Development of a linear regression model to predict carbon dioxide emissions based on key standard vehicle features

Aaron M. Carr

Shiley-Marcos School of Engineering, University of San Diego

Abstract

An examination of sample statistics and development of a linear regression model were utilized to determine whether key vehicle features (e.g., such as petroleum use, engine cylinder count, luggage volume) could predict outcomes for CO_2 emissions. The objective of this project is to provide San Diego government agency decision-makers with evidence that could be used for regulation and changes in infrastructure planning. Sample data on vehicle characteristics such as petroleum use, mechanical design, and recorded CO_2 output was collected from www.fueleconomy.gov in order to perform analyses on secondary data. After initial steps reviewing statistical measures, including consideration of associations and correlations between key variables of interest, a linear regression model was created that had nine regression coefficients. However, due to apparent issues of multicollinearity among predictor variables and a potential error introduced by categorical variables, both of which were discovered during the initial multivariate analyses, a revised regression model was developed that relies on three regression coefficients. Though the coefficient of determination on the trimmed model was lower (R^2 = .8792) than in the original model (R^2 = .9814), this was found acceptable in an attempt to mitigate overfitting of future data. The final model provided sufficient evidence, based on a 95% confidence interval (all *p*-values for the final model < .001), that an effective model could be developed to predict CO_2 based on relatively few key standard vehicle features.

Table of Contents

List of Tables

List of Figures

List of Equations

Development of a linear regression model to predict carbon dioxide emissions based on key standard vehicle features

Climate change is a global issue, with local impacts. The city of San Diego has a lot at stake as the effects of climate change become more pronounced over the coming decades, including longer and more extreme drought, poor air quality, and negative impact to its large coastal communities. The U.S. National Ocean Service (n.d.) predicts that trends in global mean sea level rise will continue in the decades to come. Increases in sea levels could have substantial effects on habitation patterns throughout both the city and the County. One of the potential drivers of climate change is the volume and usage of vehicles on the road. Transportation is big business in the U.S., especially in California where there were an estimated 32.6 million autos and trucks registered in 2019 (California Department of Motor Vehicles, n.d.). All of the vehicles that run on petroleum-based fuel emit greenhouse gases, including CO_2, as a byproduct of combustion.

Objective, Hypothesis, & Proposed Regression Model

This study will use secondary data to determine whether there is an association between primary vehicle petroleum consumption and CO_2 emissions, with the objective of providing information to key transportation policy- and decision-makers in local government agencies concerning the impact of continued petroleum use. This is part of an initiative to argue for the need to create stricter standards for vehicle emissions, and possibly by extension spur expanded development of alternative transportation methods to reduce future CO_2 emissions. The general hypothesis of this research project is that there is a strong positive association between passenger vehicle annual primary petroleum consumption and tailpipe CO_2 emissions, which will be measurably impacted by several different characteristics of the vehicles included in the sample.

The generalized object formula represents the anticipated relationship between features that individuals within the sample generally have, such that their inclusion for each individual record will

produce an output analogous to those for all other records. In other words, the model can used to perform statistical regression analyses that will provide insight into the relationship of the variables to the amount of CO_2 emissions. As the generalized formula is based on an additive model, each additional feature contributes to the final result of how much CO_2 is emitted annually. Equation 1 is the formula for the postulated hypothesis, where β_i are the regression coefficients and ϵ is the random error term that accounts for deviations between actual (x, y) values and the true regression line.

(1)

$$CO2 = \beta_0 + \beta_1 PetroleumConsumption + \beta_2 MPG + \beta_3 Cylinders + \beta_4 EngineDisplacement + \beta_5 DriveAxel \begin{pmatrix} 1 = 2\ Wheel\ Drive \\ 2 = 4\ Wheel\ Drive \\ 3 = 4\ Wheel\ Drive\ or\ All\ Wheel\ Drive \\ 4 = All\ Wheel\ Drive \\ 5 = Front\ Wheel\ Drive \\ 6 = Part\ time\ 4\ Wheel\ Drive \\ 7 = Rear\ Wheel\ Drive \end{pmatrix} + \beta_6 FuelType \begin{pmatrix} 1 = Premium\ Gasoline \\ 2 = Midgrade\ Gasoline \\ 3 = Regular\ Gasoline \\ 4 = Diesel \\ 5 = Natural\ Gas \end{pmatrix} + \beta_7 VehicleType \begin{pmatrix} 0 = Unknown \\ 1 = Hatchback \\ 2 = Passenger\ 2\ Door \\ 3 = Passenger\ 4\ Door \end{pmatrix} + \beta_8 TransmissionType \begin{pmatrix} 1 = Automatic \\ 2 = Manual \end{pmatrix} + \beta_9 LuggageDisplacement + \epsilon$$

