Modeling of Tailpipe CO_2 Emissions on Vehicle Characteristics

Amin Fesharaki

Shiley-Marcos School of Engineering, University of San Diego

Abstract

The objective of this analysis is to statistically investigate the association of tailpipe carbon dioxide emissions in grams per mile to annual primary-fuel petroleum consumption in barrels after controlling for combined miles-per-gallon for the primary fuel type, vehicle manufacturer, make, model, engine displacement, engine cylinders, transmission type, and combined luggage and passenger volume in cubic feet. The hypothesis is that multiple linear regression can accurately model CO_2 emissions with critical variables from the vehicle characteristics. A bivariate analysis found that the proportion of regular gasoline is higher than the population; regular gasoline creates a significant and most considerable impact when analyzing CO_2 tailpipe emissions for all vehicle types. Investigating the associations showed that unknown vehicles contribute the most among all vehicle types and are statistically significant with high associations to emissions, while regular gasoline has the highest association relative to all emission levels. Statistical correlations found that annual petroleum consumption is most strongly correlated with CO_2 emissions, while primary fuel type had the weakest correlation. The final model created removed significant multicollinearity concerns but still exhibited some level of multicollinearity. According to the final model, the model's intercept was 60.31 grams per mile of tailpipe CO_2 emissions. The coefficients for petroleum consumption, combined miles per gallon, engine displacement, cylinders, and volume are respectively 24.43, -1.257, 1.51, 0.51, and 0.00075. All predictor variables resulted in p-values <.0001 except volume, which was .6461. The final model had five degrees of freedom and an F value of 402,618, resulting in a p-value less than .0001 with a confidence level of 95%. These findings indicate that a final multiple regression model explains tailpipe CO_2 emission variance adequately.

Key Words: Carbon Dioxide Emissions, Fuel Economy, Vehicle Characteristic, Statistical Analysis

Table of Contents

List of Equations

List of Tables

List of Figures

Introduction

Background

Greenhouse gas emissions have been one of the primary factors to global warming and climate change. Evidence suggests that vehicle transportation caused 29% of the total greenhouse gas in 2019 (U.S. Energy Information Administration, 2019). According to Salvi et al. (2013), annual primary fuel consumption positively correlates to carbon dioxide emissions, meaning we can decrease CO_2 emissions by decreasing petroleum consumption. This notion suggests reducing carbon dioxide emissions by roughly 20% if every automobile manufacturer designed fuel-efficient vehicles. Awareness of the relationship between carbon dioxide emissions and petroleum consumption gives way to more research and hopefully identifying better ways to design cars to reduce CO_2 tailpipe emissions. Therefore, a model should be developed for CO_2 emissions in an effort for manufacturers to reduce carbon emissions from their cars when designing a new vehicle. Additionally, consumers may also be interested in viewing the model when deciding what car to buy to reduce their carbon footprint.

Purpose

The research objective for this study is to statistically investigate the association of primary fuel tailpipe carbon dioxide emissions in grams per mile to annual primary-fuel petroleum consumption in barrels after controlling for combined miles-per-gallon for the primary fuel type, vehicle manufacturer, engine displacement, transmission type, engine cylinders, and combined luggage and passenger volume in cubic feet. The goal is to create a model that accurately predicts tailpipe CO_2 emissions-based vehicle characteristics to assist automobile manufacturers with reducing their carbon footprint when designing new vehicles. The hypothesis is that multiple linear regression can predict tailpipe carbon dioxide emissions.

Generalized Objective Formula

To statistically analyze the research objective, a generalized model was made to investigate the independent variables that impact tailpipe CO_2 emissions, as seen in (1). The hypothesis is that multiple linear regression can accurately model CO_2 emissions with critical variables from the vehicle characteristics. Furthermore, the goal for the model, using a linear combination of vehicle characteristics, constant, and an error term (ϵ) will attempt to quantify each characteristic impact on carbon dioxide emissions and predict carbon dioxide emissions.

$$
\begin{aligned}
CO_2 = {} & \beta_1 + \beta_2 PetroleoumConsumption + \beta_3 CombinedMPG + \beta_4 VehicleManufacturer \\
& + \beta_5 EngineDisplacement + \beta_6 Cylinders + \beta_7 Volume \\
& + \beta_8 VehicleType \begin{pmatrix} 1 = Unknown \\ 2 = Hatchback \\ 3 = Passenger\ 2-Door \\ 4 = Passenger\ 4-Door \end{pmatrix} \\
& + \beta_9 TransmissionType \begin{pmatrix} 1 = Automatic \\ 2 = Manual \end{pmatrix} \\
& + \beta_{10} FuelType \begin{pmatrix} 1 = Premium\ Gasoline \\ 2 = Midgrade\ Gasoline \\ 3 = Regular\ Gasoline \\ 4 = Diesel \\ 5 = Natural\ Gas \\ 6 = Electricity \end{pmatrix} + \epsilon
\end{aligned}
\tag{1}
$$

Method

Data Collection

This study uses data downloaded from FuelEconomoy.gov, which includes various characteristics of 43,177 vehicle models from 1984 to 2021. Statistical data analysis was performed using a Lenovo ThinkStation P330 Tower with an Intel Xeon E-2186G 6 Core Processor on Windows 10 Pro for Workstations 64. In addition, the software used to carry out data calculations were MATLAB R2020b, Microsoft Excel for Microsoft Office 365 (64-bit), SAS Enterprise Guide 7.1 (64-bit), and Tableau 2020.3. The CSV was opened in Microsoft Excel to convert the data into a table format with emissions, fuel type, make, transmission type, vehicle type categories. The final combined table was imported to MATLAB for distribution curve fitting and SAS for statistical analysis, including bivariate frequencies, Pearson correlation coefficients, association, and chi-squared testing.

