The Clean Air Act and Auto Manufacturing Long-Run Production Strategy

Ryan S. Dunn

Shiley-Marcos School of Engineering, University of San Diego

Abstract

This study was conducted to develop a model that can provide vehicle manufacturers' a method to better understand the impact of car characteristics on emissions output, so they can make better informed decisions on which vehicles to prioritize for long-run production. The data was collected from the FuelEconomy.gov website via the U.S. Department of Energy and consisted of an initial total of 43,177 data points which, when plotted on a probability density curve, exhibited the characteristics of a normal distribution with a mean CO2 emission grams per mile of 465.538 and a standard deviation of 119.88. The initial model hypothesized that to predict tailpipe CO2 in grams per mile, we could examine annual petroleum consumption in barrels, combined MPG, manufacturer, engine displacement, engine cylinders, vehicle volume, vehicle type, emissions category (derived value from CO2 tailpipe emissions), transmission type, and primary fuel type. Evaluation of the original model's regression output provided an R-Squared value of 0.9829 and parameter estimates that were significant at the <.0001 level, however with high variance inflation factors, improperly incorporated categorical variables, and null values within the dataset, a more robust model was necessary. By removing null values and nonsensical categorical variables, the final regression model carried an R-Squared value of 0.9886 and Root MSE of 10.6, with all explanatory variables except volume having a p-value <.0001. At the 95% confidence level, therefore, we can assert that the multiple regression coefficient for annual petroleum consumption is 24.43642 ± 0.0923, combined MPG is -1.25730 ± 0.0570, engine displacement is 1.51626 ± 0.2975, engine cylinders is 0.51063 ± 0.1944, and volume is 0.00078 ± 0.0032. This allowed us to reject the null hypothesis in support of the alternate hypothesis, supporting the claim that there is a statistically significant relationship between annual petroleum consumption, MPG, displacement, and cylinders to tailpipe CO2 emissions.

Table of Contents

Abstract 2

List of Tables 4

List of Equations 5

The Clean Air Act and Auto Manufacturing Long-Run Production Strategy 6

Hypothesis and Objective of Study 6

Method 8

Bivariate Frequency Table 8

Associations of Emissions Categories 9

Results 10

Probability Density Function 10

Statistical Correlation and Lack of Correlation 12

Chi-Square Test for Independence and Homogeneity 13

Multicollinearity Concerns 15

Original Regression Model 15

Final Regression Model 16

Discussion 17

Strengths and Weaknesses of Model 18

Implications for Future Research 19

References 20

Statement of Peer Review 21

List of Tables

Table 1 Descriptive Statistics 8

Table 2 Bivarite Frequency Table 9

Table 3 Association Between Emission Category by Fuel Type, Vehicle Type, and Transmission Type 10

Table 4 Pearson Correlation Coefficients 13

Table 5 Two-Way Contingency and Chi-Squared Table 14

List of Equations

Equation 1 Generalized Object Formula 7

Equation 2 Probability Density Function 10

Equation 3 Chi-Squared Formula 15

Equation 4 Final Regression Model 16

The Clean Air Act and Auto Manufacturing Long-Run Production Strategy

The Clean Air Act (CAA) was introduced in 1970 with the intentions of combating an increasing threat of air pollution issues by providing a framework of environmental policy measure for the nation to build upon (United States Environmental Protection Agency, 2021). Revised in 1990 and with overwhelming bipartisan support, the CAA has allowed for pronounced progress in achieving national air quality standards. Displaying a clear commitment to reducing America's environmental footprint, on his first day in office, President Biden signed an executive order on protecting public health and the environment and restoring science to tackle the climate crisis. Therefore, the need for a thorough analysis of automakers' products and subsequent air emissions is essential not only to address public health and world pollution issues but also to enable automakers to make smart, environmental-driven production decisions. As the CAA allows for the trading of emission credits from companies with low pollution costs to those facing higher pollution costs (United States Environmental Protection Agency, 2021) reducing the firms' emissions rates can directly affect profitability. Through the application of data, science, and statistical-based research, the purpose of this paper is to provide a thorough analysis of products and emissions to automakers that can be leveraged in strategic planning sessions and help guide automakers to align their strategy to that of our nation's public health and environmental policy.

Hypothesis and Objective of Study

This study aims to understand the relationship between carbon dioxide emissions and various vehicle characteristics that will enable car manufacturers to make smart, long-run decisions in production. Upon reviewing the variables within the dataset, the following displays the hypothesis test:

$H_0 = \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9 = 0$

$H_a = \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9 \neq 0$

The null hypothesis states that the regression coefficient of each characteristic is 0. If there is sufficient evidence to reject the null hypothesis, it can then be inferred that there is evidence to support the assertion that the identified vehicle characteristics within the model that have a non-zero regression coefficient are related to emissions output. Automakers can then target these characteristics for research and development and stay ahead of changing environmental legislation with respect to emissions.

