Secondary Analysis of carbon dioxide emissions and petroleum consumption

Mayel Espino

University of San Diego

# Author Note

# Abstract

.

*Keywords:*

# Introduction

Global warming is ubiquitous in the news coverage, and in the political discourse. With the effects that Greenhouse gas (GHG) emissions on global warming becoming undeniable and considering that twenty eight percent of the total greenhouse emissions in the United States originate from transportation, according to the Environmental Protection Agency (n.d.) <https://www.epa.gov/transportation-air-pollution-and-climate-change/carbon-pollution-transportation>, the need for a statistical model to study this in greater detail is clear. The objective of this study is to apply the statistical tools and methods we learned in class to formulate a statistical model. The statistical model will be tested and conclusions drawn about the hypotheses postulated on the model. The hypotheses will be used to determine, for example: What are the principal characteristics of the population that have the greatest effect on the emission of greenhouse gases. Further hypotheses should be formulated and conclusions drawn from this model given its relevance to the immediate future of humanity.

Best efforts will be made to ensure that the model created for this study and the hypotheses postulated in this study are unbiased and mathematically sound, so that that they can be trusted and proven useful for future use.

This is a secondary study, the original study was done by the EPA and the sample data was acquired from said study.

The following generalized object formula or general lineal model (GLM) is a way to express multiple lineal regression models succinctly, General linear model (n.d.) Wikipedia https://en.wikipedia.org/wiki/General\_linear\_model .

co2TailpipeGpm = B0 + B1 cylinders + B2 barrels08 + B3 comb08 + B4  displ + B5 volume + B6 vehtype + B7 make + B8 mfrCode + B9 make\_id + (1)

This formula helps us describe or model the relationships between all the variables involved. The GLM also accounts for an error alongside all the factors of our equation. The error is not a mistake but rather a deviation from our model. The is the unknown that the rest of the variables B0, B1 … Bn do not account for. B0 is the baseline value for our model, when representing the location on the Y axis that our line intercepts. The B1 ..Bn are observations or basis functions of the remaining explanatory variables which are the parameters in the GLM.

The hypothesis this study postulates is that the behavior of the co2TailpipeGpm can be described in terms of variables on the righthand side of the GLM formula shown above, therefore allowing us to make good predictions based on the model, (Turner, 2008) Introduction to generalized linear models.

# Method

The data is comprised of a variety of metrics on attributes that define an automobile. We were provided a data sample obtained from FuelEconomy.gov web services. The size is considerable, with over forty three thousand records. The sample has a mix of pure electric vehicles 0.5952% , Premium and Electricity 0.2131%, Premium Gas or Electricity 0.1228%, Regular Gas and Electricity 0.1320%, Regular Gas or Electricity 0.0093%, and internal combustion vehicles 98.93%. For all vehicles the same variables are measured and recoded, there are two categories of variables: key variables and control variables. The key variables are

primary fuel tailpipe carbon dioxide emissions in grams per mile, and annual primary-fuel petroleum consumption in barrels. The control variables are: Combined miles-per-gallon for the primary fuel type, make, engine displacement in litters, engine cylinders, combined luggage and passenger volume in cubic feet, vehicle type, transmission type, and primary fuel type.

The following table provides descriptive statistics for the key variables.

**Table 1**

*Descriptive statistics, discrete data distribution and location*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *co2TailpipeGpm* | *barrels08* | *comb08* | *displ* | *volume* |
| Minimum | 0.0 | 0.06 | 7.00 | 0.000 | 0.00 |
| 1st Quartile | 386.4 | 14.33 | 17.00 | 2.200 | 0.00 |
| Median | 444.4 | 16.48 | 20.00 | 3.000 | 86.00 |
| Mean | 462.8 | 17.15 | 20.85 | 3.287 | 66.93 |
| 3rd Quartile | 522.8 | 19.39 | 23.00 | 4.300 | 113.00 |
| Maximum | 1269.6 | 47.09 | 141.00 | 8.400 | 538.00 |

**Table 2**

Descriptive statistics, categorical data counts

|  |  |  |  |
| --- | --- | --- | --- |
| make | *count* | *mfrCode* | *count* |
| Chevrolet | 0.0 | blank | 30808 |
| Ford | 386.4 | GMX | 1786 |
| Dodge | 444.4 | BMX | 1378 |
| GMC | 462.8 | FMX | 1023 |
| Toyota | 522.8 | CRX | 964 |
| BMW | 1269.6 | TYX | 941 |
|  |  | Other | 6277 |

From the counts in Table 2 we can observe that the majority of the sample population are categorized as either “Other” or “blank”. This categorial data is of limited use for our model or our hypothesis.