Methods

Sample Collection, Characteristics, & Descriptive Statistics

A sample of data on vehicle fuel economy for a variety of vehicles was retrieved from www.fueleconomy.gov on December 9, 2020 and covers the period of 1984 to December 2, 2020 (U.S. Department of Energy, 2020). The original downloaded data set contained 43,177 rows. For the current

research project, 257 rows that contained a zero-value for the CO_2 emissions variable were removed—which effectively eliminated all electric vehicles; 3 rows that contained null values for the cylinder count per vehicle variable were removed; and 1,178 rows that contained null values for the drive variable were removed. Additional processing involved assigning numerical indicators for the drive axel categories. All research analyses were performed on the remaining 41,739 records, unless noted otherwise, using Microsoft Excel software.

The vehicles included in the sample use various forms of petroleum, including diesel (2.5%; *n* = 1,039), natural gas (0.1%; *n* = 60), regular gasoline (66.4%; *n* = 27,709), midgrade gasoline (0.3%; *n* = 130), and premium gasoline (30.7%; *n* = 12,801). Variables of interest to investigate in relation to tail pipe emitted CO_2 (in grams/mile [GPM]; co2tailpipegpm) include annual petroleum consumption (in barrels; barrels08); fuel efficiency (in combined miles per gallon [MPG]) for the primary fuel type (comb08); number of engine cylinders per vehicle (cylinders); engine displacement (in liters; displ); drive axel type (drive_id); primary fuel type (prifueltype); vehicle type (vehtype); transmission type (transtype_id); and vehicle luggage volume (volume).

Descriptive sample statistics for the quantitative key characteristics are included in Table 1. The measurements of location include mean ($\bar{x}$), median, and mode. The measurements of dispersion or variability include sample standard deviation (s), sample variance, standard error, range, minimum values, and maximum values. Other measures include kurtosis and skewness. Together, the measures provide a useful snapshot about the expected values and variation for each variable. For example, co2tailpipegpm has a mean (or average) of 466.33 GPM (as calculated by summing all of the values and dividing by the *n*), as well as a standard deviation of 119.95—which is the amount generally expected a value will deviate from the mean. Generally, for normally distributed data 68.3% of values are within one standard deviation from the mean ($\bar{x} \pm 1s$) and 99.7% are within three standard deviations ($\bar{x} \pm 3s$).

