The Effect of Vehicle Specifications on Carbon Dioxide Emissions

Yunsoo Jeong

University of San Diego

ADS-500A-01: Probability and Statistics for Data Science

Professor Matthew Vanderbilt

March 1, 2021

Abstract

The use of petroleum fuels by automobiles have long been controversial due to its negative impacts on the environment. The purpose of this study is to further study the factors that lead to significant carbon dioxide emissions. It was hypothesized that if the vehicle specifications such as the primary fuel types, petroleum utilization, fuel economy, engine cylinders, and engine displacement numbers are changed, then the carbon dioxide emissions will be affected accordingly. This means that primary fuel types, petroleum utilization, engine cylinders, and engine displacement values have direct relationships with the carbon dioxide emissions of the automobiles. The objective of this study is to educate the general public about which factors may affect the production of carbon dioxide emissions and to highlight the importance of the environmentally friendly energy source for automobiles. The co2TailpipeGpm represents the tailpipe CO_2. The co2TailpipeGpm showed confidence intervals between 464.41 and 466.67 at 95% confidence level. The mean co2TailpipeGpm value was 462.77 grams per mile. The statistical significance level of the study was .005, and the p-value of the study was less than .001. The study was successful in finding significant connections between the co2TailpipeGpm and the independent variables mentioned above. For example, the barrels08, which represents the yearly petroleum utilization in barrels, showed strong correlation toward co2TailpipeGpm. The vehicles that are powered by the electricity showed least production of carbon dioxide emissions and was able to support the necessity of the more responsible energy source of automobiles.

Table of Contents

List of Tables

List of Figures

List of Equations

Introduction

The exploitation of petroleum for operations of motor vehicles have long been controversial due to air pollution issues. It is no surprise that many car buyers often take fuel economy and pollution levels into consideration when purchasing their vehicles. Some may consider the gasoline vehicles with high MPGs, and others may consider hybrid or electric vehicles. Many advocates for changes to be made in the automobile industry. Many car manufacturers are manufacturing automobiles that use different sources of energy. To replace the traditional petroleum fuels, other energy sources, including natural gas, hydrogen, propane, biodiesel, and electricity are being currently used or considered. (U.S. Department of Energy, n.d.). The purpose of this study is to analyze how much petroleum fuel is used by automobiles, and what factors have direct impacts on the levels of carbon dioxide released by automobiles. This study may be helpful to potential car buyers who would like to find out if there are correlations between types of cars and carbon dioxide emissions. This study will also give insights about the necessity of new energy sources for automobiles.

The purpose of this study is to identify factors that lead to more carbon dioxide emissions. It is hypothesized that vehicle specifications such as the fuel types, fuel economy, and engine displacements and cylinders have direct influences on the carbon dioxide emissions. The study will test various potential factors, including fuel types, engine displacements, and engine cylinders to identify the correlations between the variables and the carbon dioxide emissions. The objective of the research is to educate the general public and environment advocates about the carbon dioxide emission of the automobiles and the factors of the carbon dioxide emissions. It also aims to emphasize the necessity of the environment-friendly energy sources for motor vehicles.

The Equation 2 in the Appendix states the generalized formula for the variables that affect the carbon dioxide tailpipe emissions. As can be seen from the formula, five variables are expected to affect the carbon dioxide emissions. The five variables are the yearly petroleum uses, combined MPG, engine displacements, engine cylinders, and vehicle volumes.

Method

In order to execute the study, 43,177 vehicle information have been gathered from the U.S. Department of Energy. For example, various information such as year, vehicle volume, vehicle type, and MPG information were collected (U.S. Department of Energy, 2020). The dataset includes wide range of data, including various types of data. For example, carbon dioxide tailpipe emission information is categorized as continuous data type. The combined MPG information can be categorized as discrete data type. Furthermore, the manufacturer and models are considered nominal data type.

The population consists of various types of vehicles. There are four types of vehicle types included: unknown, hatchback, passenger 2-door, and passenger 4-door. There are two transmission types included in the population: automatic and manual. The vehicles fueled by the electricity have also been included for the purpose of this study. The data from electric cars will be helpful when comparing the carbon dioxide emission rates from the vehicles that are fueled by the gasoline, diesel, and natural gas. The years of the vehicles range from 1984 to 2021.

