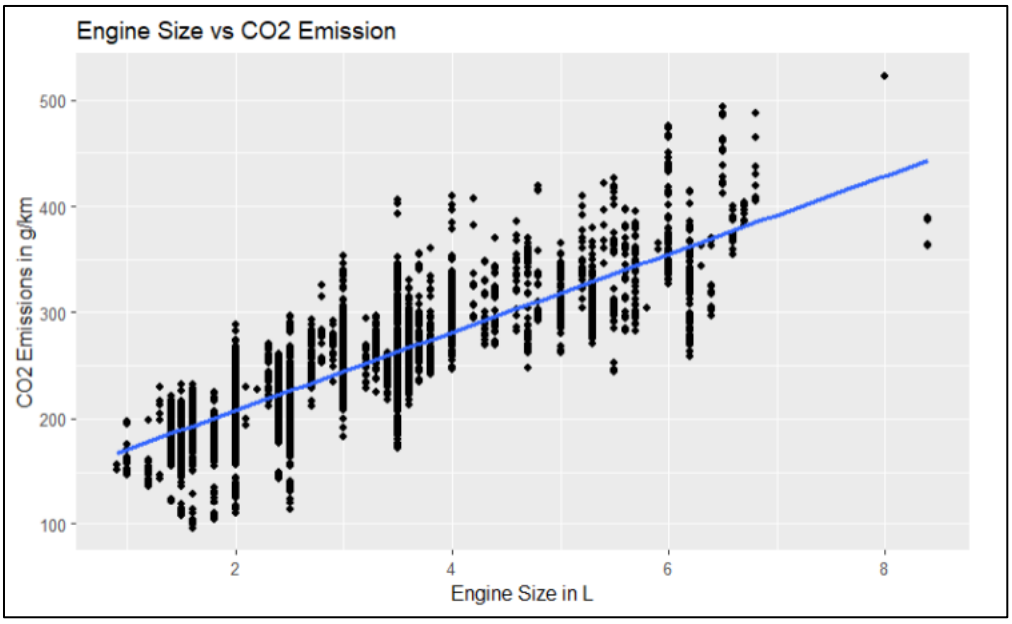
The Dataset belongs to the Canadian Government and their traffic department. They sampled multiple cars and their fuel efficiency over 7 years from 2010 to 2017. Over the course of 7 years, 12 different factors were observed from 7305 different vehicles.

The observed categories in the dataset are:

a. **Make** which includes the company that manufactured the vehicle.
b. The **Model** column contains the model of the vehicle from the specific company.
c. **Vehicle Class** contains the size class of the vehicle like whether it is an SUV, Compact, etc.
d. The **Engine size** stores the size of the vehicle's engine in liters. Cylinders reflect the number of cylinders in the vehicle's engine.
e. **Transmission Type** contains the type of transmission for that vehicle.
f. The **Fuel Type** is a categorical variable and stores the information about the fuel type consumed by the vehicle as X = Regular Gasoline, Z = Premium Gasoline, D = Diesel, E = Ethanol, N = Natural gas.
g. **Fuel Consumption** is of 4 different types. There are two columns which measure the consumption in highway and city measured in liters per 1000 km. The rest two are combined data from city and highway with two different units that is miles per gallon and liters per 100 kilometers.
h. **CO2 Emissions** is the most important variable for our study as we later that this as our response variable. This is collected from Tailpipe emissions of Carbon Dioxide in gm/km

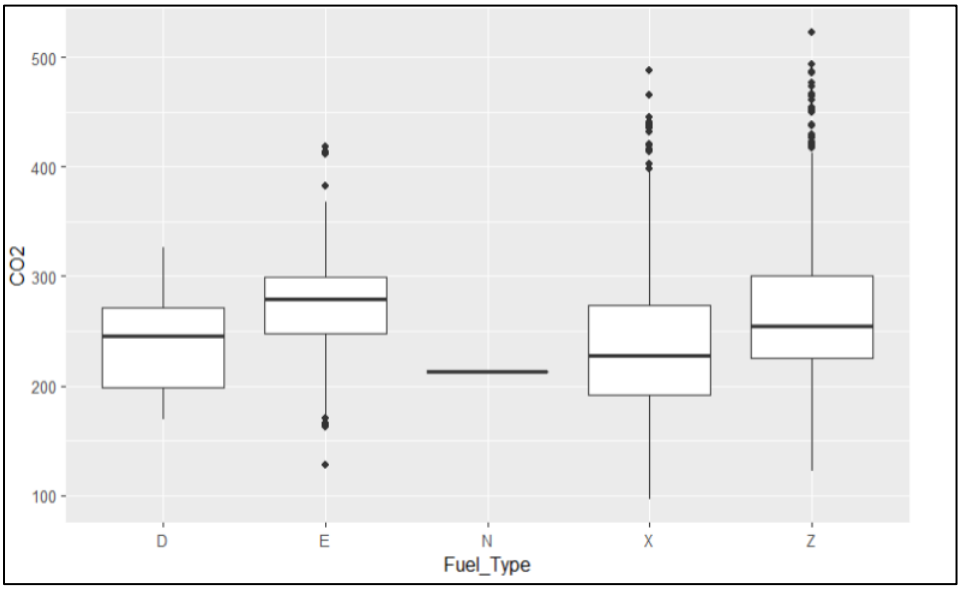
In the dataset, the Categorical variables are Vehicle class, Model, Make, Transmission, Fuel Type, Cylinders, and Quantitative variables are CO2 Emissions, Fuel Consumption, Engine Size. In the dataset, the Response Variable would be CO2 emissions. The researcher checked the trend of the different predictor variables with the response variable by plotting a scatter plot.