TABLE 1

Descriptive Statstics for Key Characteristics

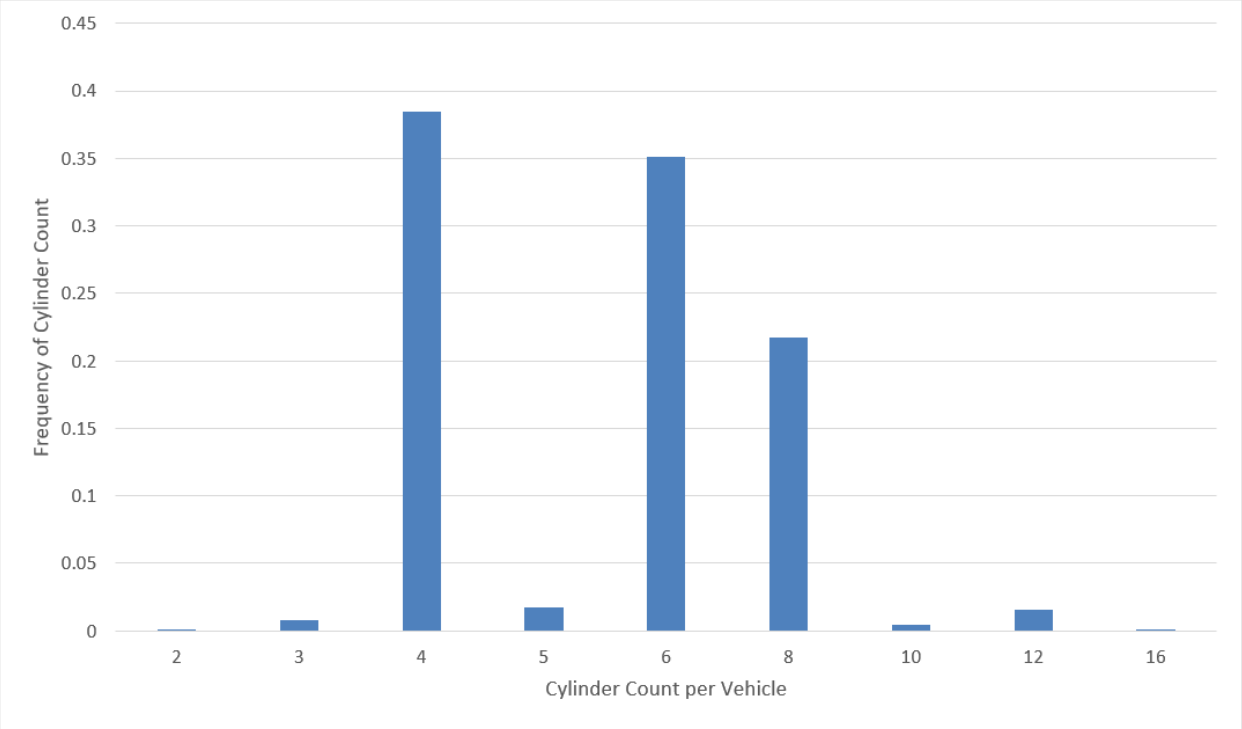
	co2tailpipegpm	barrels08	comb08	cylinders	displ	volume
Mean	466.33	17.28	20.34	5.73	3.30	64.91
Standard Error	0.59	0.02	0.03	0.01	0.01	0.33
Median	444.35	16.48	20.00	6.00	3.00	83.00
Mode	493.72	18.31	18.00	4.00	2.00	0.00
Standard Deviation	119.95	4.49	5.36	1.77	1.36	68.23
Sample Variance	14,387.85	20.18	28.68	3.12	1.84	4,656.01
Kurtosis	1.40	1.54	3.58	1.07	-0.50	-0.02
Skewness	0.71	0.63	1.20	0.89	0.65	0.67
Range	1,247.57	47.03	52.00	14.00	7.80	538.00
Minimum	22.00	0.06	7.00	2.00	0.60	0.00
Maximum	1,269.57	47.09	59.00	16.00	8.40	538.00

Note . *N* = 41,739

There are several different engine cylinder counts per vehicle included in the sample. Figure 1 shows a bar graph with the frequencies of engine cylinder counts for vehicles (which ranged from 2 to 16). The majority of vehicles had either a 4-cylinder engine (38.5%; *n* = 16,070), a 6-cylinder engine (35.1%; *n* = 14,646), or an 8-cylinder engine (21.7%; *n* = 9,054); together these three categories comprised 95.3% of the sampled vehicles. The expected value of the number of engine cylinders using the calculation for discrete random variables, $E(x) = \sum x \cdot p(x)$, is 5.73, with a standard deviation of 1.77 cylinders. Though in practicality there cannot be a non-integer value for discrete random variable integer value, this result can be interpreted as the expected number of cylinders that would be seen in a sample, based solely on probabilities.