Table 1 in the Appendix describes the descriptive statistics of main variables of the study. The barrels08 represent the yearly petroleum usage in barrels. Hence, it can be interpreted that the mean of 17.15 barrels of petroleum fuel was utilized for each vehicle. The standard deviation for barrels08 was 4.66 barrels. The co2tailpipegpm represent the carbon dioxide tailpipe emission. It can be suggested that the mean value of carbon dioxide by the vehicles of the study was about 462.77 grams per mile. The standard deviation for co2tailpipegpm was 124.77 grams per mile. As can be seen from the Table 1, the mean value for the year of the vehicles studied by this study was 2002.43 with standard deviation of 11.60. The mean value of the yearly utilization of petroleum fuel was 17.15 barrels, and the standard deviation was 4.66. The mean value for the vehicle volume was 66.93, and the standard deviation was 69.04. The mean value for the combined MPG was 20.85, and the standard deviation was 8.22.

The Figure 1 describes boxplot visualization of the yearly petroleum utilization in barrels for fuel type 1. The box plot was utilized in order to analyze the distribution of the dataset of the petroleum utilization. The use of boxplot is helpful because it allows the readers to study the distribution of the dataset by providing information such as the minimum and maximum values, the outliers, lower and upper quartiles, and the median. The outliers are values that show significant deviations from the majority of the dataset (Devore, 2016, pp. 40). The outliers are represented by the blue dots in figure 1. Hence, by looking at Figure 1, it can be predicted that most of the vehicles belong in the dataset utilize approximately 14 to 19 barrels of petroleum fuels per year. As can be interpreted from the graph, the median value can be found between 15 and 20 barrels. The figure also suggests that the upper quartile value can be identified between median and 19 barrels, and the lower quartile value can be identified between 14 barrels and the median. The minimum value is approximately 7 barrels, whereas the maximum value is approximately 26 barrels.

The Table 2 found in the Appendix shows the relationship between the variables such as the vehicle type, transmission type, and the primary fuel type. Based on the information provided by the table, it can be stated that most vehicles of the population possess automatic transmission type as the automatic transmission type comprise 70 percent of the population. The manual transmission type comprises 30 percent of the population. The unknown vehicle type is more likely to be powered by the premium gasoline. The hatchback and passenger-4-door vehicle types are more likely to be powered by regular gasoline. The passenger 2-door vehicle type is more likely to be powered by premium gasoline. Given that there are 28,733 vehicles powered by regular gasoline and 60 vehicles powered by natural gas, it can be suggested that the vehicles powered by regular gasoline are the most common in the sample, whereas the vehicles powered by natural gas are the least common in the sample.

The Table 3 found in the Appendix shows the relationships between the variables such as the fuel type, vehicle type, transmission type, and the emission levels. As the table shows, the vehicles that utilize the regular gasoline make up the majority of the sample with 66.5 percent of the sample. Vehicles that utilize the natural gas make up the least of the sample with 0.1 percent. The unknown vehicle type makes up the majority of the sample with 45.7 percent, and the automatic transmission type is more frequent than manual transmission as it makes up 70% of the sample. Based on the information provided by Table 3, it can be suggested that gasoline gas is more likely to impact the environment as the significant number of premium gasoline and regular gasoline-fueled vehicles belong in polluter and gross polluter groups. On the other hand, out of the six fuel types, it can be suggested that the electricity powered vehicles are less like to impact the environment as it makes up 80.1 percent of the ultra-low emission group.

Results

Figure 2 shows the probability density function of the data of CO_2 tailpipe gpm. As can be seen from the probability density function, the graph is showing the normal bell curve. Total

52 bins can be identified from the probability density function of tailpipe CO_2. The probability density function can be used to calculate the probability of certain interval of two numbers by calculating the area of the interval of two numbers below the curve (Holmes et al., 2017). The Equation 1 explains that the mean for CO_2 data is 465.538, and the standard deviation is 119.88. The mean value of 465.538 can be located around the peak of the curve in Figure 2.