Engine size vs CO2 emissions



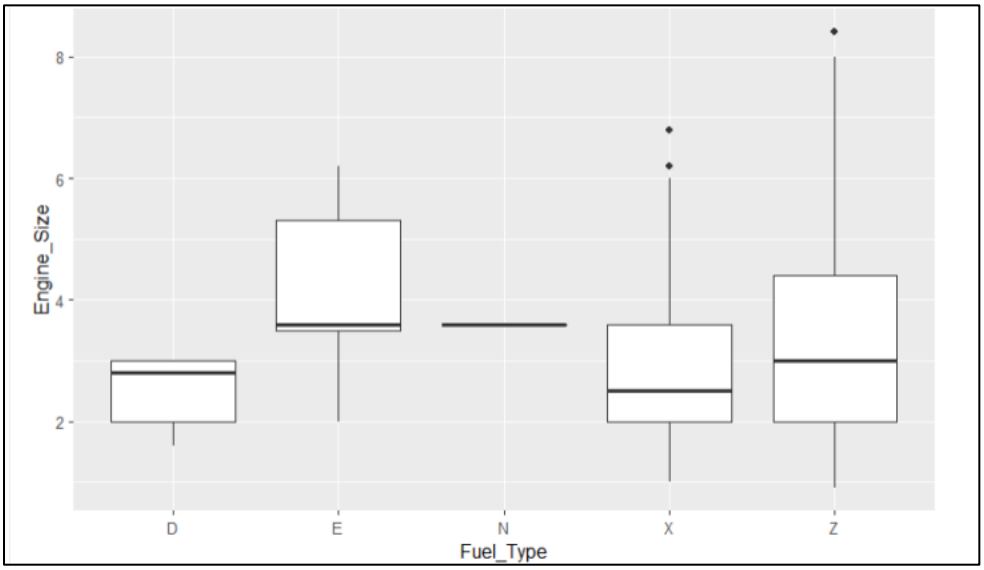
From the scatter plot above we can see that Engine Size has a linear relationship with CO2 emissions. This means that as engine size increases in Liters the CO2 emission of the particular vehicle rises as well. We can also see that there are also a lot of outliers. To determine the reason for bias we checked the relationship between fuel type and Co2 emissions.

Fuel Type vs CO2 emissions



From the box blot above we can see the 5 quantiles of the 5 different fuel types when plotted against the CO2 emissions. Here Fuel type of premium gasoline surprisingly has greater outliers and Natural gas has none. This could raise a bias in the study but can be answered through further analysis.

Engine Size vs Fuel Type

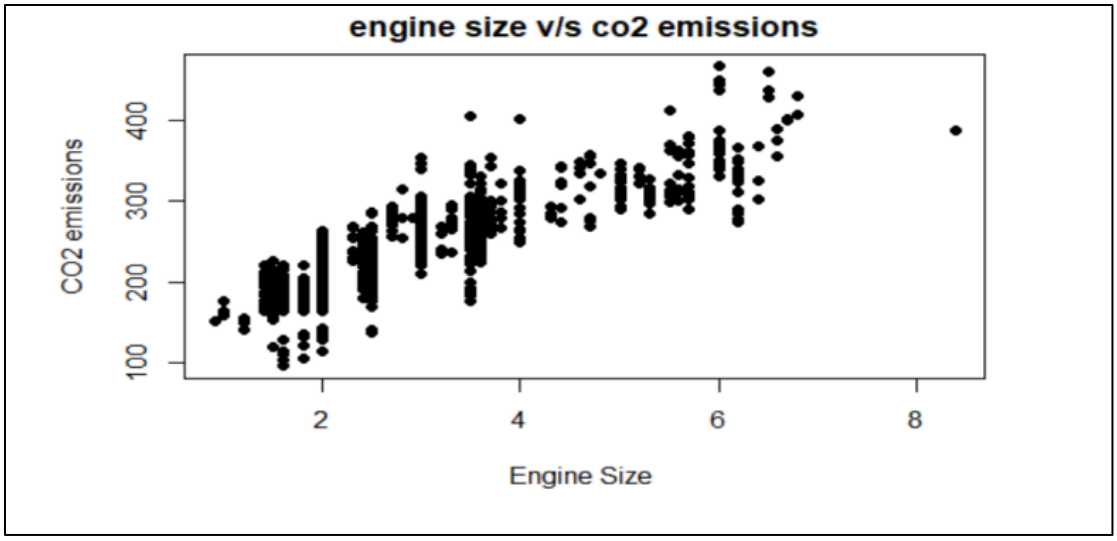


Now plotting the Engine Size and Fuel type in a box plot we can see the reason for the bias observed earlier. As we can see from the plot the reason for premium gasoline has so many outliers as it is also used in vehicles with bigger engines and since engine size has a linear trend with CO2 emissions it can explain the outliers in premium gasoline fuel type. To further study the dataset we will be looking at the fuel consumption predictor value and try to fit all these predictions in a model to see which features of a car lead to higher CO2 emissions.