The boxplot graph in Figure 1, clearly shows the distribution of the key variables side by side and from the comparison we can derive that the sample population distributions are very dissimilar. Most are discrete variables, a few are categorical, such as make. What each variable measures is quite different and loosely coupled. This means we will need to process and model each of the variables independently.

Table 3 and Table 4 are the bivariate table which illustrates the sample proportions of two of our categorical variables in relationship to each other. The variables being considered are vehicle type and transmission type.

**Table 3**

Bivariate table, part one

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | *Population*  *N(%)* | *Premium Gasoline*  *N(%)* | *Midgrade Gasoline*  *N(%)* | *Regular Gasoline*  *N(%)* |
| Vehicle Type |  |  |  |  |
| Unknown | 19,730(45.7%) | 3,491(27.3%) | 90(69.2%) | 15,346(53.4%) |
| Hatchback | 5,070(11.17%) | 1,313(10.3%) | 0(0%) | 3,535(12.3%) |
| 2 door | 6,394(14.8%) | 3,157(24.7%) | 12(9.2%) | 3,120(10.9%) |
| 4 door | 11,983(27.8%) | 4,840(37.8%) | 28(21.5%) | 6,732(23.4%) |
| Transmission Type |  |  |  |  |
| Automatic | 30,210(70.0%) | 9,411(73.5%) | 130(100%) | 19,588(68.2%) |
| Manual | 12,956(30.0%) | 3,390(26.5%) | 0(0%) | 9,143(31.8%) |

**Table 4**

Bivariate table, part two

|  |  |  |
| --- | --- | --- |
| Variable | *Diesel*  *N(%)* | *Electricity*  *N(%)* |
| Vehicle Type |  |  |
| Unknown | 34(56.7%) | 84(32.7%) |
| Hatchback | 2(3.3%) | 105(40.9%) |
| 2 door | 1(1.7%) | 1(0.4%) |
| 4 door | 23(38.3%) | 67(26.1%) |
| Transmission Type |  |  |
| Automatic | 60(100%) | 248(100%) |
| Manual | 0(0%) | 0(0%) |

From the tables we should note some key relationships between the proportions of the sample population and specific categories. One key relationship is the proportional amount of unknown vehicle type to the all the fuel types. From this I conclude that comparing fuel type to vehicle type is of very limited value. On the other hand, the comparison of transmission type to fuel type is very useful.

Table 5 shows a similarly the associations between four variables, in this case the fuel type, vehicle type and transmission type against emissions category.

**Table 5**

Association of Emissions Category by Fuel Type and Other Characteristics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | *Population*  *N(%)* | *Ultra-low emissions*  *N(%)* | *Very-low emissions*  *N(%)* | *Low emissions*  *N(%)* | *Standard*  *N(%)* |
| Primary Fuel Type |  |  |  |  |  |
| Premium Gasoline | 12,801 (29.6%) | 24 (7.5%) | 70(18.2%) | 1,169(21.0%) | 9,798(33.2%) |
| Midgrade Gasoline | 130(0.3%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 124 (0.4%) |
| Regular Gasoline | 28,733 (66.5%) | 40(12.5%) | 311(81.0%) | 4,066(73.2%) | 18,971(64.2%) |
| Diesel | 1,196 (2.8%) | 0(0.0%) | 0(0.0%) | 303(5.5%) | 629(2.1%) |
| Vehicle Type |  |  |  |  |  |
| Unknown | 19,730 (45.7%) | 91(28.3%) | 50(13.0%) | 739(13.3%) | 12,579(42.6%) |
| Hatchback | 5,070 (11.7%) | 122(38.0%) | 128(33.3%) | 1,820(32.8%) | 2,952(10.0%) |
| 2 door | 6,394 (14.8%) | 7(2.2%) | 11(2.9%) | 703(12.7%) | 5,193(17.6%) |
| 4 door | |  | | --- | | 11,983 (27.8%) | | 101(31.5%) | 195(50.8%) | 2,294(41.3%) | 8,819(29.9%) |
| Transmission Type |  |  |  |  |  |
| Automatic | 30,210 (70.0%) | 312(100.0) | 301(78.4%) | 3,202(57.6%) | 20,730(70.2%) |
| Manual | 12,956 (30.0%) | 0(0.0%) | 83(21.6%) | 2,354(42.4%) | 8,813(29.8) |