Equation 1 is the original multiple linear regression model used to assess the relationships between CO2 emissions and various characteristics within the FuelEconomy.gov dataset. By conducting a multiple regression analysis with the below parameters, we can develop a linear equation that can be used to then determine the effect that each variable has on the output of CO2 emissions.

$$\begin{aligned} CO_2 = {} & \beta_0 + \beta_1 barrels08 + \beta_2 combo + \beta_3 make_{id} + \beta_4 displ + \beta_5 cylinders + \beta_6 volume \\ & + \beta_7 vehtype \begin{pmatrix} 0 = Unknown \\ 1 = Hatchback \\ 2 = Passenger\ 2 - Door \\ 3 = Passenger\ 3 - Door \end{pmatrix} + \beta_8 emissionscat \\ & + \beta_9 prifueltype \begin{pmatrix} 1 = Premium\ Gasoline \\ 2 = Midgrade\ Gasoline \\ 3 = Regular\ Gasoline \\ 4 = Diesel \\ 5 = Natural\ Gas \\ 6 = Electricity \end{pmatrix} + \varepsilon \end{aligned} \tag{1}$$

β_0 : y-intercept

ε : error term

Method

The data was retrieved from the U.S. Department of Energy and contained fuel economy data of automobiles from 1984 to current model year vehicles through December 2, 2020. Vehicle characteristics included drive axel type, fuel type, transmission type, EPA vehicle size class, year, make and model. Cylinders of vehicles surveyed included 16-cylinder (0.03%), 12-cylinder (1.55%), 10-cylinder (0.41%), 8-cylinder (21.35%), 6-cylinder (34.42%), 5-cylinder (1.79%), 4-cylinder (38.96%), 3-cylinder (0.74%), 2-cylinder (0.14%), and unknown (0.60%). Top vehicle manufacturers were Chevrolet (9.62%), Ford (8.03%), Dodge (6.09%), GMC (6.01%), and Toyota (5.07%).

Table 1

Descriptive Statistics

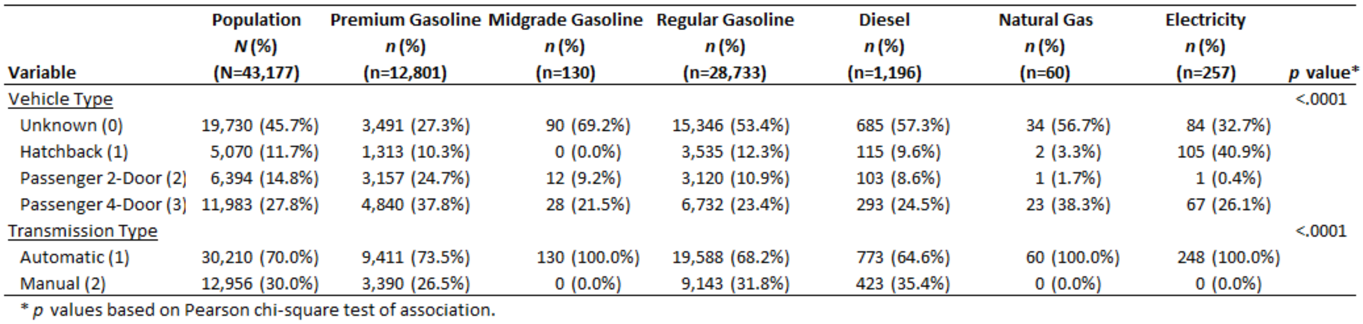
Variable	Mean	Standard Error	Median	Standard Deviation	Variance	Skewness	Range	Min	Max	Sum	Count
Barrels08	17.15	0.02	16.48	4.66	21.74	0.37	47.03	0.06	47.09	740622.26	43177
Co2TailpipeGpm	462.76	0.60	444.35	124.77	155568.05	0.41	1269.57	0	1269.57	19980884	43177
Comb08	20.85	0.04	20	8.22	67.64	6.35	134	7	141	900064	43177
Cylinders	5.71	0.01	6	1.76	3.11	0.90	14	2	16	245181	42917
Displ	3.29	0.01	3	1.36	1.84	0.66	8.4	0	8.4	141063.5	42919

Bivariate Frequency Table

Table 2 displays a bivariate frequency table that compares four vehicle and two transmission types across six sub-categories of fuel types. For the 43,177 vehicle type data points examined, 28,733 use regular gasoline, 12,801 use premium gasoline, 1,196 use diesels, 257 are electric, 130 use midgrade gasoline, and 60 are natural gas. Unknown vehicle types make up most of the vehicle types with 19,730 (45.7%) data points, followed by passenger 4-Door with 11,983 (27.8%), passenger 2-Door with 6,394 (14.8%), and hatchbacks with 5,070 (11.7%). Notable relationships are passenger 4-Door vehicle types accounting for 38.3% of natural gas types and 37.8% of premium gasoline types despite accounting for only 27.8% of the population, hatchbacks accounting for 40.9% of electric vehicles though only accounting for 11.7% of the population, and passenger 2-Door making up 24.7% of premium gasoline despite being only 14.8% of the population. For the 43,166 data points in transmission type, 100% of midgrade gasoline, natural gas, and electric fuel types were automatic transmission types, versus premium gasoline, regular gasoline, and diesel being 73.5%, 68.2%, and 64.6, respectively.