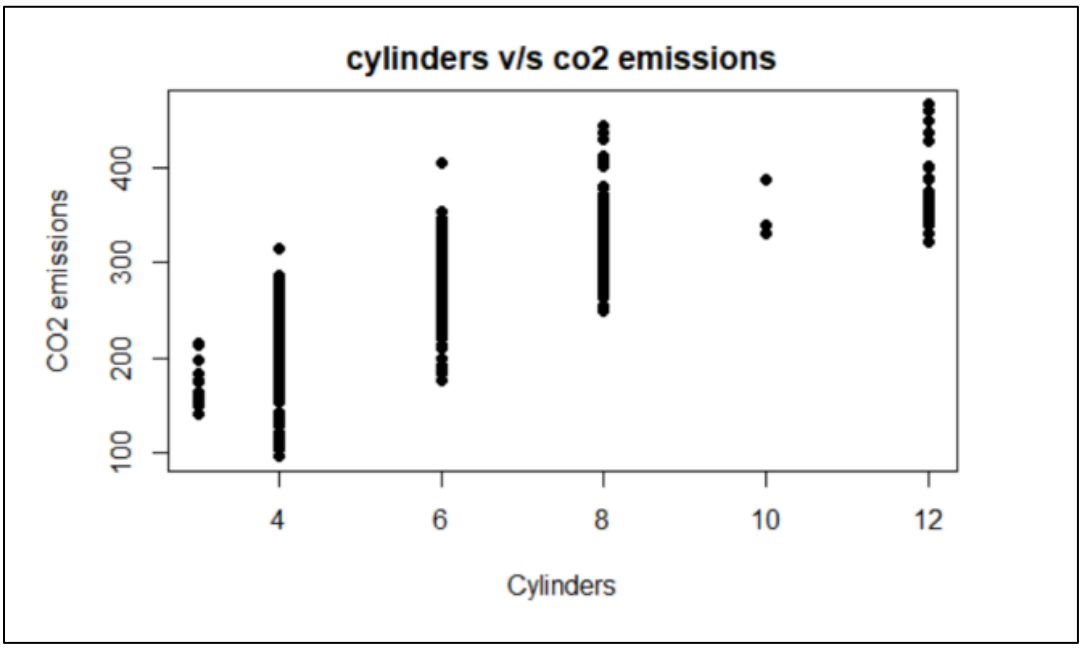
Results and Discussion of analysis

Our data set, the Canadian government's official data on the complete information about cars- number of cylinders, make, model, year, and many more, with the most influential column for our analyses, the CO2 emissions for each car over the span of years. In our initial analysis, I plotted the relationship between different components against the CO2 emissions, sorted by their categories.

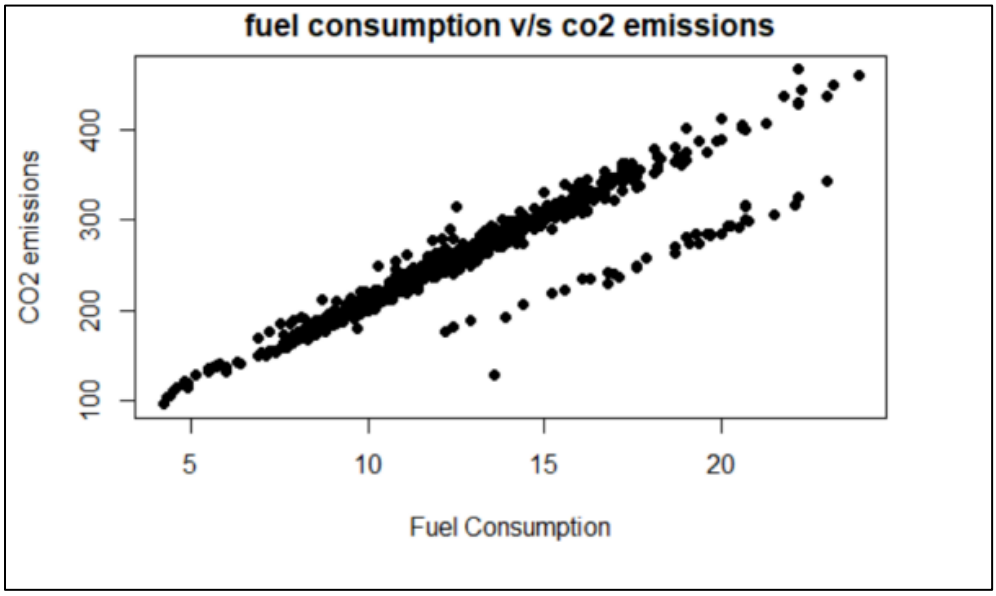
1. The first scatterplot represents the relationship between the engine size and the level of CO2 emissions by the corresponding size of the engine in the car. The scatterplot clearly shows a strong positive correlation that is close to 1, however we can also see some outliers. The residue for some data points are scattered a bit more than others in their engine size category. This behavior can be explained by the new models of cars that some companies have made which are really fuel efficient while maintaining a lower CO2 emissions. The other case for outliers can be explained by the older cars in the same category that generally produced more emissions, as the newer engines have developed to emit lesser over time.



2. The second scatterplot is for the relationship between number of cylinders and emissions. We can see a clear trend that is rising and shows a strong positive correlation. There are some outliers that are again explainable by the fact that older cars had lower efficiency with the same number of cylinders, so they emit more even for the same number of cylinders.



3. The third one is relating the fuel consumption and the CO2 emissions. We can see that there is a very strong (almost 1) correlation between these variables. At a time, after the fuel consumption reaches 12ish, there is a split, which in residuals, which shows two positive streams of predictor variables. This is understandable as a clear differentiation between the SUV models and expensive, models that consume more and emit more, whereas smaller cars that consume more and emit lesser into the environment.



All in all, these are very promising results as we can see very strong positive correlations among the variables that are clearly indicative of our initial assumptions being right. Excited to see where we can find abnormalities that are not easily thinkable by simple logic.