Table 4 describes the correlations between each variables of this study. As can be observed from the table, the co2TailpipeGpm is most correlated with the petroleum utilization in barrels for fuel type 1. The combined MPG for fuel type 1 is most correlated with the CO_2 Tailpipe. The engine displacement in liters is most correlated with the engine cylinders. The vehicle volume is most correlated with the vehicle type.

While it may seem like the primary fuel type is mostly correlated with engine cylinders as the correlation between the primary fuel type and engine cylinders is the highest in the, the correlation is weak as the correlation is only 21.49 percent. This goes the same with manufacturer ID as there is no significant correlation with other variables. The primary fuel type indicates low correlations to the carbon dioxide tailpipe, vehicle volume, vehicle type, petroleum utilization, and the manufacturer id.

Table 5 from the Appendix represent the contingency table for types of vehicles and the levels of polluters. The contingency table above was utilized to calculate the chi-square value. The chi-square test is basically the calculation of the probability of the anticipated variations given that the hypothesis is accepted (Griffiths et al., 2000). The chi-square value was 9,406.19, and the degree of freedom was 15. Considering that the chi-square value is high, it can be suggested that there is significant discrepancy between the anticipated and actual values

(McDonald, 2020). The p-value was less than .001. Because the p-value is less than the significance level, the null hypothesis is not accepted (Devore, 2016, pp. 643)

The multicollinearity occurs when the independent variable is shown to have significant relationships with other independent variables (Devore, 2016, pp. 606). As can be seen from the Table 4 in the Appendix, there are independent variables in the dataset that have multicollinearity. One example is the engine cylinders and engine displacement. The Table 4 shows the correlation value of .90 between the two variables. It is not possible to resolve the multicollinearity issues entirely as these variables are critical for the purpose of this research. In order to alleviate the multicollinearity in the dataset, four of the independent variables have been removed. Hence, the Trimming Regression Model without the categorical non-indicator variables has been selected for the final model.

The Figure 3 describes the Originally Planned Model. The figure definitely has many differences compared to the Trimming Regression Model, which is represented by Figure 4. While determining which regression model will be most suitable for the purpose of this study, one of the factors that was considered was the R-squared. The strength of the chosen regression model is that the model has high coefficient of determination. The R-squared, also known as the coefficient of determination, is used to describe the accuracy of the regression model by comparing the expected value of y variable and the actual data of the y variable (Schneider et al., 2010, pp. 780). If the coefficient of determination is close to 1, it suggests that the regression model is precise. Considering the R-Squared value for the Trimming Regression Model was almost .99, it can be suggested that the chosen regression model is successful in describing the deviations of the dependent variables (Schneider et al, 2010, pp. 780).

Another factor that was considered when choosing the regression model was whether or not to include the categorical variables. The categorical variables can be included into the regression by utilizing a dummy variable (Devore, 2016, pp. 574). However, it is difficult to include ordinal data such as the fuel type. Hence, it has been determined not to include categorical variables that cannot become the dummy variable. Therefore, the Trimming Regression Model without indicator variable and null has been chosen as the final model as it is best suited for the purpose of this study. The Equation 3 in the Appendix represents final research equation of the study.

Discussion

Initially, it was hypothesized that if the primary fuel types, petroleum utilization, engine cylinders, and engine displacement numbers are changed, then the carbon dioxide emissions will be changed accordingly. The hypothesis was supported. As can be seen from Table 4 of the Appendix, many correlations can be made. First, there is a very strong correlation between the utilization of the petroleum fuel and the carbon dioxide emissions. This is because the correlation coefficient for co2TailpipeGpm and barrels08 is .92. In addition, engine displacements and engine cylinders both show strong correlations with the carbon dioxide emission as the correlation coefficients for engine displacements and engine cylinders are about .90.

By studying which primary fuel type, vehicle type, and transmission type are linked to polluter and gross polluter groups, it was identified that regular gasoline and premium gasoline are more likely to impact the environment, whereas vehicles that are powered by the electricity are less likely to impact the environment. Hence, it can be concluded that the usage of petroleum fuels by the automobiles does indeed impact environment due its direct association with the carbon dioxide emissions.