Table 2

Bivariate Frequency Table



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.

Associations of Emissions Categories

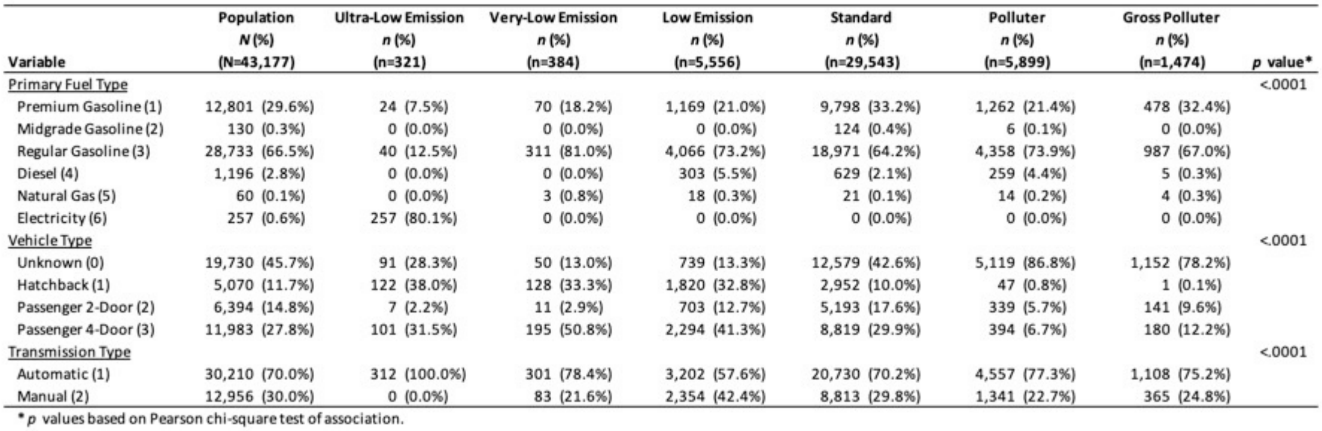
Table 3 displays the associations of primary fuel type, vehicle type, and transmission type to 6 emissions sub-categories for the 43,177 data points within the dataset. Standard emissions output account for the largest proportion of emission sub-categories with 29,543 vehicles, followed by polluter (n = 5,899), low emission (n = 5,556), gross polluter (n = 1,474), very-low emissions (n = 321), and ultra-low emissions (n = 321).

For the ultra-low emission sub-category, 80.1% of the primary fuel types were electric, whereas diesel, natural gas, and midgrade gasoline all accounted for 0% of ultra-low emission vehicles. Hatchback vehicles accounted for the largest subset of vehicle types with ultra-low emissions with 38.0%, and all 312 ultra-low emissions vehicles had an automatic transmission type.

Gross polluter emissions were generally distributed between regular gasoline (67.0%) and premium gasoline (32.4%), with all other primary fuel types accounting for <1.0%. The unknown vehicle type had the most gross polluter vehicles (78.2%), and only a single hatchback vehicle (0.1%) was deemed a gross polluter. Manual transmission types accounted for 24.8% of gross polluters.

Table 3

Associations Between Emissions Category by Fuel Type, Vehicle Type, and Transmission Type



	Population	Ultra-Low Emission	Very-Low Emission	Low Emission	Standard	Polluter	Gross Polluter	
	N (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	
Variable	(N=43,177)	(n=321)	(n=384)	(n=5,556)	(n=29,543)	(n=5,899)	(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.