Basic Analysis

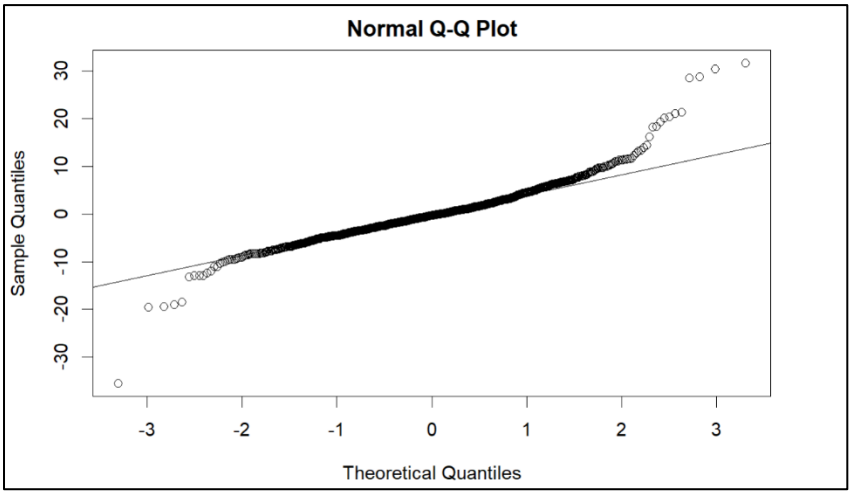
Larger Model

(CO2 emission ~ engine size + cylinders + fuel consumption + co2 rating + smog rating + make + class + fuel type)

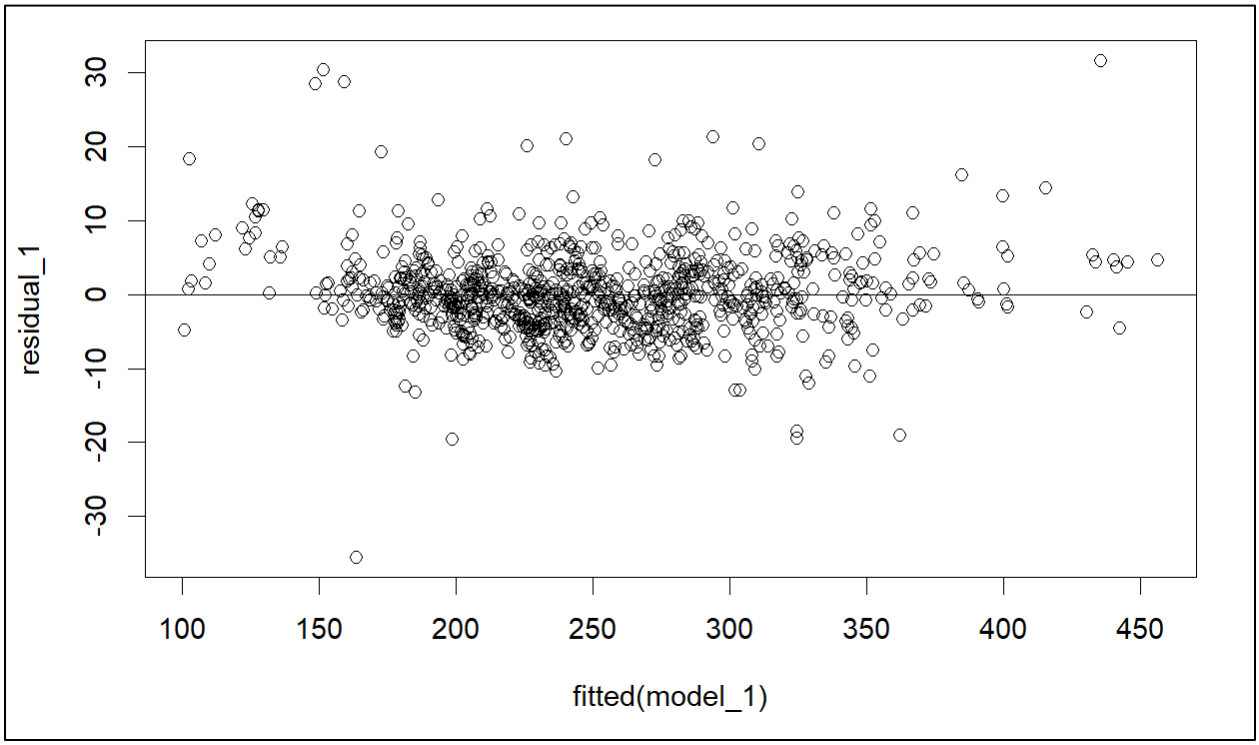
The first model that the researcher studied used CO2 emissions as the response variable and used engine size, cylinders, fuel consumption, CO2 and smog rating of the vehicle as the quantitative variables and used vehicle class, vehicle make, and fuel type as the categorical variables. The researcher used dummy variables for class, make, and fuel type to see the effect of different vehicles from different brands with different classes and engine types on the CO2 emissions made by the vehicle.

QQ-Plot interpretation

The QQ- plot for the residuals shows a heavy-tailed distribution, which means the probability of larger numbers is much more likely than normal distribution.



Residual-Plot interpretation



The residual plot is normal with low number of outliers which can be explained due to the model year and the transmission type of the car.

Interpretation of Coefficients

Engine Size = As the size of engine increases by 1 liter, the CO2 emissions decrease by 0.3094 gm/km.

Cylinders = as the number of cylinders increase by 1, the CO2 emissions increase by 0.5508 gm/km.