Figure 1

Bar Graph of Cylinder Characteristic Frequencies



Bivariate Examination

Bivariate frequency distribution tables are helpful in providing useful information about the comparison of two variables with a relatively quick glance. Table 2 provides information for the vehicle type and transmission type variables relative to the primary fuel type variable. Focusing first on the population for vehicle type, there a significant proportion that are unknown. This may have a significant effect on the results, but without having additional info, only the proportions of known categories will be interpreted. Passenger 4-door has the highest proportion of the population (27.8%), followed by passenger 2-door (14.8%), and hatchback types (11.7%). As noted already, the data is also broken down by primary fuel type. When compared directly against the population, items of note include increased proportions for the 4-door type in both premium gasoline and natural gas vehicles. For the former, passenger 4-door is the highest proportion within the fuel type (37.8%), followed by the passenger 2-

door (24.7%), but the unknown category is significantly lower meaning that more vehicles with that fuel type have a known vehicle type. For natural gas, even though passenger 4-door is higher (38.3%), the other types are much lower than in the population based on the fact that the unknown type is significantly higher (56.7%). It is interesting to note that a significant proportion of the electric vehicles with the natural gas fuel type were hatchbacks (40.9%), which also resulted in a very large difference relative to the population.

TABLE 2

Characteristics of 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%)	
Variable	**(N = 43,166)**	**(n = 12,801)**	**(n = 130)**	**(n = 28,731)**	**(n = 1,196)**	**(n = 60)**	**(n = 248)**	***p*- value***
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-squared test of association.

Examination of Associations Between Key Variables

Looking at associations between features is a critical step in the early phases of exploratory data analysis. Table 3 compares the proportional distribution of emission categories against the primary fuel type, vehicle type, and transmission type features. The population proportion for premium gasoline is 29.6% relative to primary fuel type, but when compared to individual emission categories, it is almost a stair-step increase starting at ultra-low emission category (7.5%), with standard and gross polluter being higher than the population (33.2% and 32.4% respectively). Interestingly, the proportion for the polluter category is actually lower than the population (21.4%). Regular gasoline makes up the largest proportion of the population within the primary fuel type (66.5%). But all other delineated categories except ultra-low emission and standard gasoline are actually higher, indicating that there are significant differences

within each category for the very-low emission (81.0%), low emission (73.2%), polluter (73.9%), and gross polluter (67.0%) categories—regarding the ultra-low emission exception, there was a sizable decrease in proportion, which makes sense when also considering that the electricity variable proportion went from 0.6% in the population to 80.1% in the ultra-low emission category.

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%)	
Variable	**(N = 43,166)**	**(n = 312)**	**(n = 384)**	**(n = 5,556)**	**(n = 29,543)**	**(n = 5,898)**	**(n = 1,473)**	***p* -value***
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-squared test of association.