Population Characteristics

The data initially consisted of 43,177 samples. After removing the blank and null values through SAS, a total of 42,917 samples were used for descriptive statistical analysis. The samples include vehicles from 138 different manufacturers. Important population characteristics have been identified for each sample. Barrels08 represents the annual petroleum consumption in barrels for each primary fuel type: premium gasoline, midgrade gasoline, regular gasoline, diesel, natural gas, and electricity. Combo08 represents the combined MPG, displ indicates engine displacement in liters, and cylinders show each vehicle's number of engine cylinders. Transmission types include manual or automatic for each vehicle sample.

Moreover, the dataset also includes unknown vehicle types, hatchback, passenger two-door, and passenger four-door, and the vehicle volume in cubic feet. However, it is essential to note that FuelEconomy.gov did not require interior volume dimensions for two-seater passenger cars or any vehicle classified as trucks: vans, pickups, special purpose vehicles, minivans, and sport utility vehicles. Lastly, the vehicles are grouped into the following categories based on how much tailpipe emissions were emitted: ultra-low emissions, very-low emissions, low emissions, standard, polluter, and gross polluter. All research fields (variables) with their respective field description are discussed in Table 1.

Descriptive Statistics

The continuous data variables include CO_2 emissions, petroleum consumption, volume, engine displacement. The discrete variables in the dataset are the combined MPG / MPGe and engine cylinders. The mean value of CO_2 Tailpipe emissions was 465.54 grams per mile with a standard deviation of 119.88, indicating a positively skewed distribution. For the annual petroleum consumption in barrels, the mean value was 17.25, with a standard deviation of 4.49. The conversion from barrels to gallons is 1 barrel equals 42 gallons. In addition, the mean value of volume was 66.88 cubic feet with a standard deviation of 69.12. As mentioned previously, this value is not accurate since interior dimensions for two passenger vehicles and trucks were not required to be recorded in the sample set. Lastly, the mean value of engine displacement was 3.29 liters with a standard deviation of 1.36. In addition to descriptive statistics for each variable**,** Table 2 shows the ranges for each variable.

Categorical variables of interest include primary fuel type, vehicle type, transmission type, manufacturer. From analyzing the entire dataset of 43,177 samples, primary fuel types

included 29.6% premium gasoline, 0.3% midgrade gasoline, 66.5% regular gasoline, 2.8% diesel, 0.1% natural gas, and 0.6% electricity. Vehicle types include 11.7% hatchback, 14.8% passenger two-door, 27.8% passenger four-door, and 45.7% were considered unknown. Transmission types include 70.0% automatic and 30.0% manual. Furthermore, the emissions category is an ordinal value based on the value of each car's respective tailpipe CO_2 emissions. Emission level categories are as follows: 0.72% ultra-low emission, 0.89% very-low emission, 12.87% low emission, 68.42% standard, 13.66% polluter, and 3.41% gross polluter. The emission category in which each vehicle fell into was determined by the population mean and standard deviation, as shown in Table 3.

The number of cylinders is a discrete variable, and therefore determining the mean and standard deviation holds no value. Thus, a box was made for tailpipe CO_2 Emissions relative to the number of engine cylinders, shown in Figure 1, to analyze the discrete variable further. Electric vehicles are not included in the visualization since electric cars typically have no cylinders. Furthermore, there are only 61 data points of 2-cylinder vehicles, including Mazda RX7/8 models and BMW I3 models. In addition, the 16-cylinder vehicles were exclusive to only the Bugatti manufacturer with only 13 data points. Thus, the 2-cylinder and 16-cylinder vehicles will not be compared to the 3-cylinder, 4-cylinder, 5-cylinder, 6-cylinder, 8-cylinder, 10-cylinder, and 12-cylinder vehicles. Moreover, the average tailpipe CO_2 emissions increase as the number of cylinders increases, as seen in Figure 1**.** Excluding 2-cylinder and 16-cylinder vehicles, 3-cylinder vehicles showed the slightest variance among all other cylinder values and had the lowest average tailpipe emissions. After 3-cylinder vehicles, the 5-cylinder vehicles have the second-lowest variation with a low number of outliers. By contrast, the 4-, 5-, 6-, 8-, 10-, 12-

cylinder models exhibit significant variations across all vehicle types with many more sample points.

Bivariate Frequency

Table 3 represents the bivariate frequency of four different vehicle types and two different transmission types across six categories of fuel types for a total of 43,177 samples. Within the table, the independent variables are unknown, hatchback, passenger two-door, and passenger four-door vehicle types and automatic and manual transmission types. In contrast, the dependent variables are the fuel types of premium gasoline, midgrade gasoline, regular gasoline, diesel, natural gas, and electricity. The percentages of each variable for the total population are shown in Table 3. There are proportionally more automatic transmissions than manual transmissions across all fuel types regarding transmission types.