Fuel Consumption = As the fuel consumption in city increases by 1 liters/1000 km , the CO2 emission increase by 15.3203 gm/km.

CO2 rating = As CO2 rating increases by 1, the CO2 emission decreases by 5.2706 gm/km.

Smog Rating = As smog rating increases by 1, the CO2 emission increase by 0.4374 gm/km.

Make = The change in CO2 is measured as the coefficient higher/lower than the variable used to make the dummy for the make class which is volvo.

Example = Acura has a CO2 emission of 5.8282 gm/km higher than volvo, on average and all else constant.

Class = The change in CO2 is measured as the coefficient higher/lower than the variable used to make the dummy for the class which is passenger van.

Example = compact class has a CO2 emission of 22.8364 gm/km lower than passenger van, on average and all else constant.

Fuel Type = The change in CO2 is measured as the coefficient higher/lower than the variable used to make the dummy for the fuel type which is type d.

Example = Fuel Type Z has a CO2 emission of 20.244 gm/km lower than type D, on average and all else constant.

Hypothesis Test

$H_{A:}$ At least one variable out of engine size, cylinders, fuel consumption, CO2 rating, smog rating, make, class, and fuel type are significant predictors of CO2 emission.

$H_{0:}$ None of the variables out of engine size, cylinders, fuel consumption, CO2 rating, smog rating, make, class, and fuel type are significant predictors of CO2 emission.

Test Statistic = t-distribution = 1.646

p-value = 2.2e-16

df = 997

Here, we reject the null and prove the alternative hypothesis which states that at least one variable out of engine size, cylinders, fuel consumption, CO2 rating, smog rating, make, class, and fuel type are significant predictors of CO2 emission. Here we see that engine size, cylinders, make, and smog rating are not significant predictors of CO2 emissions as their p-values are higher than 0.05. We reject the null hypothesis as the p-value (2.2e-16) is less than our 0.05 significance level, the p-value being almost equivalent to 0. This implies that we are almost about 100% confident in rejecting the null and proving our claim.

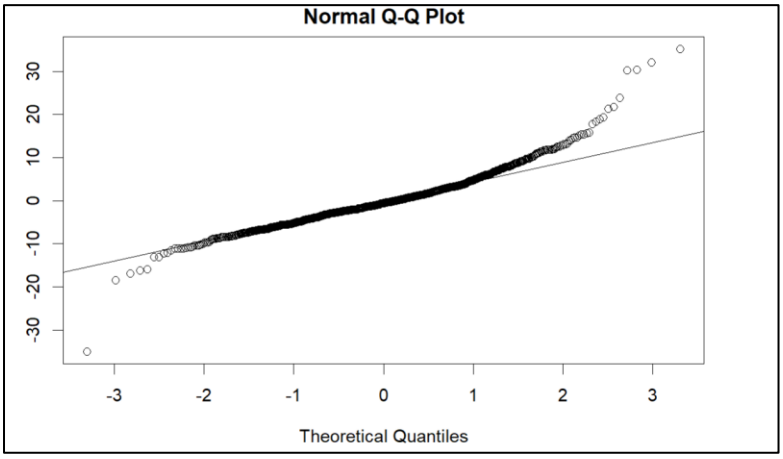
Reduced Model

(CO2 emission ~ fuel consumption + co2 rating + class + fuel type)

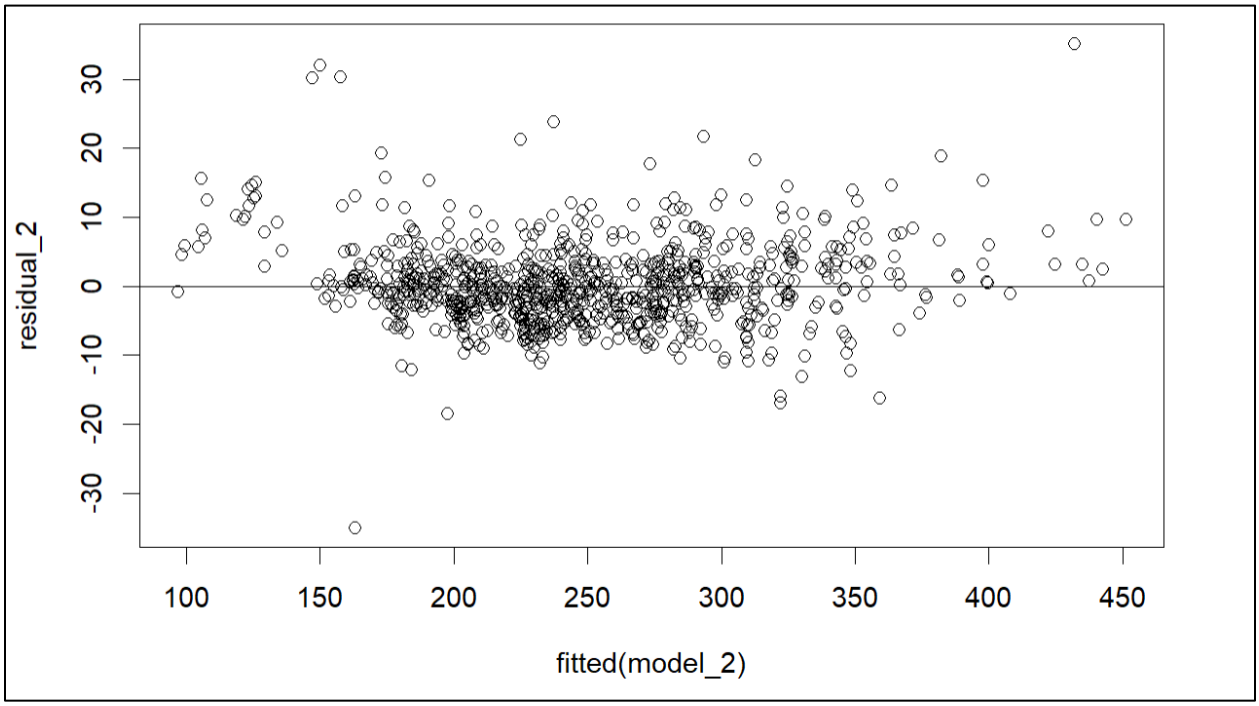
The reduced model that the researcher studied used CO2 emissions as the response variable and used fuel consumption and CO2 rating of the vehicle as the quantitative variables and used vehicle class and fuel type as the categorical variables. The researcher used dummy variables for class and fuel type to see the effect of different vehicles with different classes and engine types on the CO2 emissions made by the vehicle.