One of the most important strength of this dataset is that it has a huge dataset. This is very beneficial because having a large dataset lowers the chances of having errors. Because the mean, standard deviation, maximum, and minimum values are important to achieve the objectives of this study, having more datasets definitely aid in lowering errors. As Duan pointed out, having a large sample allows more precise estimation of the population mean (Duan, 2021). Another strength of this dataset it incorporates various independent variables that may affect the dependent variables. For example, the dataset contains various variables such as fuel types, vehicle types, engine displacements and cylinders, and vehicle volume. Having extensive variables is very beneficial for the purpose of this study as the research studies the potential factors that have direct linkages to the carbon dioxide emissions.

The limitation of this dataset is that it does not contain information about every car that has been built from 1984 to 2021. The dataset focuses on the vehicles that are manufactured in North America regions. Hence, it is possible to claim that the dataset is biased. In addition, while having a large dataset may help lowering errors as stated above, it is also possible to have more errors due to having a large dataset. This is because there is a chance that mistakes could be made while collecting information (Lloyd, 2021). Another limitation of this dataset is that there is not a strong control group. Table 3 in the Appendix states that 12,801 samples of premium gasoline were collected, and 28,733 regular gasoline-powered vehicles samples were collected. On the other hand, only 257 samples of electricity-powered vehicles, 1,196 samples of diesel-powered vehicle samples, 60 samples of natural gas-powered vehicles were collected. While the dataset itself is successful in suggesting which type of primary fuel type is responsible for polluting the environment, the dataset would have been more accurate if more samples were collected for diesel, natural gas, and electricity-powered vehicles.

Appendix

Table 1

Descriptive Statistics of Main Variables

Variable	Observations	Mean	Standard Deivation	Min	Max
co2TailpipeGpm	43177	462.77	124.77	0	1269.57
barrels08	43177	17.15	4.66	0.06	47.09
displ	43177	3.29	1.36	0	8.4
cylinders	43177	5.71	1.76	2	16
year	43177	2002.43	11.6	1984	2021
volume	43177	66.93	69.04	0	538
comb08	43177	20.85	8.22	7	141

Table 2

Characteristics of 43,166 Sample Vehicle Models by Primary Fuel Type

Variable	Population *N* (%) (N=43,177)	Premium Gasoline *n* (%) (n=12,801)	Midgrade Gasoline *n* (%) (n=130)	Regular Gasoline *n* (%) (n=28,733)	Diesel *n* (%) (n=1,196)	Natural Gas *n* (%) (n=60)	Electricity *n* (%) (n=257)	*p* value*
Vehicle Type								<.0001
Unknown (0)	19,730 (45.7%)	3,491 (27.3%)	90 (69.2%)	15,346 (53.4%)	685 (57.3%)	34 (56.7%)	84 (32.7%)	
Hatchback (1)	5,070 (11.7%)	1,313 (10.3%)	0 (0.0%)	3,535 (12.3%)	115 (9.6%)	2 (3.3%)	105 (40.9%)	
Passenger 2-Door (2)	6,394 (14.8%)	3,157 (24.7%)	12 (9.2%)	3,120 (10.9%)	103 (8.6%)	1 (1.7%)	1 (0.4%)	
Passenger 4-Door (3)	11,983 (27.8%)	4,840 (37.8%)	28 (21.5%)	6,732 (23.4%)	293 (24.5%)	23 (38.3%)	67 (26.1%)	
Transmission Type								<.0001
Automatic (1)	30,210 (70.0%)	9,411 (73.5%)	130 (100.0%)	19,588 (68.2%)	773 (64.6%)	60 (100.0%)	248 (100.0%)	
Manual (2)	12,956 (30.0%)	3,390 (26.5%)	0 (0.0%)	9,143 (31.8%)	423 (35.4%)	0 (0.0%)	0 (0.0%)	

* *p* values based on Pearson chi-square test of association.

Table 3

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)	*p* value*
Primary Fuel Type								<.0001
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								<.0001
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								<.0001
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%)	

* *p* values based on Pearson chi-square test of association.

Table 4

Pearson Correlation Coefficients

	co2TailpipeGpm	barrels08	comb08	make_id	displ	cylinders	volume	vehtype	emissionscat	transtype_id
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 resulted in a *p*-value < .0001.