Moreover, premium gasoline accounts for 12,801 samples (29.64% of the total population), with passenger four-door vehicle types having the most significant proportion of 37.8%. Unknown vehicles and passenger two-door vehicles have similar proportions of 27.3% and 24.7%, respectively. Midgrade gasoline has 130 samples (0.30% of the total population), with unknown vehicle types having the most significant proportion of 69.2% and passenger four-door having the second largest proportion of 21.5%. Regular gasoline has 28,733 samples (66.54% of the total population), with unknown vehicle types having the majority proportion of 53.4% and passenger four-door having the second highest of 21.5%. Diesel accounts for 1,196 samples (2.77% of the total population), with unknown having the largest proportion of 57.3% while the passenger four-door has the second-highest proportion of 21.5%. Natural gas has the lowest sample points of 60 (0.14% of the total population), with unknown having the largest proportion of 56.7% and passenger four-door with 38.3%. Lastly, electric vehicles have a total of

257 sample points (0.60% of the total population), with hatchbacks having the largest proportion, 40.9%, and the second highest being unknown vehicle types with 32.7%.

In other words, unknown vehicles account for the highest percentage of all vehicle types for midgrade gasoline, regular gasoline, diesel, and natural gas. Furthermore, regular gasoline has the highest proportion (66.54%) across all other fuel types. Thus, the proportion of regular gasoline is higher than the population; regular gasoline creates a significant and most considerable impact when analyzing CO_2 tailpipe emissions for all vehicle types. The high bivariate frequency can cause the average to sway more in favor of regular gasoline towards its average tailpipe emission value. In addition, the p-value is extremely low (<.0001) for all these proportions, which indicates that there is a statistically significant difference between the proportions of each independent variable and the dependent variable.

Associations

Table 4 outlines the association of emissions category by fuel type and other characteristics: primary fuel type, vehicle type, and transmission type. The CO_2 emissions are converted into categorical data to show that our dependent variable is relative to some of the independent variables in the model. The dependent variable in this table is the CO_2 emission levels, while the independent variables are the primary fuel type, vehicle type, and transmission type. These characteristics have also acquired a p-value of less than .0001, which indicates a statistically significant association concerning the different levels of emission categories. Of the total population, 68.42% of vehicles fell under the standard category, and 45.7% of all vehicle types are classified as unknown. Additionally, as mentioned previously, 70% of vehicles are automatic transmission types. In short, unknown vehicles contribute the most among all vehicle

types and are statistically significant with high associations to emissions, while regular gasoline has the highest association relative to all emission levels.

Results

Probability Density Function

A normal probability density function was determined for the CO_2 emissions, as seen in (2). Additionally, the independent variable in the density function is CO_2 tailpipe emissions. When considering the practicality of this density function, it was determined that any value less than 0 CO_2 emissions would result in zero chance since vehicles having a negative carbon dioxide emission is impossible because a vehicle cannot have negative carbon dioxide emissions based on current technology. In addition, the probability distribution exhibits a normal bell-shaped curve which allows the density function to predict what percentage of CO_2 emissions will fall at a specific value, as shown in Figure 2. For research purposes, the data was forced into a normal curve by creating a modified number of bins to get the data points as close as possible to a normal distribution. The center of the curve for the CO_2 probability distribution function with a frequency histogram has an approximate value of 450 grams per mile, similar to the mean value of 465.54 calculated through SAS. Due to the normally distributed histogram and probability chart, the density function predicts 95% of CO_2 emissions values to fall between 225.75 and 705.27 by calculating two standard deviations above and below the mean as shown in Figure #.

$$PDF_{CO2} = f(\mathrm{x}; \mu = 465.538, \sigma = 119.88)$$

$$= \begin{cases} \dfrac{1}{[(\sqrt{2\pi})(119.88)]} e^{-\{(x-465.538)^2/[(2)(119.88)]^2\}}, & \mathrm{x} < 0 \\ 0, & \mathrm{x} \geq 0 \end{cases} \quad (2)$$

Where PDF_{CO2} = probability distribution function of tailpipe CO_2 emissions

μ = population mean

σ = population standard deviation

Pearson Correlation Coefficients

Table 5 displayed the Pearson correlation coefficients (r) for all the vehicle characteristics in the data set. The correlation values resulted in a p-value <0.0001, indicating that the variable was statistically significant at a confidence level of 95%. With respect to statistical correlation, absolute values of r > .75 describe a strong relationship while absolute values of .25 < r < .50 indicate a weak relationship and r <.25 values are considered very weak relationships (Statology, 2020). The strongest positive correlation relative to CO_2 emissions was the annual petroleum consumption in barrels for fuel type (.9885), which supports the notion that a higher amount of CO_2 is associated with higher levels of annual petroleum consumption. The strongest negative correlation was the combined MPG (-.9184), indicating a strong negative linear correlation concerning CO_2 emissions. The inverse correlation shows that the amount of CO_2 emitted tends to decrease as the value of MPG increases. Other variables with relatively strong positive correlations concerning CO_2 tailpipe emissions include the derived value based on classified CO_2 tailpipe emissions (.8894), engine displacement in liters (.7954), and the number of engine cylinders (.7438).

From the Pearson correlation table, two variables exhibit weak correlations and two variables that exhibit a very weak relationship to the CO_2 tailpipe emission, shown in Table #. The two weak correlations are vehicle volume (-.4342) and vehicle type (-.3636). Very weak statistical correlations include make ID (-.2157) and primary fuel type (-.1128) being the weakest

correlated across all variables. Note that negative correlation values indicate an inverse relationship, as previously mentioned. Furthermore, statistical correlations of make ID and primary fuel type are also weak across all other variables. Thus, the weak correlations support the notion that the nominal and ordinal variables tend to impact the CO_2 tailpipe emissions the least among other data types based on the Pearson Correlation Coefficient Table.