QQ-Plot interpretation

The QQ- plot for the residuals shows a heavy-tailed distribution, which means the probability of larger numbers is much more likely than normal distribution.



Residual Plot Interpretation



The residual plot is normal with low number of outliers which can be explained due to the model year of the car as older cars would have lower fuel efficiency irrespective of the class, fuel type, CO2 rating, and fuel consumption of the car.

Interpretation of Coefficients

Fuel consumption: As the fuel consumption in city increases by 1 liters/1000 km , the CO2 emission increase by 15.5385 gm/km.

CO2 rating: As CO2 rating increases by 1, the CO2 emission decreases by 5.0428 gm/km.

Class = The change in CO2 is measured as the coefficient higher/lower than the variable used to make the dummy for the class which is passenger van.

Example = compact class has a CO2 emission of 19.2677 gm/km lower than passenger van, on average and all else constant.

Fuel Type = The change in CO2 is measured as the coefficient higher/lower than the variable used to make the dummy for the fuel type which is type d.

Example = Fuel Type Z has a CO2 emission of 18.6209 gm/km lower than type D, on average and all else constant.

Hypothesis Test

H_A : At least one variable out of fuel consumption, CO2 rating, class, and fuel type are significant predictors of CO2 emission.

H_0 : None of the variables out of out of fuel consumption, CO2 rating, class, and fuel type are significant predictors of CO2 emission.

Test Statistic = t-distribution = 1.646

p-value = 2.2e-16

df = 1038

Here, we reject the null and prove the alternative hypothesis which states that at least one variable out of fuel consumption, CO2 rating, class, and fuel type are significant predictors of CO2 emission. We reject the null hypothesis as the p-value (2.2e-16) is less than our 0.05 significance level, the p-value being almost equivalent to 0. This implies that we are almost about 100% confident in rejecting the null and proving our claim.

Comparing larger and smaller model

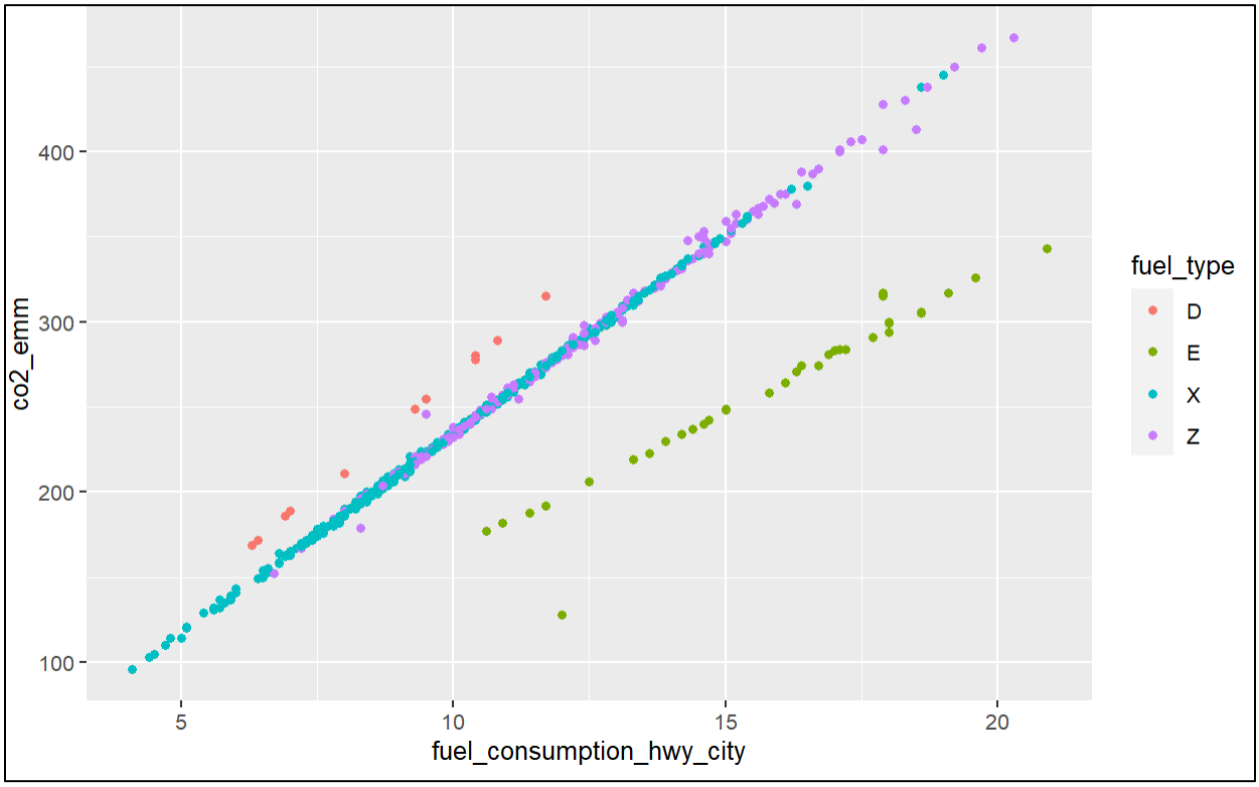
H_0: $\sigma^2(Full) < \sigma^2\ (Reduced)$