Results

Probability Density Function

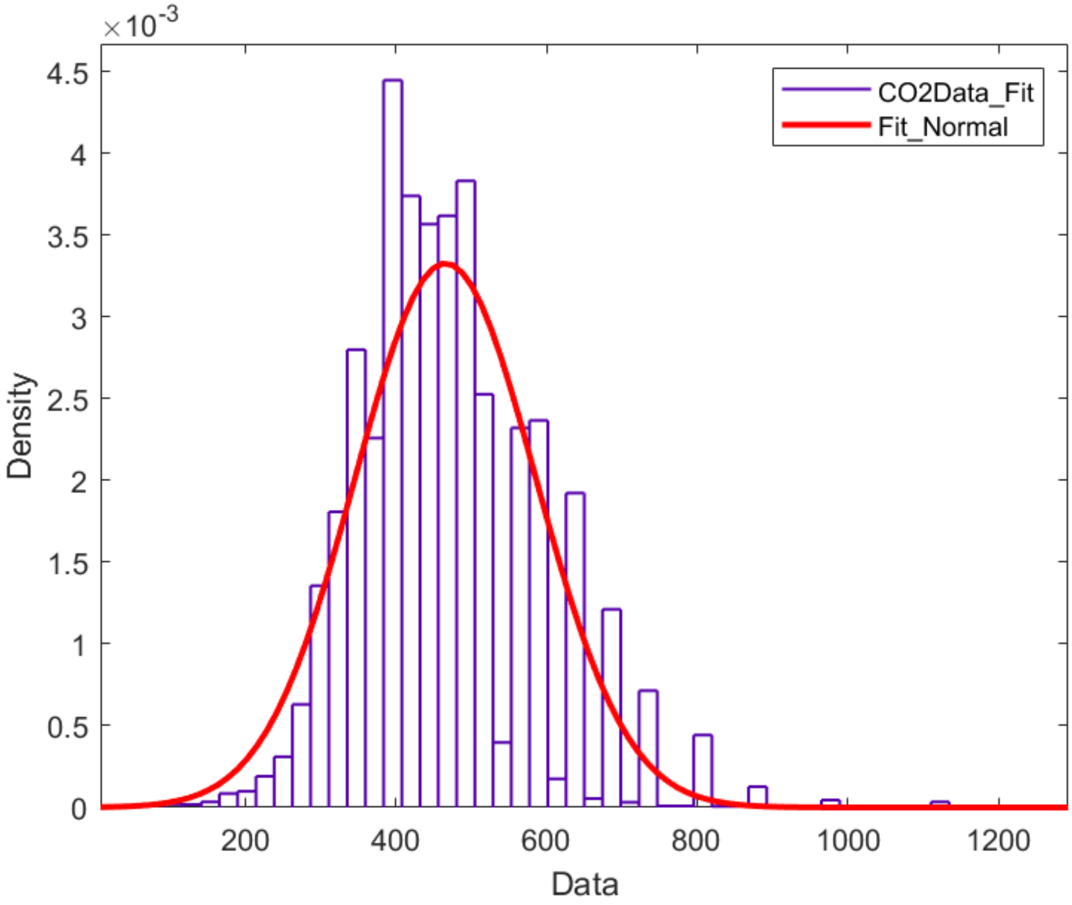
$$PDF_{CO2} = f(x, \mu = 465.538, \sigma = 119.88) = \begin{cases} \dfrac{1}{[(\sqrt{2\pi})(119.88)]} e^{-[(x-465.538)^2/[(2)(119.88)]^2]} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (2)$$

Figure 1 depicts the Probability Density Function (PDF) used for the CO2PipelineGpm variable. The data is normally distributed, therefore, we utilize the associated Normal Probability

Density Function, where $\pi \approx 3.14159$, $e \approx 2.71828$, the mean (μ) is 465.538, and the standard deviation (σ) is 119.88. Since no vehicle in the dataset contains a CO2PipelineGpm value less than zero, $\mathrm{P}(x < 0) = 0$, and the remaining cumulative probability that x takes on a value in the interval is 1.

Figure 1

Probability Density Function



The PDF depicted in Figure 1 plots the distribution of the CO2TailpipeGpm variable within the FuelEconomy.gov data set. The x-axis depicts the value of the corresponding CO2TailpipeGpm value in the dataset, and the y-axis displays the density of each corresponding bin along the x-axis. Consisting of 52 bins total, each bin represents a collection of data points within a pre-defined range. There are 43,177 data points within the original dataset; however, when accounting for null values, only 42,919 data points exist. Accounting for the removal of these 198 zero values, the mean (μ) occurs at 465.538, and the standard deviation (σ) is 119.88. Because the data is normally distributed, roughly 68% of the data exists within one standard deviation, 95% exists within two standard deviations, and 99.7% exists within three standard deviations (Bruce et al., 2020).

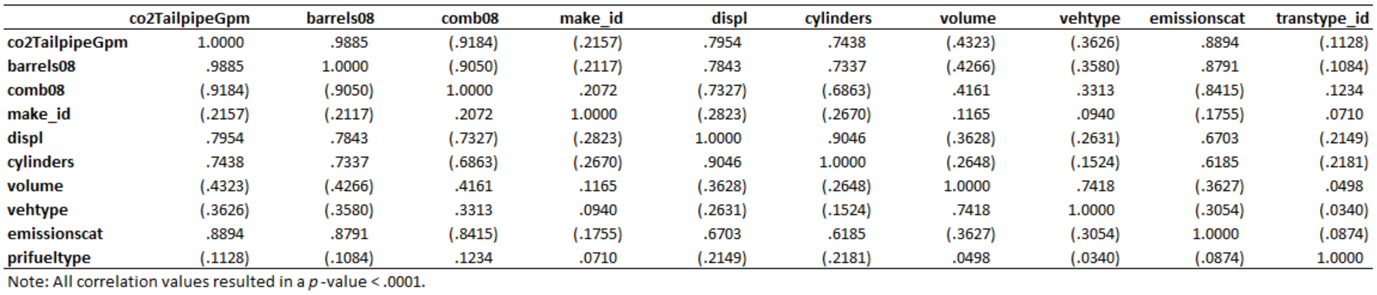
Statistical Correlation and Lack of Correlation

Table 4 displays the Pearson correlation coefficients amongst the vehicle characteristics within the emissions data set. Correlation coefficient values range from -1 to 1, where a value of -1 indicates a perfect, inverse linear relationship, 1 indicates a perfect, positive linear relationship, and 0 indicates an absence of a linear relationship where variables are said to be uncorrelated (Devore, 2016). According to Devore (2016), a weak correlation exists when the correlation coefficient r is $-.5 \leq r \leq .5$, moderate when either $-.8 < r < .5 \ or \ .5 < r < .8$, and strong when either $r \geq .8 \ or \ r \leq -.8$. Coefficients must be interpreted with caution, however, as a weak correlation coefficient does not indicate a total lack of a relationship, only a lack of a linear relationship (Devore, 2016). With all correlation values resulting in a p-value <.0001, the correlation coefficients are statistically significant.