Pearson's Chi-Square Test

Emissions category and vehicle types had a correlation value of .3054, indicating a weak correlation indicated Table. Therefore, a chi-squared test for independence or homogeneity was necessary to investigate further the significance of the association between the association of vehicle types relative to the emissions category. The null hypothesis of homogeneity states that the proportion of the emissions category is the same for each population (Devore, 2016). In contrast, the alternative hypothesis states that the emissions categories are not homogeneously relative to vehicle types. A chi-squared test of independence focuses on the lack of association between the variables. Therefore, emissions category and vehicle type are independent of one another is the null hypothesis for independence. Furthermore, the chi-squared homogeneity test was performed for this study, resulting in the null hypothesis claiming that emissions categories are homogenous with respect to the four vehicle types.

In addition, a table of expected values, as shown in Table 7, was also created in order to perform chi-squared testing. Category combination was unnecessary because all cells in the expected count table are at least five, allowing the chi-squared test for homogeneity to be applied safely. As indicated in Table 8, the chi-squared test resulted in a total value of 9,406.17, and with 15 degrees of freedom, the p-value was less than .001. With a p-value this small, the null hypothesis would be rejected at any practicable significance level. Therefore, the hypothesis of

homogeneity is rejected in favor of the alternative hypothesis that the distribution of reason for nonconformity is different for the emission categories.

On the other hand, an immense sample set can potentially cause inflated chip-squared statistics and decrease the p-value, and in this case, the total chi-squared value was so large that the p-value could be incorrectly low. The most notable example in the dataset is that 173 gross polluters were expected, but only one was observed. Therefore, a secondary analysis is needed to review the validity of the chi-squared test further.

Multicollinearity

Multicollinearity issues can arise when two or more predictor (explanatory) variables in a multiple regression model are related to each other and likewise related to the response variable (Akinwande et al., 2015). In other words, multicollinearity can occur when independent variables in a regression model are correlated. As such, Table 5 identifies predictor variables that are strongly correlated with each other ($r < |.75|$): emissions category with petroleum consumption (.8791) and combined mpg (-.8415), engine displacement with petroleum consumption (.7843), MPG (-.7327) and cylinders (.9406), volume with vehicle type (.7418), and cylinders with petroleum consumption (.7438). However, strongly correlated independent variables do not always imply that there will be a significant estimation problem. In other words, correlation is not collinearity but instead an indication of potential multicollinearity. Therefore, further investigation is required before removing the variables.

In addition, the initially planned model includes categorical data non-indicator variables, which were assigned numerical IDs. The model contains categorical predictor variables: make ID, vehicle type, emissions category, transmission type, and primary fuel type. Converting categorical data to numerical IDs is not sufficient enough to be used in regression models. SAS

converted the categorical data into discrete numbers. Therefore, the residuals, the vertical deviation from the estimated regression line, treat the data as discrete rather than categorical data. Instead, the categorical data should be in values of 0 and 1 or as dummy variables. The model's categorical data as numerical IDs that increase is incorrect because it imposes an ordering on the categories. Categorical variables with C possible categories into a multiple regression model require the use of C-1 indicator variables which are not present in the models provided (Devore, 2016).

Initial Regression Model

Based on (1), a multi regression model was made (3). The response, or dependent variable, was CO_2 emissions, while the predictor, or explanatory or independent variables, are petroleum consumption in, combined miles-per-gallon, vehicle manufacturer, engine displacement, engine cylinders, and combined luggage and passenger volume. The originally planned model used 42,917 samples and yielded a dependent mean of 465.54 and an adjusted R^2 value of .9828; the R^2 value refers to the variation in CO_2 emissions by the variance of the independent variables. However, this does not necessarily represent a near-perfect model and instead raises over-fitting concerns. Over-fitting refers to situations where the model fails to generalize estimates outside its own dataset (Rigg & Hankins, 2015). Thus, having a very high R^2 value should be investigated further to see if the predictor variables may not necessarily belong.

$$CO_2 Tailpipegpm = 79.5534 + 20.7437(barrels08) - 2.5356(combo08) + 0.0033(make_{id}) + 2.1722(displ) + 2.2629(cylinders) - .0078(volume) - 0.4136(vehtype) + 12.8766(emissionscat) + 0.7799(transtype_{id}) + 2.9902(prifueltype) + \in \quad (3)$$

Moreover, the original model contained variables with high variance inflation factors (VIF) which "measures how much multicollinearity has increased the variance of a slope estimate" (Stine, 2011). The VIF values between five and ten are typically used to indicate problems with multicollinearity (Tay, 2017). Using this threshold, high VIFs in the model were include annual petroleum consumption (9.34), combined MPG (6.08), engine displacement (7.29), and cylinders (6.47). However, further investigation of these variables was not determined to have a severe enough impact on removing the variables for the final model. To summarize, the initially planned model can predict data based on that data set. However, it would not predict anything else about the other data sets due to the issues surrounding categorical indicator variables, overfitting, and multicollinearity issues.