HA: $\sigma^2(Full) > \sigma^2\ (Reduced)$

F-statistic test of 4.4617 on 41 df with p = 2.2e−16

At a 95% confidence interval, we fail to reject the null hypothesis, which implies that our reduced model is a better fit to the data than our larger model.

Advanced Analysis



The researcher found an interesting pattern of an outlier band while observing the fuel consumption and the CO2 emissions against each other. On comparing the fuel consumption relative to the different fuel types, we observed that for a higher fuel consumption of fuel type E, the CO2 emissions were lower and for a lower consumption of fuel type D, the CO2 emissions were higher. Due to the correlation between fuel type and the CO2 emissions relative to the fuel consumption, the researcher had reason to believe that mean CO2 emission of different fuel type cars are different from each other. The researcher conducted the following hypothesis test to determine this relationship:

Hypothesis Testing 1

$$H_0: \sigma^2(\text{v}) = 0$$

$$H_A: \sigma^2(\mathrm{v}) \neq 0$$

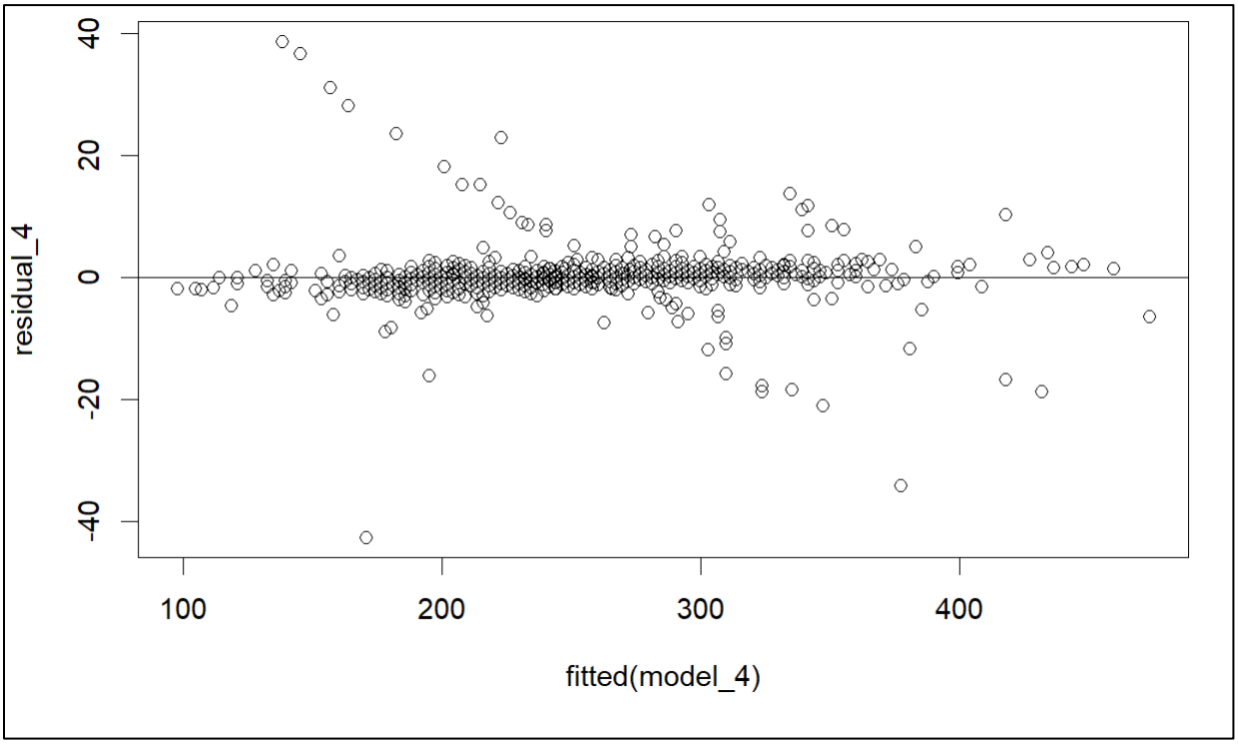
$$F - test\ statistic = \frac{\frac{1058}{4} \times 3784.52 + 17.58}{17.58} = 56{,}941.0193$$

$$p - value = 0$$

Since we have a very small p-value, we reject the null hypothesis, and have evidence to believe that the mean CO2 emissions of different fuel types are different from each other.

However, the researcher had some doubts about the predictive ability of the model as there are less than 10 levels of the fuel type, and hence, decided to use class as the random effect.

As can be seen in the residual graph below, it is evident that independent means of the variable level are interfering with the prediction model and hence the researcher believes it to be an inefficient model to predict CO2 emissions.