Within the emissions dataset, strong linear relationships exist between barrels08 and CO2TailpipeGpm (.9885) and emissionscat and CO2TailpipeGpm (.8894). It can then be inferred that as annual petroleum consumption and the derived value based on classified CO2 tailpipe emission increase, tailpipe CO2 in grams per mile follows. A moderate positive relationship exists between displ (.7954) and cylinders (.7438) to CO2TailpipeGpm, which intuitively aligns with the assumption that as an engine size increases, so does the output of carbon dioxide of that engine. Conversely, a strong negative linear relationship exists between combo08 and CO2TailpipleGpm (-.9184), indicating that as the miles per gallon of a vehicle increases, carbon dioxide output decreases. Uncorrelated variables include make_id, volume, vehtype, and prifueltype variables, as a weak negative linear relationship exists with respect to CO2TailpipeGpm.

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.

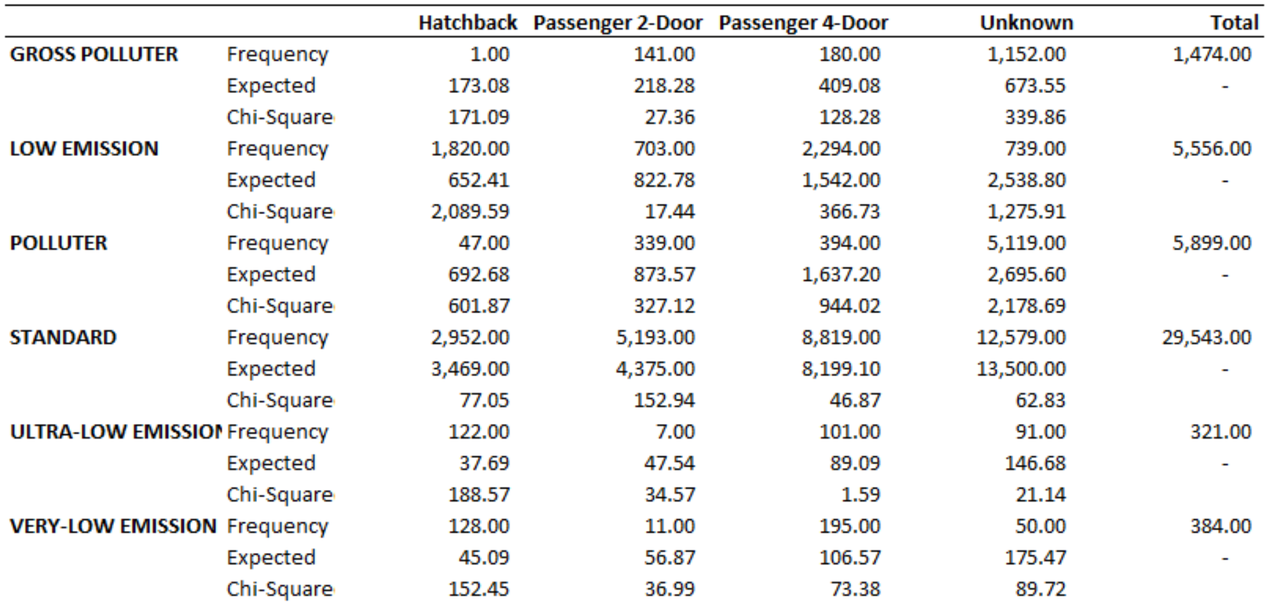
Chi-Square Test for Independence and Homogeneity

The Person Correlation Coefficient derived for emissions category and vehicle types was -.0340, which suggested no relationship. Further analysis was conducted to analyze this relationship using a Chi-Squared test for independence using a two-way contingency table, where the Chi-Squared statistic is a measurement of the extent that data is different from an expected value (Bruce et al., 2020). Two tests were conducted using the Chi-Squared statistic, the test of homogeneity, and the test of independence. The test of homogeneity is performed on two samples of whether they came from populations with similar distributions where the null hypothesis asserts that the population distributions are similar, and the test for independence to test statistical independence of attributes, where the null hypothesis asserts that the categories are independent of one another. For both tests, the expected cell count for all populations of interests and categories is greater than five.

The test for independence is useful in evaluating the relationship between two different factors in a single population (Devore, 2016). In Table 5, the two-way contingency table is constructed, displaying the observed occurrences, expected occurrences, and the test statistic value for each intersection. The test statistic value of the Chi-Squared statistic is 9,406 with 15 degrees of freedom, where the degrees of freedom are determined by multiplying the number of rows minus one by the number of columns minus one. For a p-value of .001 and with 15 degrees of freedom, our test statistic must be more than 37.69 to allow us to reject the null hypothesis of both the test for homogeneity and independence. As the Chi-Statistic is 9,406, we can reject both null hypotheses in favor of the alternate hypotheses that the distributions of the variables are not the same for each subgroup, and that the variables are indeed dependent.