Final Regression Model

Categorical non-indicator variables (make ID, vehicle type, emissions category, transmission type, and primary fuel type) were removed for the final model due to multicollinearity concerns. In addition, the model removes 19,903 observations due to null values in order to provide the most accurate model possible given the dataset from FuelEconomy.gov. Missing volume samples (19,731), all associated with unknown vehicle types, accounted for the majority (99.1%) of the null observations. (4) represents the final model to predict tailpipe CO_2 emissions with the

following predictor variables: annual petroleum consumption (barrels08), combined MPG (combo08), engine displacement (displ), cylinders, and volume. The model was determined to have an adjusted R^2 value of 0.9886, suggesting 98.86% of the variance in CO_2 emissions in relation to the predictor variables. The dependent mean is 415.94, which is lower than the original planned model's dependent mean (465.54). The model also resulted in a p-value less than .0001 with 5 degrees of freedom, as opposed to 10 degrees of freedom in the initial model. Therefore, the final model had a lower variance coefficient, lower error value, and no categorical variables indicating a more accurate model when comparing both models.

$$CO_2 Tailpipegpm = 60.2159 + 24.4364(barrels08) - 1.2573(combo08) + 1.5163(displ) + 0.5106(cylinders) + 0.0008(volume) \quad (4)$$

Although, reducing the variable count in the regression model does not entirely remove the multicollinearity concerns. The new model statistically investigates the association of tailpipe carbon dioxide emissions to annual consumption in barrels after controlling combined miles per gallon, engine displacement, engine cylinders, and volume. Table 5 reveals a multicollinearity issue within the final model due to the strong inverse correlation of .9050 between petroleum consumption and combined MPG, causing variable inflation. Intuitively, this makes sense in practical applications where fuel-efficient cars with lower miles per gallon will probably consume less petroleum.

Discussion

Hypothesis

According to the final model, the model's intercept was 60.31 grams per mile of tailpipe CO_2 emissions. The coefficients for petroleum consumption, combined miles per gallon, engine displacement, cylinders, and volume are respectively 24.43, -1.257, 1.51, 0.51, and 0.00075. All predictor variables resulted in p-values <.0001 except volume, which was .6461. The relatively high p-value from volume could be from removing 19,731 samples which made up 45.97% of all samples (42,917) used in the dataset.

The objective of this analysis is to statistically investigate the association of tailpipe carbon dioxide emissions in grams per mile to annual primary-fuel petroleum consumption in barrels after controlling for combined miles-per-gallon for the primary fuel type, vehicle manufacturer, make, model, engine displacement, engine cylinders, transmission type, and combined luggage and passenger volume in cubic feet. This study's null hypothesis is that there are no correlations between CO_2 tailpipe emissions and petroleum consumption; in other words, vehicle characteristics do not impact CO_2 emissions. In contrast, the alternative hypothesis suggests a relationship between the independent variables from the vehicle characteristics and CO_2 emissions.

The final model had five degrees of freedom and an F value of 402,618, resulting in a p-value less than .0001 with a confidence level of 95%. Thus, the null hypothesis is rejected, and the alternative hypothesis is accepted. We can further conclude that there is a significant association between CO_2 emissions and annual petroleum consumption. In other words, for every 1 barrel of annual petroleum consumed, tailpipe CO_2 emissions are expected to increase by 24.43 grams per mile if all other predictor variables were fixed. However, by accepting the alternative

hypothesis, there is a risk of having a Type I error if the null hypothesis should have been accepted instead. In conclusion, the final multiple regression model proved to explain tailpipe CO_2 emission variance adequately.

Strengths and Limitations

The research project has a voluminous data set of 43,177; this can be seen as a strength because of the enormous magnitude of sample points which can indicate a good representation of the population. By having a large sample size, we can move closer to having a normal distribution in the dataset. Additionally, larger samples are more robust to outliers and provide a more accurate population parameter estimate. Furthermore, another strength of the study is that it is continuously updated to provide more accurate results. For example, FuelEconomy.gov mentions that the sample set contains the MPG for all 1985-2997 models, and some 2011-2016 model years have been updated. Similarly, the fuel prices for gasoline and diesel are updated every week. However, having such a large sample size can cause some problems due to missing information. According to Devore (2016), large sample sizes can amplify the differences in estimated P values, causing a potential problem when analyzing the values in conjunction with the null hypothesis.

Moreover, the data notes in the excel file highlight potential data limitation issues if the whole data set is used, which can impact the precision and accuracy when modeling the data. For example, one limitation would be the interior volume dimensions since the volume dimensions are not required for two-seater passenger cars or any vehicle classified as a truck. From the data provided by fuel economy, there are 19,731 data points with a volume of 0. Another limitation is the number of data points (261) with no cylinder values, and that unrounded MPG values are not available for some vehicles. Lastly, according to FuelEconomy.gov, the tailpipe CO_2 is based on

EPA testing for model years 2013 and beyond and using an EPA emission factor of -1 (not available) for previous model years. Thus, if the whole dataset (43,177), the limitations far exceed the strengths due to missing data values.

Future Research

Future research should consider the potential effects of these limitations more carefully. One example of avoiding a limitation is creating a dataset that uses the same EPA testing across all observations. Future research should also examine strategically better sampling methods to avoid potential bias. For example, the 16-cylinder vehicles were exclusive to 14 Bugatti's only. Therefore, providing more samples of other supercars can be helpful when including 16-cylinder vehicles in the dataset. In addition, a different model can be created to predict CO_2 tailpipe emissions of two-seater passenger cars where the interior volume dimensions were provided, resulting in null values skewing the data points. Similarly could be done to any vehicle classified as trucks: vans, pickups, special purpose vehicles, minivan, and sport utility vehicles. Moreover, future studies should also further develop and confirm these initial findings with more samples to provide manufacturers with a model of vehicle characteristics to meet their reduced CO2 emission goals.

References

Akinwande, M. O., Dikko, H. G., & Samson, A. (2015). Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis. *Open Journal of Statistics*, *05*(07), 754–767. https://doi.org/10.4236/ojs.2015.57075

B.L.Salvi, B. L. (1993). Alternative fuels for transportation vehicles: A technical review. *ScienceDirect*. https://doi.org/10.4271/931816

Devore, J. L. (2016). *Probability and statistics for engineering and the sciences*. Nelson.