Table 5

Contingency Table

	Hatchback	Passenger 2-Door	Passenger 4-Door	Unknown	Total
Gross Polluter	1.00	141.00	180.00	1152.00	1474.00
Low Emission	1820.00	703.00	2294.00	739.00	5556.00
Polluter	47.00	339.00	394.00	5119.00	5899.00
Standard	2952.00	5193.00	8819.00	12579.00	29543.00
Ultra-Low Emissior	122.00	7.00	101.00	91.00	321.00
Very-Low Emission	128.00	11.00	195.00	50.00	384.00
Total	5070.00	6394.00	11983.00	19730.00	43177.00

Figure 1

Boxplot Visualization for Annual Petroleum Consumptions in Barrels for Fuel Type 1

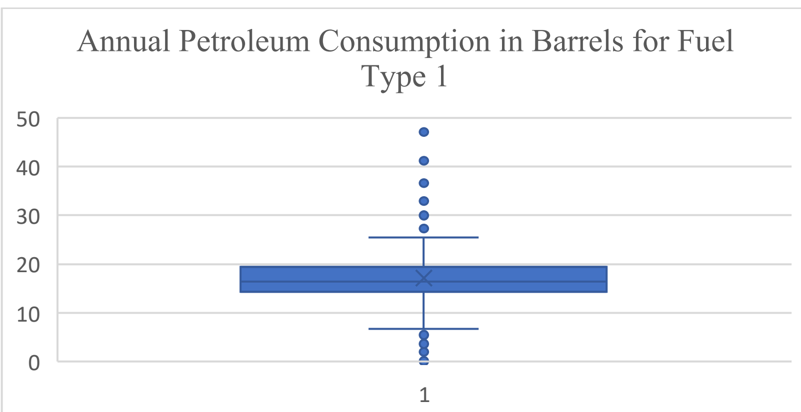


Figure 2

Probability Distribution for CO_2

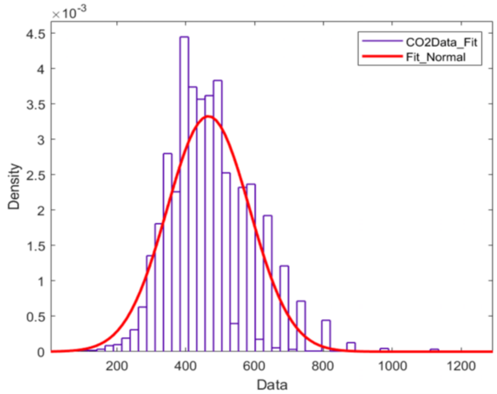


Figure 3

Originally Planned Model- Nulls Excluded

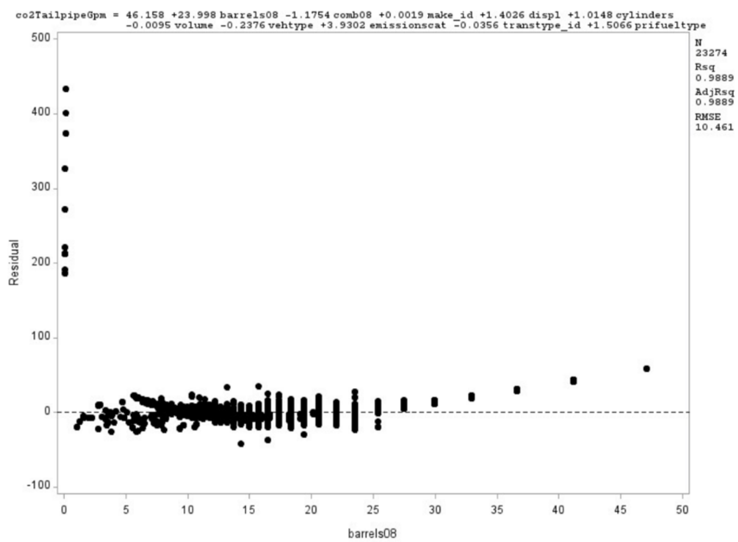
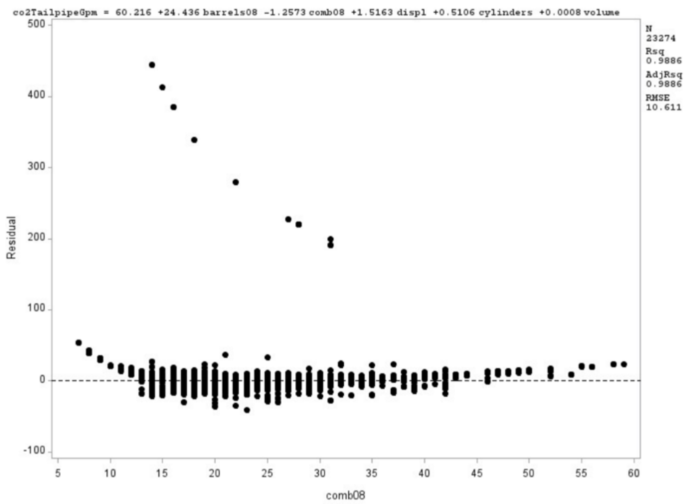