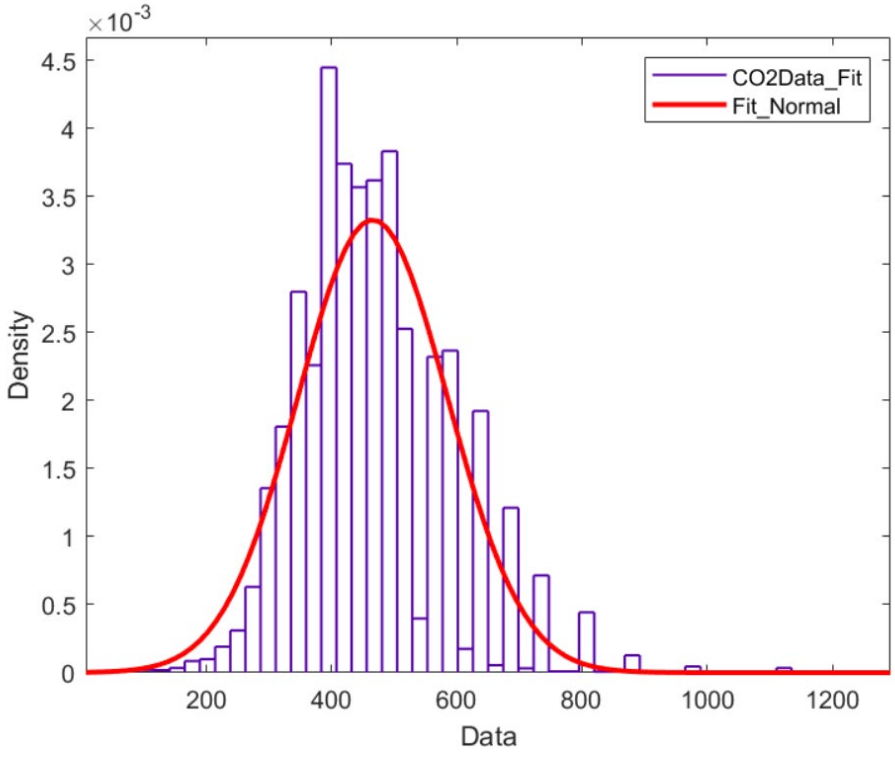
Results

CO_2 Emissions Probability Distribution

Figure 2 is the density curve graph of the CO_2 tailpipe emissions data. A normal curve has been added to the histogram to show that the bulk of the values fit within a normal distribution. The mean is the topmost of the normal curve and is approximately 475 GPM as seen by the visualization, which is indeed close to the actual mean. As a rule, the mean of a normal curve is also the median, such that there is a .5000 probability that half the values are below it. However, in this instance since there is clearly some positive skew, the mean does not exactly equal the median (444.35 GPM). As the sample size is very large, and much of the data does conform to a normal distribution, it will be assumed for the remainder of analyses that the data is normally distributed. Equation 2, which is the probability density

function (PDF) for normally distributed data, represents a density curve under which the probabilities of all values of continuous random variable X lie. The entire range under the PDF equals 1, as it covers all possible probabilities. However, because it is assumed the data has a normal probability distribution, the value of random variable X can actually be standardized to estimate the probability that it is within a certain area of the data. Where $\bar{X}$ is the mean of X and s is the sample standard deviation, this is done by converting X to Z using the formula $Z = (X - \bar{X})/s$. The probability is then denoted by $P(Z \leq z)$, and can be determined by referring to a standard normal curves lookup chart. For instance, the estimated probability that a CO_2 emission value is less than or equal to one standard deviation above the mean (466.33 + 119.95 = 586.28) can be calculated to be .8413.

Figure 2. CO_2 Data Distribution Fit to a Normal Curve



(2)

$$PDF_{CO2} = f(x;\ \mu = 466.33, \sigma = 119.95) = \begin{cases} \dfrac{1}{\left[\left(\sqrt{2\pi}\right)(119.95)\right]} e^{-\left[(x-466.33)^2/\left[(2)(119.95)^2\right]\right]}, & X \geq 0 \\ 0, & X < 0 \end{cases}$$

Correlations Between Key Variables

When examining the correlation coefficient matrix in Table 4, there are several variable pairs that stand out as having significant correlations, meaning that there is strong evidence for a linear relationship to exist between them. For instance, a strong positive correlation between CO_2 emissions (co2TailpipeGpm) and annual petroleum consumption (barrels08) is indicated (r = .9885), meaning that as values of one rise, so do values of the other. There is a strong negative correlation between CO_2 emissions and combined MPG for fuel type 1 (comb08; r = -.9184)—the higher the MPG, the lower the CO_2 emissions. Another strong correlation includes engine displacement (displ) and number of cylinders (r = .9046), which makes sense given that more cylinders ostensibly will require a heavier engine. Conversely, vehicle volume does not have very high correlations to any of the other variables, except for with vehicle type (vehtype), which is still only moderate (r = .7418)—the next highest correlation for volume is with CO_2 emissions, but it is weak (r = -.4323).

TABLE 4

Pearson Correlation Coefficients (N = 42,917)

	co2TailpipeGpm	barrels08	comb08	make_id	displ	cylinders	volume	vehtype	emissionscat	prifueltype
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 -values < .0001.

.1128), which does not necessarily indicate that there is no relationship, just that if there is, it is not linear.

Test for Variable Independence

Comparison of two discrete variables can be accomplished with a two-way contingency table, such as Table 5, which includes emission group as the rows (designated by random variable I) and vehicle type as the columns (random variable J). By computing expected values of cells within the table, tests for heterogeneity and independence can be performed. The latter applies to this dataset since the focus is on two different factors from a single population. The test begins by stating the null hypothesis that the two factors are independent (H_0: $p_{ij} = p_i \cdot p_j$), and then a chi-squared test is performed to determine the *p*-value to compare against the level of significance (α)—equation 3 is the formula to compute chi-squared (χ^2). Several criteria must be met for the test statistic to be useful, including the assumptions that n_{ij} represents the number of individuals in the (i, j)th cell of the table and the estimated expected value ($\hat{e}_{ij}$) of all cells must be greater than or equal to five. The calculated χ^2 using the emission group and vehicle type variables was 9,017.56, and a *p*-value < .001 was calculated based on $(I - 1)(J - 1)$ degrees of freedom. The results indicate that the null hypothesis should be rejected, but it can also be argued that the test is inconclusive based on the calculated *p*-value alone. Additional analyses would need to be performed to tell whether the two variables are indeed dependent.