Environmental Protection Agency. (2021, April 14). *Sources of Greenhouse Gas Emissions*. EPA. https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions#:~:text=In%202019%2C%20greenhouse%20gas%20emissions,of%20U.S.%20greenhouse%20gas%20emissions).

The official U.S. government source for fuel economy information. Fuel Economy. (n.d.). https://fueleconomy.gov/.

Rigg, J., & Hankins, M. (2015). Research on Modeling Methods Study. *Value in Health*.

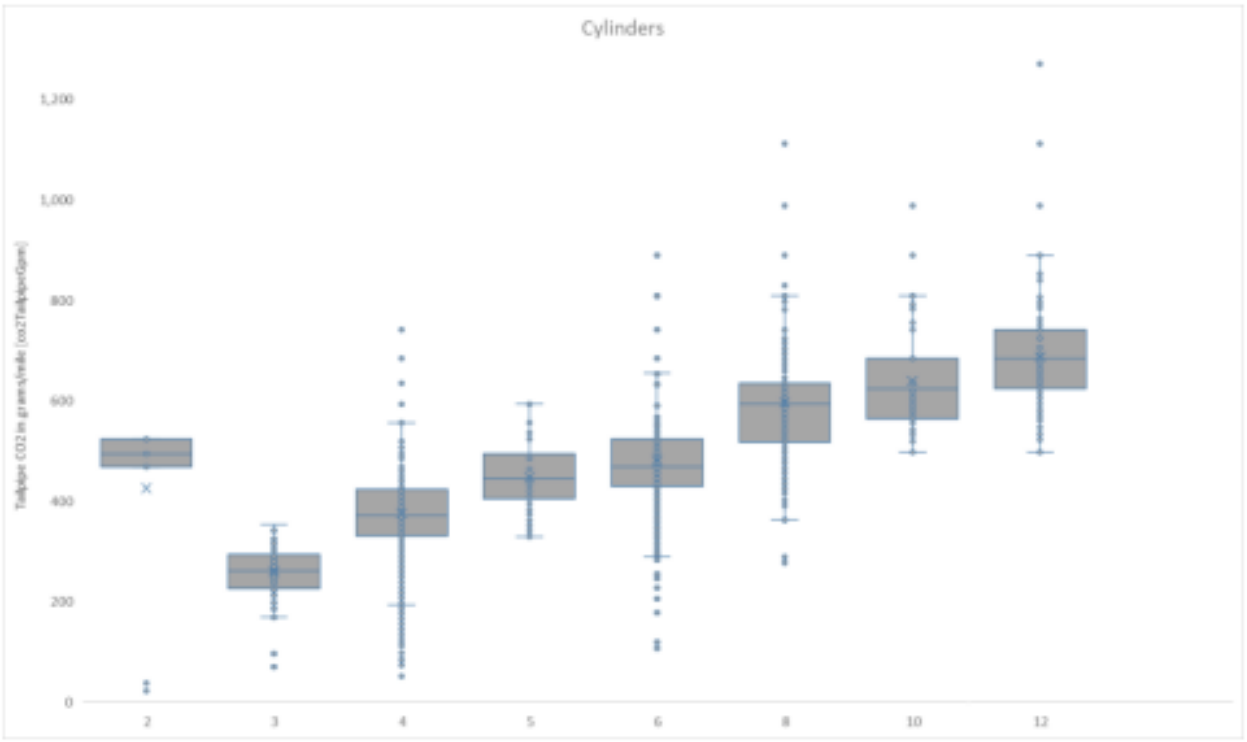
Tay, R. J. (2017). Correlation, Variance Inflation and Multicollinearity in Regression Models. *Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis*. https://doi.org/10.1063/1.4823947

Zach. (2021, April 27). *What is Considered to Be a "Strong" Correlation?* Statology. https://www.statology.org/what-is-a-strong-correlation/.

Appendix

Figure 1

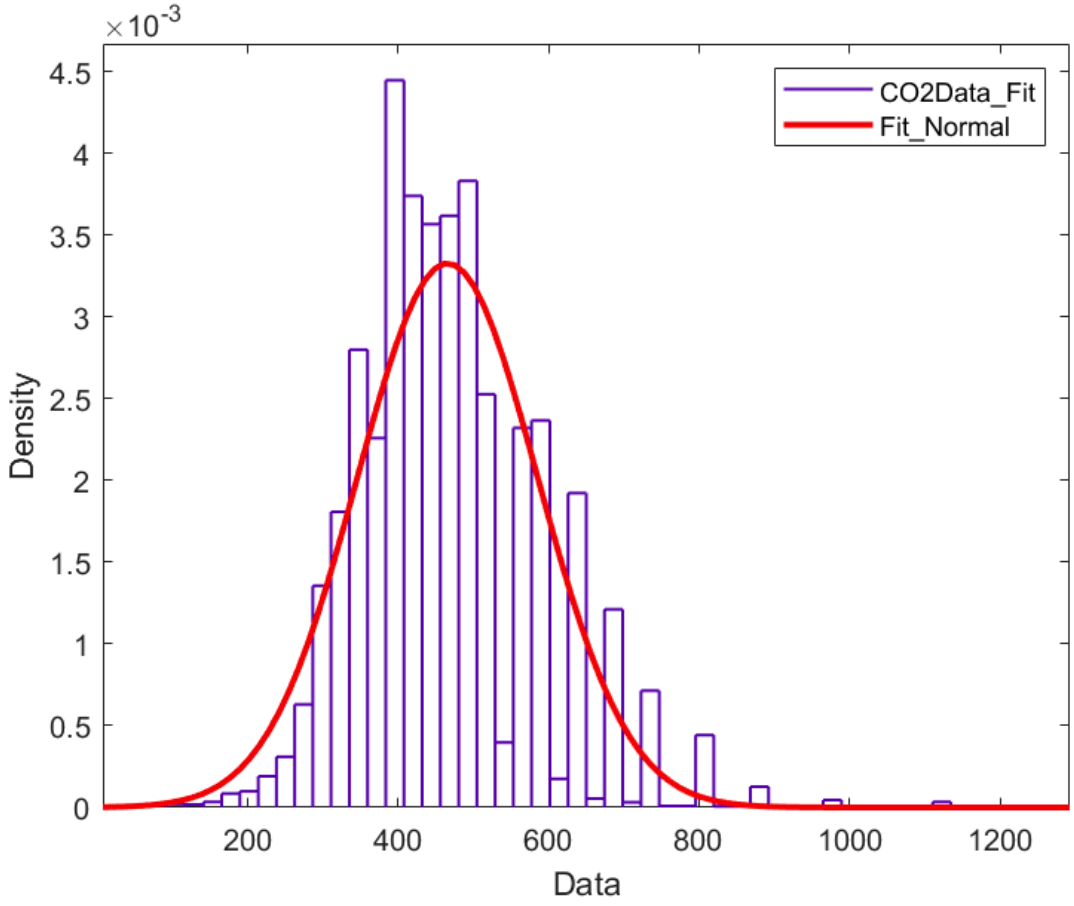
Tailpipe CO2 Emissions in Grams Per Mile Relative to Engine Cylinders