Table 5

Two-Way Contingency and Chi-Squared Table



		Hatchback	Passenger 2-Door	Passenger 4-Door	Unknown	Total
GROSS POLLUTER	Frequency	1.00	141.00	180.00	1,152.00	1,474.00
	Expected	173.08	218.28	409.08	673.55	-
	Chi-Square	171.09	27.36	128.28	339.86	
LOW EMISSION	Frequency	1,820.00	703.00	2,294.00	739.00	5,556.00
	Expected	652.41	822.78	1,542.00	2,538.80	-
	Chi-Square	2,089.59	17.44	366.73	1,275.91	
POLLUTER	Frequency	47.00	339.00	394.00	5,119.00	5,899.00
	Expected	692.68	873.57	1,637.20	2,695.60	-
	Chi-Square	601.87	327.12	944.02	2,178.69	
STANDARD	Frequency	2,952.00	5,193.00	8,819.00	12,579.00	29,543.00
	Expected	3,469.00	4,375.00	8,199.10	13,500.00	-
	Chi-Square	77.05	152.94	46.87	62.83	
ULTRA-LOW EMISSIOI	Frequency	122.00	7.00	101.00	91.00	321.00
	Expected	37.69	47.54	89.09	146.68	-
	Chi-Square	188.57	34.57	1.59	21.14	
VERY-LOW EMISSION	Frequency	128.00	11.00	195.00	50.00	384.00
	Expected	45.09	56.87	106.57	175.47	-
	Chi-Square	152.45	36.99	73.38	89.72	

$$\chi^2 = \Sigma \frac{(observed - estimated)^2}{estimated\ expected} \tag{3}$$

Multicollinearity Concerns

Predictor variables are often highly interdependent in multiple regression data sets – when these relationships exist and independent variables can be predicted by other independent variables, the data is said to exhibit multicollinearity (Devore, 2016). Therefore, this can make it difficult to interpret regression coefficients. Within the emissions data set, multicollinearity can be observed when reviewing the linear regression – originally planned model where $R^2 = .9829$ and high variance inflation factors for barrels08 (9.34681), displ (7.29952), and combo08 (6.08008) exist, where these variables are inflated by the associated variance inflation factor, due to being highly correlated to at least one other variable within the model. Additional exploration of the data reveals that the volume calculation is comprised of elements of the vehtype, and the emissionscat being a derived value based on classified CO2 tailpipe emissions. This allows volume and emissionscat, which are both independent variables in the model, to be predicted by other independent variables. This assumption was validated by the correlation coefficient of volume to vehtype being .7418 and emissioncat and CO2 tailpipe emissions being .8894, and both with p-values <.0001, noting a statistically significant relationship.

Original Regression Model

The originally planned linear regression model aimed to predict tailpipe CO2 in grams/mile by modeling annual petroleum consumption in barrels, combined MPG, manufacturer, engine displacement, engine cylinders, vehicle volume, vehicle type, emissions category (derived value from CO2 tailpipe emissions), transmission type, and primary fuel type. By containing many observations used (N = 42,917), and an R-Squared value of 0.9829, corresponding to 98.29% of the proportion of variance being able to be explained by the model, the model initially appeared to be an excellent choice to predict CO2 emissions. However, with further examination and as discussed previously, the originally planned linear regression model suffered from various issues, such as the multicollinearity concerns discussed in the paragraph above. Values that were not recorded or not applicable were coded with a zero or negative one. The initial model also included categorical variables that had not been properly introduced into the model. To include a categorical variable into a regression model, special attention must be taken as they cannot be entered into the model as they exist. However, with the use of techniques such as the use of dummy coding, where variables can have a value of either zero or one, these categories could be incorporated (Institute for Digital Research & Education, 2021). Both the vehicle type and primary fuel categorical variables were assigned values of 1-6 and 0-3, respectively. This imposes an ordering onto the categories that are not necessarily implied by the problem context (Devore, 2016). To appropriately incorporate categorical variables, the researcher must develop $c - 1$ indicator variables, where *c = possible categories.* Finally, the emissions category was a derived value that was based on classified CO2 tailpipe emission in grams per mile. It was therefore removed as an independent variable as it has a direct dependency on the dependent variable.