Table 5

Two-Way Contingency Table Including Output of Chi-Squared Test

		Hatchback	Passenger 2-door	Passenger 4-door	Unknown	Total
Gross Polluter	**Frequency**	1.00	137.00	167.00	1,150.00	1,455.00
	Expected	159.62	210.41	401.83	683.14	
	Chi-squared	157.63	25.61	137.23	319.05	639.52
Low Emission	**Frequency**	1,691.00	686.00	2,255.00	734.00	5,366.00
	Expected	588.68	775.99	1,481.92	2,519.41	
	Chi-squared	2,064.13	10.44	403.30	1,265.25	3,743.11
Polluter	**Frequency**	37.00	318.00	368.00	5,110.00	5,833.00
	Expected	639.91	843.53	1,610.89	2,738.67	
	Chi-squared	568.05	327.41	958.96	2,053.26	3,907.68
Standard	**Frequency**	2,705.00	4,878.00	8,508.00	12,547.00	28,638.00
	Expected	3,141.75	4,141.43	7,908.92	13,445.91	
	Chi-squared	60.71	131.00	45.38	60.10	297.19
Ultra-low Emission	**Frequency**	17.00	6.00	34.00	7.00	64.00
	Expected	7.02	9.26	17.67	30.05	
	Chi-squared	14.18	1.14	15.08	17.68	48.09
Very-low Emission	**Frequency**	128.00	11.00	195.00	49.00	383.00
	Expected	42.02	55.39	105.77	179.82	
	Chi-squared	175.95	35.57	75.27	95.18	381.97
TOTAL	**Frequency**	4,579.00	6,036.00	11,527.00	19,597.00	41,739.00
	Expected	4,579.00	6,036.00	11,527.00	19,597.00	
	Chi-squared	3,040.65	531.18	1,635.21	3,810.52	9,017.56

Note. $p < .0001$

(3)

$$\chi^2 = \sum_{all\ cells} \frac{(observed - estimated\ expected)^2}{estimated\ expected} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - \hat{e}_{ij}\right)^2}{\hat{e}_{ij}}$$

Multicollinearity

Table 6 shows the output of a linear regression analysis based on the original generalized model. Though the coefficient of determination (R^2 = .9814) indicates strongly that most of the amount of observed y variation can be explained by use of the multiple regression model, there is potentially an issue with multicollinearity as indicated by the variance inflation factor (VIF) values. Multicollinearity occurs when there are relationships present between some or all of the predictor variables ($x_1, x_2, \ldots, x_n$) that impact observed values, in addition to the relationships each explanatory variable has with the dependent variable y (Devore, 2016, p. 606). Four of the variables exceed the VIF threshold

of five, while the barrels08 and displ variables exceed seven (7.13 and 7.24, respectively), which warrants further examination into whether multicollinearity exists (Cody & Smith, 2006). While none of the VIF values are above 10, there is a very strong indication that some independent variable values can be predicted from other independent variables. This is also visible in Table 4, which shows strong correlations between barrels08 and combo08 (r = -.9050) as well as cylinders and displ (r = .9046).