**Table 6**

Association of Emissions Category by Fuel Type and Other Characteristics

|  |  |  |
| --- | --- | --- |
| Variable | *Polluter*  *N(%)* | *Gross Polluter*  *N(%)* |
| Primary Fuel Type |  |  |
| Premium Gasoline | 1,262(21.4%) | 478(32.4%) |
| Midgrade Gasoline | 6(0.1%) | 0(0.0%) |
| Regular Gasoline | 4,358(73.9%) | 987(67.0%) |
| Diesel | 259(4.4%) | 5 (0.3%) |
| Natural Gas | 14(0.2) | 4 (0.3%) |
| Electricity | 0(0.0%) | 0(0.0%) |
| Vehicle Type |  |  |
| Unknown | 5,119(86.8%) | 1,152(78.2%) |
| Hatchback | 47(0.8%) | 1(0.1%) |
| 2 door | 339 (5.7%) | 141 (9.6%) |
| 4 door | 394 (6.7%) | 180 (12.2%) |
| Transmission Type |  |  |
| Automatic | 4,557(77.3%) | 1,108(75.2%) |
| Manual | 1,341(22.7%) | 365(24.8%) |

There is more useful information in these table, we can see intuitive patterns that are intuitive and logical in the relationships. For example: The fact that the lowest emissions category has the largest proportion in the electrical vehicle variable, or that the gross polluters have the highest proportion in the most popular fuel time, regular gasoline. Additionally the way that the proportions are related to the transmission type is also expected, for example no electric vehicles have manual transmission.

Because the patterns seen in this associations table are consistent to what we expect to see in the general population, the sample data seems to be representative of the hypothetical population. Another useful conclusion is that the relationship between fuel type and emissions category is useful for our model.

# Results

Now we consider the probability density function for the CO2 emissions:

PDFco2 = f(x;µ = 465.538, (2)

The probability density function (PDF) describes mathematically all possible values of our variable of interest, in this case CO2 emissions, these values plotted on a graph give us the distribution of all the probabilities. Considering the PDF we select the most appropriate statistical formulas and processes to apply.

From the CO2 emissions probability plot on Figure 2 we see that the sample data is very nearly normally distributed. We can see that the values at either end, that is the tails, are the farthest from the normal distribution. This is consistent from the boxplot for CO2 in Figure 1. The values in the tails, the outliers, correspond to the measurements from pure electric and hybrid automobiles and internal combustion automobiles that are extreme polluters.

Table 7 and Table 8 contain the correlation coefficients of CO2 emissions, fuel consumption, fuel type, automobile make, engine displacement, emissions category and transmission type. From the tables we arrive at a few key observations, there is a strong correlation between The CO2 emissions and : The fuel consumption, the engine displacement, the number of cylinders, and the emissions category. There is also strong but inverse correlation between the CO2 emissions and the combined fuel type.

On the other hand there is a marked lack of correlation between CO2 emissions and: automobile make, automobile type and primary fuel type. This is unsurprising and a good indication that these variables may not be very useful for our model or hypothesis.

There are other strong correlations, positive and negative between annual fuel consumption and: combined fuel consumption, engine displacement, cylinders and emissions category. These are unsurprising and not very useful for our model or hypothesis.