Hypothesis Testing 2

$$H_0: \sigma^2(\text{v}) = 0$$

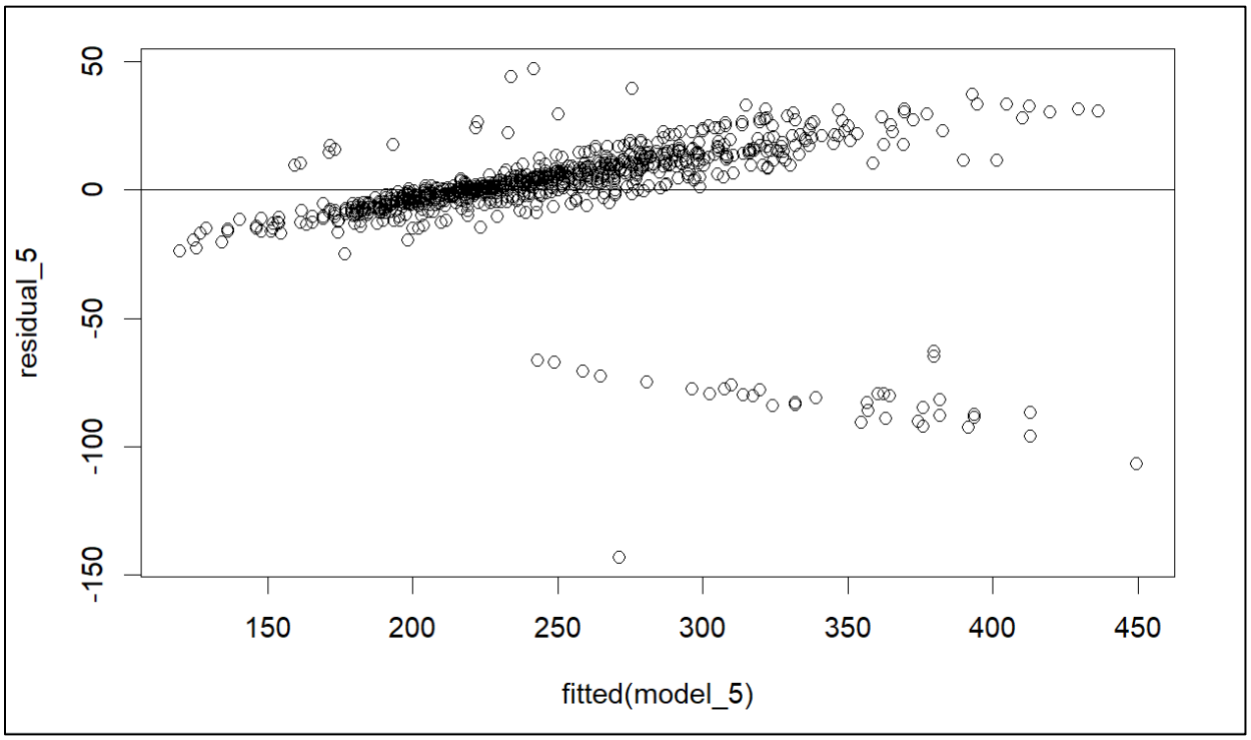
$$H_A: \sigma^2(\text{v}) \neq 0$$

$$F - test\ statistic = \frac{\frac{1058}{15} \times 28.37 + 456.12}{456.12} = 5.39$$

$$p - value = 4.903746e - 10$$

Since we have a very small p-value, we reject the null hypothesis, and have evidence to believe that the mean CO2 emissions of different classes of cars are different from each other.

As is evident from the graph presented below, the model is heteroscedastic as the value of residual increases as the fitted model values increase which implies that an important variable has been omitted, the researcher believes that in this model fuel type is an important predictor of CO2 emissions and might be causing heteroscedasticity.



Conclusions

The researcher studied the Canadian Government's Traffic department's data that summarized the finding for fuel efficiency of multiple cars over the period of 7 years and found to find the Variables that are significant in increasing fuel emissions. After studying the behaviors between the variables of the dataset, the initial variables of interest in predicting the car's CO2 emission were engine size, fuel types, cylinders, and fuel consumption. Through careful analysis, the researcher found that Cars with bigger engine size used premium gasoline as fuel type, and hence had high CO2 emissions.

To study their relationships in depth, the research conducted basic hypothesis testing using regression analysis. The researcher used dummy variables to incorporate the categorical variables in the model. First, the researched tried to predict the CO2 emissions using engine size, cylinder, fuel consumption, CO2 rating, smog rating, make, class, and fuel type of the car and found that engine size, cylinders, make, and smog rating are not significant predictors of CO2 emissions. To improve the predictive ability of the model, the researcher eliminated the insignificant predictors and tested a reduced model including fuel consumption, CO2 rating, class, and fuel type of the car and found that the reduced model is a better predictor of the CO2 emissions.

However, on comparing the fuel consumption with the CO2 emissions, relative to the fuel type, the researcher could see a banded pattern in the graph which motivate the researcher to conduct liner regression using the mixed effects model. In the mixed effects model Fuel Type was taken as the random effect and the fuel consumption was taken as the fixed effect. On conducting advanced analysis, the researcher found that the mean CO2 emissions of different fuel types are different from each other. To get a better estimate of the random variation in the

dataset, the researcher used class of the car as a random effect as well and found that the mean CO2 emissions of different classes of cars are different from each other. Further, the researcher would like to study the variation in the dataset as an impact of the correlation between engine size and the fuel type to get a better idea of which variables to control to reduce the CO2 emission.

From the present analysis, the researcher has evidence to believe that fuel consumption, CO2 rating, class, and fuel type of the car can determine the CO2 emission of the car and can be controlled to reduce the CO2 emission per car.

References

https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64#wb-auto-6