Note: A box and whisker plot was made to investigate engine cylinders relative to carbon dioxide emissions. From "Carbon Dioxide Emissions and Personal Vehicle Choices," by E. Belekanich, 2020.

Figure 2

Probability Density Function of Carbon Dioxide Tailpipe Emissions



Note: The number of bins was modified in order to provide a normal distributed curve

Table 1

Description of each research variable

Research Fields	Field Description
co2TailpipeGpm	tailpipe CO2 in grams/mile
barrels08	annual petroleum consumption in barrels
comb08	combined MPG for fuel Type
make_id	manufacturer (division) assigned ID
displ	engine displacement in liters
cylinders	engine cylinders
volume	vehicle volume (cubic feet) = hlv + hpv + lv2 + lv4 + pv2 + pv4
vehtype	categorized vehicle type (Hatchback | 2-Door Passenger | 4-Door Passenger | Other)
emissionscat	derived value based on classified CO2 tailpipe emissions in grams per mile
transtype_id	transmission type
prifueltype	primary fuel type based

Table 2

Descriptive Statistics for Continuous Vehicle Characteristics

Variable	N	Mean	Std Dev	Minimum	Maximum	Range
co2TailpipeGpm	42917	465.5412	119.8828	22	1270	1248
barrels08	42917	17.2547	4.48843	0.06	47.08714	47.02714
displ	42917	3.28686	1.35693	0.6	8.4	7.8
volume	42917	66.88492	69.12347	0	538	538

Table 3

Characteristics of 43,177 Sample Vehicle Models by Primary Fuel Type

Characteristic	Population		Premium Gasoline		Midgrade Gasoline		Regular Gasoline		Diesel		Natural Gas		Electricity	
	N	%	n	%	n	%	n	%	n	%	n	%	n	%
	43177	100	12,801	29.6	130	0.3	28733	66.5	1,196	2.8	60	0.1	257	0.6
Vehicle Type														
Unknown (0)	19730	45.7	3491	27.3	90	69.2	15346	53.4	685	57.3	34	56.7	84	33
Hatchback (1)	5070	11.7	1313	10.3	0	0	3535	12.3	115	9.6	2	3.3	105	41
Passenger 2-Door (2)	6394	14.8	3157	24.7	12	9.2	3120	10.9	103	8.6	1	1.7	1	0.4
Passenger 4-Door (3)	11983	27.8	4840	37.8	28	21.5	6732	23.4	293	24.5	23	38.3	67	26
Transmission Type														
Automatic (1)	30210	70	9411	73.5	130	100	19588	68.2	773	64.6	60	100	248	100
Manual (2)	12956	30	3390	26.5	0	0	9143	31.8	423	35.4	0	0	0	0

Note: Table 3 indicates the bivariate frequencies with p-values <.0001 based on the Pearson chi-square test of association