Final Regression Model

$$CO_2 = 60.21 + \ 24.43\beta_1 - 1.25\beta_2 + 1.52\beta_3 + .51\beta_4 + \varepsilon \tag{4}$$

β_0 = y-intercept

β_1 = Annual Petroleum Consumption

β_2 = Combined MPG

$\beta_3 =$ Engine Displacement in Liters

$\beta_4 =$ Engine Cylinders

$\varepsilon =$ error term

Equation 4 displays the final model used to predict emissions output. To address severe multicollinearity concerns, categorical variables improperly introduced, and null values, the independent variables were reduced from eleven in the initial model to five in the final model, where only annual petroleum consumption, combined MPG, engine displacement, and engine cylinder were used. The number of observations decreased to 23,274 from the 43,177 in the original model, degrees of freedom to 3 from 5, and R-Squared to .9886 from .9829 in the original model. While this represents a significant decrease in the explained variance from the original model, it is important to note that in the original model, multicollinearity contributed to the increased R-Squared value and inaccurately explained the variance that was due to explanatory variables interdependency with one another. The final model also has a sum of squares value of 226,644,309, compared to the original mode's sum of squares value of 606,206,477. This interprets to the deviation of observed data points away from the mean value being significantly less than those within the original dataset, meaning our linear regression line is a better fit to the included data points. Our selected model's Root MSE was 10.61, compared to the originally planned model's Root MSE of 15.70, which can further support that our data fits the model well. The Root MSE is the square root of the average squared error within the regression model and is the most widely used metric to assess the overall accuracy of the model (Bruce et al., 2020). It can also be used to compare models against one another, where a smaller Root MSE translates to a model with better fit.

Discussion

The null hypothesis associated with this study suggested that the regression coefficient of each explanatory variable in the initial model was zero. Exploratory data analysis was necessary to eliminate noise in the initial model by removing variables that contributed to multicollinearity and categorical variables incorrectly introduced, such as vehicle types, emissions category, transmission type, and primary fuel type. With an F-Value of 207,194, numerator degrees of freedom of 10, denominator degrees of freedom of 23,263, and corresponding p-value of <.0001, we can reject the null hypothesis of the parameter estimates for all independent variables being zero with an $\alpha = .001$. Further interpretation of the final model's output confirms that at the 95% confidence level we can assert that the multiple regression coefficient for annual petroleum consumption is 24.43642 ± 0.0923, combined MPG is -1.25730 ± 0.0570, engine displacement is 1.51626 ± 0.2975, engine cylinders is 0.51063 ± 0.1944, and volume is 0.00078 ± 0.0032. All parameter estimates have an associated p-value of <.0001, apart from the volume, which has a p-value of 0.6461. For this reason, it was determined to leave volume out of the final regression model, as at the 95% confidence level, the parameter estimate is not statistically significant.

Strengths and Weaknesses of Model

The original emissions dataset consists of 43,177 datapoints, making it more robust to outliers and able to provide more precise population parameter estimates. Additionally, with larger datasets, the data will tend to be more normally distributed, allowing for better predictive and prescriptive statistics. While a large dataset is typically seen as a strength, it does not come without cost. Large sample sizes can magnify small differences of p-values, leading a researcher to incorrectly reject the null hypothesis when the true practicality of the difference is insignificant. Therefore, care must be taken to understand the practicality of the observed test statistics to avoid making incorrect assertions (Lin et al., 2013). When reviewing the emissions dataset and counting the number of records by vehicle year (1984 to 2021), the average number of vehicle records per year is 1,136, with a standard deviation of 227, where the minimum count for a year was 762 for 1997, and the maximum was 1,964 for 1984. With the vehicle year distributions being inconsistent year over year and 1984 being more than three standard deviations away from the mean, for example, the possibility for the data to be skewed by time periods exists. Further limitations of the study include unrounded MPG values not being available for some vehicles, interior volume dimensions not being required for two-seater passenger cars, or any vehicle classified as a truck, and tailpipe CO2 based on EPA tests for model years 2013 and beyond versus the previous year models using an EPA emissions factor.

Implications for Future Research

This study was conducted to predict CO2 emissions based on key, controllable vehicle characteristics that car manufacturers may use to determine long-run production strategies. To further enhance the validity of this study, the first possibility is to determine a valid method to incorporate the null/unknown values from the population that were removed when creating the final regression model. With these values included, the regression model would include the full population of vehicle models versus the 23,274 that were used in this study. Furthermore, to provide more accurate estimates, an individual model could be constructed by emissions category, vehicle type, and manufacturer, which would provide a more in-depth look into detailed vehicle characteristics that would allow engineering teams within their respective manufacturer to directly affect vehicle engineering norms and innovate vehicle manufacturing to align with strategic, global carbon reduction efforts.

References

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical statistics for data scientists: 50+ essential concepts using R and Python.* O'Rielly Media, Incorporated.