**Table 7**

Pearson Correlation Coefficients (N=42,917) – Part 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | co2TailpipeGpm | barrels08 | comb08 | make\_id | displ |
| co2TailpipeGpm | 1.0000 | .9885 | (.9184) | (.2157) | .7954 |
| barrels08 | .9885 | 1.0000 | (.9050) | (.2117) | .7843 |
| comb08 | (.9184) | (.9050) | 1.0000 | .2072 | (.7327) |
| make\_id | (.2157) | (.2117) | .2072 | 1.0000 | (.2823) |
| displ | .7954 | .7843 | (.7327) | (.2823) | 1.0000 |
| cylinders | .7438 | .7337 | (.6863) | (.2670) | .9046 |
| volume | (.4323) | (.4266) | .4161 | .1165 | (.3628) |
| vehtype | (.3626) | (.3580) | .3313 | .0940 | (.2631) |
| emissionscat | .8894 | .8791 | (.8415) | (.1755) | .6703 |
| prifueltype | (.1128) | (.1084) | .1234 | .0710 | (.2149) |

**Table 8**

Pearson Correlation Coefficients (N=42,917) – Part 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | cylinders | volume | vehtype | emissionscat | ??? |
| co2TailpipeGpm | .7438 | (.4323) | (.3626) | .8894 | (.1128) |
| barrels08 | .7337 | (.4266) | (.3580) | .8791 | (.1084) |
| comb08 | (.6863) | .4161 | .3313 | (.8415) | .1234 |
| make\_id | (.2670) | .1165 | .0940 | (.1755) | .0710 |
| displ | .9046 | (.3628) | (.2631) | .6703 | (.2149) |
| cylinders | 1.0000 | (.2648) | (.1524) | .6185 | (.2181) |
| volume | (.2648) | 1.0000 | .7418 | (.3627) | .0498 |
| vehtype | (.1524) | .7418 | 1.0000 | (.3054) | (.0340) |
| emissionscat | .6185 | (.3627) | (.3054) | 1.0000 | (.0874) |
| prifueltype | (.2181) | .0498 | (.0340) | (.0874) | 1.0000 |

**Table 8**

Contingency and chi-square table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Hatchback | 2 door | 4 door | Unknown | Total |
| Gross polluter | | | | | |
| Observed | 1.00 | 141.00 | 180.00 | 1152.00 | 1474.00 |
| Expected | 173.00 | 218.28 | 409.08 | 673.55 | 1473.99 |
| Chi-square | 171.00 | 27.36 | 128.28 | 339.86 | 666.59 |
| Low emission | | | | | |
| Observed | 1820.00 | 703.00 | 2294.00 | 739.00 | 5556.00 |
| Expected | 652.41 | 822.78 | 1542.00 | 2538.80 | 5555.99 |
| Chi-square | 2089.59 | 17.44 | 366.74 | 1275.91 | 3749.67 |
| Polluter | | | | | |
| Observed | 47.00 | 339.00 | 394.00 | 5119.00 | 5899.00 |
| Expected | 692.68 | 873.57 | 1637.20 | 2695.60 | 5899.05 |
| Chi-square | 601.87 | 327.12 | 944.02 | 2178.69 | 4051.70 |
| Standard | | | | | |
| Observed | 2952.00 | 5193.00 | 8819.00 | 12579.00 | 29543.00 |
| Expected | 3469.00 | 4375.00 | 8199.10 | 13500.00 | 29543.10 |
| Chi-square | 77.05 | 152.94 | 46.97 | 62.84 | 339.69 |
| Ultra-low emissions | | | | | |
| Observed | 122.00 | 7.00 | 101.00 | 91.00 | 321.00 |
| Expected | 37.69 | 47.54 | 89.09 | 146.68 | 321.00 |
| Chi-square | 188.57 | 34.57 | 1.59 | 21.14 | 245.89 |
| Very-low emission | | | | | |
| Observed | 128.00 | 11.00 | 195.00 | 50.00 | 384.00 |
| Expected | 45.09 | 56.87 | 106.57 | 175.47 | 384.00 |
| Chi-square | 152.45 | 36.99 | 73.38 | 89.72 | 352.53 |
| Total |  |  |  |  |  |
| Observed |  |  |  |  | 43177.00 |
| Expected |  |  |  |  | 43177.12 |
| Chi-square |  |  |  |  | 9406.05 |

Degrees of freedom = 15, = 0.5, we get a p-value < 2.2 e-16 (nearly zero).