Table 4

Association of Emissions Category by Fuel Type and Other Characteristics

Variable	Population *N*(%) (N=43,177)		Ultra-Low Emission *n*(%) (n=321)		Very-Low Emission *n*(%) (n=384)		Low Emission *n*(%) (n=5,556)		Standard *n*(%) (n=29,543)		Polluter *n*(%) (n=5,899)		Gross Polluter *n*(%) (n=1,474)	
Primary Fuel Type														
Premium Gasoline (1)	12,801	(29.6%)	24	(7.5%)	70	(18.2%)	1,169	(21.0%)	9,798	(33.2%)	1,262	(21.4%)	478	(32.4%)
Midgrade Gasoline (2)	130	(0.3%)	0	(0.0%)	0	(0.0%)	0	(0.0%)	124	(0.4%)	6	(0.1%)	0	(0.0%)
Regular Gasoline (3)	28,733	(66.5%)	40	(12.5%)	311	(81.0%)	4,066	(73.2%)	18,971	(64.2%)	4,358	(73.9%)	987	(67.0%)
Diesel (4)	1,196	(2.8%)	0	(0.0%)	0	(0.0%)	303	(5.5%)	629	(2.1%)	259	(4.4%)	5	(0.3%)
Natural Gas (5)	60	(0.1%)	0	(0.0%)	3	(0.8%)	18	(0.3%)	21	(0.1%)	14	(0.2%)	4	(0.3%)
Electricity (6)	257	(0.6%)	257	(80.1%)	0	(0.0%)	0	(0.0%)	0	(0.0%)	0	(0.0%)	0	(0.0%)
Vehicle Type														
Unknown (0)	19,730	(45.7%)	91	(28.3%)	50	(13.0%)	739	(13.3%)	12,579	(42.6%)	5,119	(86.8%)	1,152	(78.2%)
Hatchback (1)	5,070	(11.7%)	122	(38.0%)	128	(33.3%)	1,820	(32.8%)	2,952	(10.0%)	47	(0.8%)	1	(0.1%)
Passenger 2-Door (2)	6,394	(14.8%)	7	(2.2%)	11	(2.9%)	703	(12.7%)	5,193	(17.6%)	339	(5.7%)	141	(9.6%)
Passenger 4-Door (3)	11,983	(27.8%)	101	(31.5%)	195	(50.8%)	2,294	(41.3%)	8,819	(29.9%)	394	(6.7%)	180	(12.2%)
Transmission Type														
Automatic (1)	30,210	(70.0%)	312	(100.0%)	301	(78.4%)	3,202	(57.6%)	20,730	(70.2%)	4,557	(77.3%)	1,108	(75.2%)
Manual (2)	12,956	(30.0%)	0	(0.0%)	83	(21.6%)	2,354	(42.4%)	8,813	(29.8%)	1,341	(22.7%)	365	(24.8%)

Note: All p-values were less than .0001 based on Pearson chi-square test of association

Table 5

Pearson Correlation Coefficients for 42,917 Observations

	co2Tailpipe	barrels08	comb08	make	displ	cylinder	volume	vehtype	emissionscat	transtype_
co2TailpipeGpm	1.0000	.9885	(.9184)	(.2157)	.7954	.7438	(.4323)	(.3626)	.8894	(.1128)
barrels08	.9885	1.0000	(.9050)	(.2117)	.7843	.7337	(.4266)	(.3580)	.8791	(.1084)
comb08	(.9184)	(.9050)	1.0000	.2072	(.7327)	(.6863)	.4161	.3313	(.8415)	.1234
make_id	(.2157)	(.2117)	.2072	1.0000	(.2823)	(.2670)	.1165	.0940	(.1755)	.0710
displ	.7954	.7843	(.7327)	(.2823)	1.0000	.9046	(.3628)	(.2631)	.6703	(.2149)
cylinders	.7438	.7337	(.6863)	(.2670)	.9046	1.0000	(.2648)	(.1524)	.6185	(.2181)
volume	(.4323)	(.4266)	.4161	.1165	(.3628)	(.2648)	1.0000	.7418	(.3627)	.0498
vehtype	(.3626)	(.3580)	.3313	.0940	(.2631)	(.1524)	.7418	1.0000	(.3054)	(.0340)
emissionscat	.8894	.8791	(.8415)	(.1755)	.6703	.6185	(.3627)	(.3054)	1.0000	(.0874)
prifueltype	(.1128)	(.1084)	.1234	.0710	(.2149)	(.2181)	.0498	(.0340)	(.0874)	1.0000

Note: All correlation values in a p-value <.0001

Table 6

Two-way Contingency Table of Vehicle Type and Emissions Category

Frequency Table	**Hatchback**	**Passenger 2-Door**	**Passenger 4-Door**	**Unknown**	**Total**
GROSS POLLUTER	1.00	141.00	180.00	1,152.00	1,474.00
Low Emission	1,820.00	703.00	2,294.00	739.00	5,556.00
Polluter	47.00	339.00	394.00	5,119.00	5,899.00
Standard	2,952.00	5,193.00	8,819.00	12,579.00	29,543.00
Ultra-Low Emission	122.00	7.00	101.00	91.00	321.00
Very Low Emission	128.00	11.00	195.00	50.00	384.00
Total	5,070.00	6,394.00	11,983.00	19,730.00	43,177.00

Table 7

Expected Value Table for Vehicle Types and Emission Category

Expected Value Table	**Hatchback**	**Passenger 2-Door**	**Passenger 4-Door**	**Unknown**
GROSS POLLUTER	173.08	218.28	409.08	673.55
Low Emission	652.41	822.78	1,541.97	2,538.85
Polluter	692.68	873.57	1,637.16	2,695.58
Standard	3,469.05	4,374.97	8,199.13	13,499.86
Ultra-Low Emission	37.69	47.54	89.09	146.68
Very Low Emission	45.09	56.87	106.57	175.47

Table 8

Chi-Squared Table for Emission Category and Vehicle Type

Chisq Table	**Hatchback**	**Passenger 2-Door**	**Passenger 4-Door**	**Unknown**	**Total**
GROSS POLLUTER	171.09	27.36	128.28	339.86	666.59
Low Emission	2,089.61	17.44	366.77	1,275.95	3,749.78
Polluter	601.87	327.12	943.98	2,178.73	4,051.70
Standard	77.06	152.96	46.86	62.81	339.70
Ultra-Low Emission	188.57	34.57	1.59	21.14	245.87
Very Low Emission	152.45	36.99	73.37	89.72	352.53
Total	-	-	-	-	9,406.17

Note: The p-value of the table resulted in less than .001