Devore, J. (2016). *Probability and statistics for engineering and the sciences* (8th ed.). Boston, MA: Cengage Learning

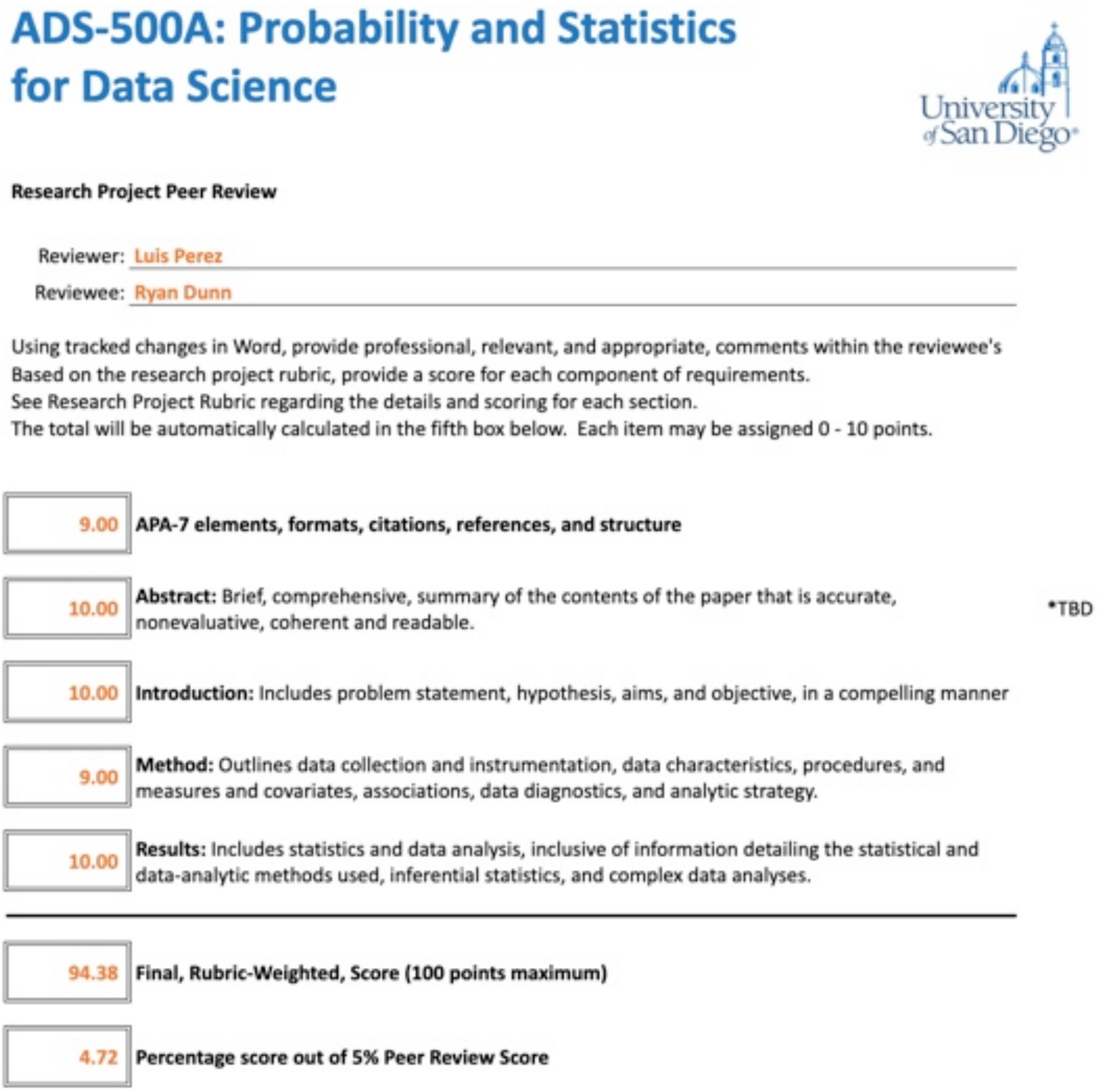
Institute for Digital Research & Education (2021). *Coding systems for categorical variables in regression analysis.* https://stats.idre.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis-2/

Lin, M., Lucas, H. C., & Shmueli, G. (2013, December). Research commentary: Too big to fail: Large samples and the p-value problem. *Information Systems Research*, *24*(4), 906–917. http://www.jstor.org/stable/24700283

United States Environmental Protection Agency. (2021, March 24). *Clean Air Act text*. https://www.epa.gov/clean-air-act-overview/clean-air-act-text

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

Statement of Peer Review



ADS-500A: Probability and Statistics for Data Science

University of San Diego

Research Project Peer Review

Reviewer: Luis Perez

Reviewee: Ryan Dunn

Using tracked changes in Word, provide professional, relevant, and appropriate, comments within the reviewee's
Based on the research project rubric, provide a score for each component of requirements.
See Research Project Rubric regarding the details and scoring for each section.
The total will be automatically calculated in the fifth box below. Each item may be assigned 0 - 10 points.

9.00	**APA-7 elements, formats, citations, references, and structure**	
10.00	**Abstract:** Brief, comprehensive, summary of the contents of the paper that is accurate, nonevaluative, coherent and readable.	*TBD
10.00	**Introduction:** Includes problem statement, hypothesis, aims, and objective, in a compelling manner	
9.00	**Method:** Outlines data collection and instrumentation, data characteristics, procedures, and measures and covariates, associations, data diagnostics, and analytic strategy.	
10.00	**Results:** Includes statistics and data analysis, inclusive of information detailing the statistical and data-analytic methods used, inferential statistics, and complex data analyses.	
94.38	**Final, Rubric-Weighted, Score (100 points maximum)**	
4.72	**Percentage score out of 5% Peer Review Score**	