Table 6

Original Planned Linear Regression Model Measures

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Variance Inflation Factor*
Intercept	122.20	1.48	82.63	0.0000	119.30	125.10	
barrels08	21.94	0.05	460.95	0.0000	21.85	22.04	7.13
comb08	-3.06	0.04	-85.27	0.0000	-3.13	-2.99	5.76
cylinders	2.03	0.12	17.68	0.0000	1.81	2.26	6.44
displ	2.13	0.16	13.42	0.0000	1.82	2.44	7.24
drive_id	0.06	0.05	1.25	0.2118	-0.04	0.16	1.04
prifueltype	3.22	0.09	34.83	0.0000	3.04	3.40	1.22
vehtype	-0.40	0.10	-4.04	0.0001	-0.59	-0.21	2.53
transtype_id	0.87	0.18	4.79	0.0000	0.51	1.22	1.07
volume	0.00	0.00	-2.26	0.0239	-0.01	0.00	2.55

Note. N = 41,739

Regression Model Comparisons

Initial Model

After performing the regression analysis that resulted in the output in Table 6, equation 4 was achieved. As noted, the coefficients of determination indicate that the model is a strong fit for the values of x_n to predict y, however having nine parameters (k = 9) makes the model overly complicated and not very easy to interpret. If a similar level of CO_2 emissions variation explanation could be achieved with fewer parameters, that model would be preferred (Devore, 2016, p. 599). Also, reducing specific key characteristics may be desired to address the issues both with multicollinearity, as well as those arising from recoding categorical variables—see the Discussion Section for further details.

(4)

$$CO2 = 122.20 + (21.94)PetroleumConsumption - (3.06)MPG + (2.03)Cylinders + (2.13)EngineDisplacement + (0.06)DriveAxel + (3.22)FuelType - (0.40)VehicleType + (0.87)TransmissionType - (0.00)LuggageDisplacement + \epsilon$$

Final Model

The final model was based on trimmed regression model with only three parameters, including fuel efficiency (combo08), engine displacement (displ), and engine volume. The coefficient of determination (R^2 = .8792) was lower than the originally planned model. Moreover, the mean square error (MSE), which is a measure of average deviation of values within the entire sample, was significantly higher in the trimmed model than the original (1,737.49 vs. 267.56). Along with the coefficients, Table 7 provides the following measures for the model: t test statistic values, *p*-values, the upper and lower bounds based on a 95% confidence level, and VIF values. In order to assess the model's utility, a statistical test was done to determine whether at least one regression parameter did not equal zero and would therefore be associated with predicting outcomes for CO_2 emission values. The test statistic for the final model (*f* = 101,296.90) resulted in a *p*-value < .001, leading to us to reject the null hypothesis (H_0: $\beta_1 = \beta_2 = \beta_3 = 0$) at the 95% confidence interval in favor of the alternative, i.e., at least one of the regression coefficients has an appreciable effect on the dependent variable.

Table 7

Final Linear Regression Model Measures

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Variance Inflation Factor*
Intercept	720.00	1.76	408.63	0.0000	716.54	723.45	
comb08	-15.92	0.06	-275.30	0.0000	-16.04	-15.81	2.30
displ	22.81	0.22	102.61	0.0000	22.38	23.25	2.19
volume	-0.08	0.00	-24.21	0.0000	-0.09	-0.07	1.22

Note. N = 41,739

Discussion

Conclusion

The final regression model that was chosen (see equation 5) significantly reduced the number of parameters and was able to reduce all of the VIF values below five, which eliminated the issues of multicollinearity that were present in the original regression model. Additionally, it does not include any of the qualitative variables, thereby eliminating the issue related to incorrect implementation of recoding with indicators procedures. As noted, the R^2 was lower than in the original model, but this was done intentionally to avoid an issue with overfitting; this occurs when a model does really well at predicting outcomes from the current data, but as soon as new observation is added, it becomes less effective at predicting independent variable values (Bruce et al., 2020, p. 236).

(5)

$$CO2 = 720.00 - (15.92)MPG + (22.81)EngineDisplacement - (0.08)LuggageDisplacement + \epsilon$$

As seen from Table 4, there is a strong correlation between CO_2 emission and petroleum use (*r* = .9885). However, it did not factor into the final chose model, as it was causing potential issues of multicollinearity with fuel efficiency. Though this strictly means that we cannot accept our initial hypothesis completely, a model was achieved that provides strong evidence that it can provide an

effective prediction mechanism for values of CO_2 emissions based on inputted values of key independent values. Not overfitting the model also argues for being able to apply it to either multiple samples, or to completely new data, to achieve analogous results. An interpretation of equation 5 is that as the values for MPG and volume decrease, CO_2 emissions increase, though for the latter the change is very slight. For engine displacement, as it increases (which ostensibly results in more space and weight), CO_2 emissions increase. The direction of effect for MPG and engine displacement make sense intuitively based on the nature of what the variables represent: higher MPG means a more efficient engine, and therefore lower emissions, while more engine displacement means higher weight and therefor higher emissions.