What we are testing here is if the polluter category is independent from the vehicle classification: two door, four door, hatchback and unknown. So we postulate the null and alternate hypotheses as follows:

H0: emissions category is independent from automobile type

H1: emissions category is dependent from automobile type

# Discussion

The strengths of this study are in the statistical tools and methods that were applied, and in particular in the model, the hypothesis formulation, the tests and determination of the hypothesis. The statistical methodology we followed insures that we are unbiased in our conclusions, free from any preconceived notions we may or may not be aware of. To that end we follow reliable techniques and statistical tests based on mathematical principles and a history of experimental and/or empirical knowledge.

model and hypotesis

\*\* The weakness in this study are in the data. Not expect to be perfect

Sampling methods

Silent

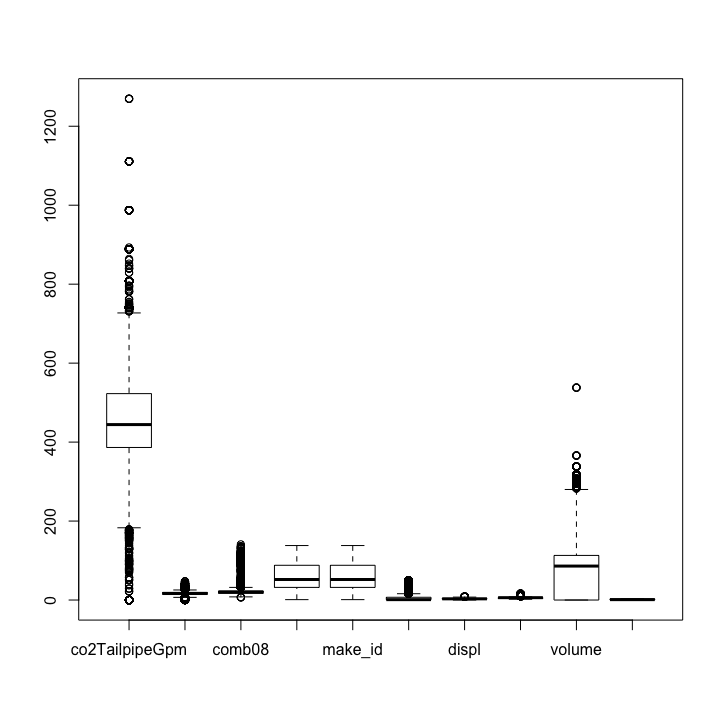
Cross vectoring

\*\*Implications for future research or application

# Appendix

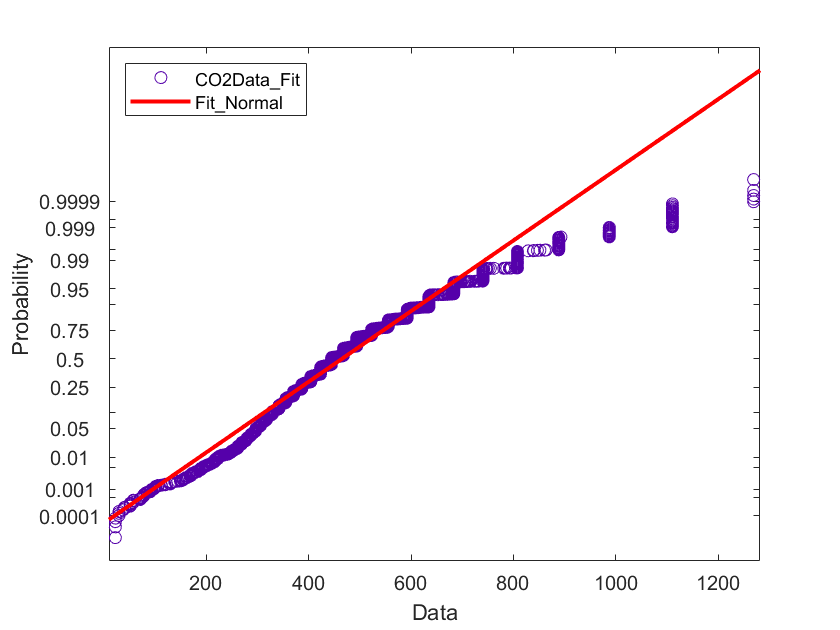
**Figure 1**

Boxplot of key variables



**Figure 2**

Normal probability plot



References

<https://www.epa.gov/transportation-air-pollution-and-climate-change/carbon-pollution-transportation>

<https://www.epa.gov/greenvehicles/fast-facts-transportation-greenhouse-gas-emissions>

Turner, Heather (2008) Introduction to generalized linear models <http://statmath.wu.ac.at/courses/heather_turner/glmCourse_001.pdf>