Figure 4

Trimming Regression Model- Categorical Non-Indicator Variables and Nulls Excluded



Equation 1

Probability Density Function for CO_2 Emission

$$PDF_{CO2} = f(x; \mu = 465.538, \sigma = 119.88) = \begin{cases} \frac{1}{[(\sqrt{2\pi})(119.88)]} e^{-[\frac{(x-465.538)^2}{[(2)(119.88)^2]}]} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Equation 2

Expected Generalized Formula for CO_2 Emission

$$CO_2 = \mathcal{X}_0 + \mathcal{X}_1 Petroleum\ Consumption - \mathcal{X}_2 MPG + \mathcal{X}_3\ Displacement + \mathcal{X}_4\ Cylinder + Volume\ \mathcal{X}_5 + \varepsilon$$

Equation 3

Final Formula for CO_2 Emission

$$CO_2 = 60.22 + 24.44\ \mathcal{X}_1 - 1.26\ \mathcal{X}_2 + 1.52\ \mathcal{X}_3 + .51\ \mathcal{X}_4 + .00075\ \mathcal{X}_5 + \varepsilon$$

References

Devore, J. (2016). *Probability and Statistics for Engineering and the Sciences* (9th ed.). Boston, MA: Cengage Learning.

Duan, D. (2021). *RE: Discussion 5.1: Research Project Data Discussion.* Unpublished Manuscript. University of San Diego. Retrieved from https://ole.sandiego.edu/ultra/courses/_79544_1/cl/outline

Griffiths, A., Gelbart, W., Lewontin, R., Suzuki, D., & Miller, J. (2000). *An Introduction to Genetic Analysis 7th edition*. New York: W H Freeman &. Retrieved February 27, 2021, from https://www.ncbi.nlm.nih.gov/books/NBK21907/

Holmes, A., Illowsky, B., & Dean, S. (2017). *Introductory Business Statistics*. Houston, Texas: OpenStax. Retrieved February 27, 2021, from https://openstax.org/books/introductory-business-statistics/pages/5-1-properties-of-continuous-probability-density-functions?query=probability%20density%20function&target=%7B%22index%22%3A0%2C%22type%22%3A%22search%22%7D#eip-33

Lloyd, T. (2021). *RE: Discussion 5.1: Research Project Data Discussion*. Unpublished Manuscript. University of San Diego. Retrieved from https://ole.sandiego.edu/ultra/courses/_79544_1/cl/outline

McDonald, J. (2020). 2.3: Chi-square test Of goodness-of-fit. Retrieved February 28, 2021, from https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Biological_Statistics_(McDonald)/02%3A_Tests_for_Nominal_Variables/2.03%3A_Chi-Square_Test_of_Goodness-of-Fit

Schneider, A., Hommel, G., Blettner, M. (2010) Linear regression analysis—part 14 of a series on evaluation of scientific publications. Deutsches Ärzteblatt International 2010, 107(44): 776–82. DOI: 10.3238/arztebl.2010.0776

U.S. Department of Energy. (2020, December). *FuelEconomy.gov web services*. Retrieved from https://www.fueleconomy.gov/feg/ws/index.shtml

U.S. Department of Energy. (n.d.). *Alternative Fuels and Advanced Vehicles*. Alternative Fuels Data Center: Alternative Fuels and Advanced Vehicles. Retrieved from https://afdc.energy.gov/fuels/.