Study Strengths & Weaknesses

This study has several strengths and weaknesses. One of the key aspects, that represents both is the use of a pre-existing data set that was collected by investigators other than those performing the current research. This can be a strength from the stand-point of efficient resource allocation. The cost in relation to this paper was zero, which meant all of the time necessary to complete the research could be devoted to statistical analysis. On the other hand, the drawback is that at this point, it is impossible to determine whether there are issues contained in the data that arose from collection or sampling bias. Moreover, the current research must assume that all values were recorded correctly and consistently. This is a tall ask, especially given the size of the dataset and how many variables there are. That is a lot of data points to grasp fully.

Additionally, there were some key assumptions made during the processing and analyses of the data that have potentially introduced errors. These include: The categories (c) of the qualitative variables were not recoded correctly, such that the number of dummy indicators are not based on the correct methodology of using $c - 1$; as noted in the in the Sample Collection & Characteristics section of the Methods, calculating measurements of location and dispersion on discrete variables introduces

ambiguity and error, since often the results will be fully representative of values with the possible range; and finally a lot of the analyses performed was based on the assumption that all data was normally distributed. To this last point, though the number of observations included in the sample is a generally acceptable basis to assume normality, the limitations of that assumption are seen in Figure 2.

Overall the amount of observations analyzed in the data set (N = 41,739) represents an overall strength. Having more data to examine increases the chances of being able to determine whether meaningful patterns exist in the data, given certain conditions are met. Additionally, where it is determined that the data represent a sample of a population, having more data can enable more accurate estimations of population parameters. It could be argued that having too many records would make it more likely to conclude statistical significance between the point estimate and hypothesized value when there is no practical significance—this is due to the fact that as n becomes increasingly larger, the standard error that resides in the denominator of the calculation of p-values get smaller (Devore, 2016, p. 352). However, to combat a potential misrepresentation of actual significance, there are other measurements that can be reported along with p-values that give a broader picture—allowing readers to make their own interpretations of the data—which include confidence intervals and effect sizes. As Kaplan et al. (2014) point out, there are other issues that can arise from examining large datasets, including retrospective bias, in which no hypotheses are established before analyzing the data; multiple comparisons bias, which results from having so many variables to compare that eventually something statistically significant will be determined; and measurement errors, which introduce issues related to missing or incorrectly entered data. The first two issues are more easily controlled by making sure hypotheses and variables of interest are determined prior to doing exploratory data analysis, as well as being clear and up-front about the process in the research write-up. The last potential issue, related to actual data values issue, is harder to fully address, especially in the case of an observational study. To off-set impact, the investigator must make every effort possible to make sure they understand

the data and must attempt to both understand and clearly communicate the implications of measurement pitfalls.

References

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical Statistics for Data Scientists* (2nd ed.). O'Reilly Media.

California Department of Motor Vehicles. (n.d.). *Estimated vehicles registered by County for the period of January 1 through December 31, 2019*. https://www.dmv.ca.gov/portal/uploads/2020/06/2019-Estimated-Vehicles-Registered-by-County-1.pdf

Cody, R. P., & Smith, J. K. (2006). *Applied Statistics and the SAS Programming Language* (5th ed.). Pearson Prentice Hall.

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

Kaplan, R. M., Chambers, D. A., & Glasgow, R. E. (2014). Big data and large sample size: A cautionary note on the potential for bias. *Clinical and Translational Science*, *7*(4), 342-346. https://doi.org/10.1111/cts.12178

U.S. Department of Energy. (2020, December). FuelEconomy.gov web services. Retrieved December 9, 2020, from https://www.fueleconomy.gov/feg/ws/index.shtml

U.S. National Ocean Service. (n.d.). *How is sea level rise related to climate change?* Department of Commerce, National Oceanic and Atmospheric Administration. Retrieved October 18. 2021, from https://oceanservice.noaa.gov/facts/sealevelclimate.html