IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Regressive Analytical Approach to Automobile Carbon Emission

Sihyun Lee

Asia Pacific International School (APIS), 57 Wolgye-ro, Nowon-gu, Seoul, 01874, South Korea; sihyunlee1010@gmail.com

ABSTRACT: This study predicted automobile carbon emission via linear regression by evaluating the prediction error contribution based on the statistical analysis. Carbon emissions are the most critical concern for recent global climate change, and automobiles are a significant contributor. We elucidated the relationship between automobile emission predictions and their classifications since a data-driven regression method that enables a practical forecast of high-emission vehicles concerning the historical emission categories. The applied historical data comprises the following regulatory features: engine size, number of cylinders, highway fuel consumption, and combined fuel consumption. We set up the experiment using the linear regression model to predict CO_2 emission (g/km). Our coefficient of determination came out to 0.939, which is a good fit. In addition, we found that the residual distribution of predicted errors is generally distributed with a 2% prediction error. Among the features, fuel consumption shows the most substantial relation to CO_2 emissions.

KEYWORDS: Environmental Engineering, Pollution Control, Regressive Analysis, Machine Learning Algorithm.

Introduction

Climate change is the long-term change in Earth's local and regional weather conditions based on historical trends.[1] It results in alterations in the environment, including a rise in temperature in land and sea, a rise in sea levels, loss of ice at Earth's poles, changes in frequency and severity of extreme weathers — such as hurricanes, heatwaves, wildfires, droughts, floods, and precipitation — and changes in cloud and vegetation cover, which negatively impact humans' activities like residence or agriculture.[2] Paradoxically, human activities mainly drive such change; for example, burning fossil fuels like coal or oil emits gases, which contributes to the greenhouse effect.[2] Its significant effect is the global temperature increase induced by trapping the heat in the Earth's atmosphere. One of the critical parameters for the greenhouse effect is carbon dioxide.

Automobiles are the major contributor to carbon emissions, occupying 28% of US greenhouse gas emissions.[3] They contribute by combusting fossil fuels like gasoline or diesel, releasing carbon dioxide as a byproduct. For the EU emissions, road transport has been calculated to cause about a fifth of overall emissions.[4] Transport methods outside of automobiles varied on how much they affected carbon emissions. While passenger cars accounted for 61% of carbon emissions from road transport, alternative transportation methods such as electric cars, public transport, bicycles, or walking had significantly less emissions.[3] Another reason automobiles contribute so much to greenhouse gas emissions is the increased travel demand, which has resulted due to growth in population, economy, and cities.[3]

The emerging method to measure greenhouse gases in the atmosphere is spectroscopy,[5] which utilizes electromagnetic radiation. By measuring the wavelength of molecules in the atmosphere through the absorption of electromagnetic radiation, it is possible to identify greenhouse gases. Another method exists, which uses satellites. The basic principle is not different, but it simply makes work on a bigger scale. The satellites provide measurements of greenhouse gases like CO_2 in all the places the sunlight casts on the Earth's surface, using spectrometers to detect light passing through the atmosphere.[6] Multiple methods have been utilized to examine CO_2 emissions from passenger automobiles. Chassis dynamometer (CD), remote sensing (RS), plume chasing (PC), and portable emissions measurement systems (PEMS) are mainly carried out for measuring emitted CO_2 in grams per kilometer (g/km).[7] In North America, pounds of carbon dioxide per megawatt-hour (lbs. CO_2/MWhr) is used as the vehicle emission performance standard.[8]

Understanding CO_2 emissions related to automobile classified performance is strongly recommended to assess the rapidly growing transition of the greenhouse effect. Thus, finding a possible solution to slow down the growth of the greenhouse effect associated with automobile usage is essential. In this study, we attempted to evaluate the automobile carbon emission rate using linear regression with machine learning algorithms. Since linear regression is a straightforward analysis with the statistical approach, which can reflect the effect of independent variables on the model. We hypothesized that the linear regression model would enable us to estimate the extent to which individual automobile classifying parameters can contribute to overall CO_2 emissions. The analysis used historical data in Canada that classifies emission rates related to automobile performance. We also conducted a statistical analysis to predict the most severe case of carbon emission (in grams per kilometer).

■ Methods

Dataset:

The dataset applied in this study consists of fuel consumption rates and estimated carbon dioxide (CO_2) emissions classified for the light-duty vehicles registered in Canada in 2023. The raw dataset was downloaded via Natural Resources Canada[9] or Kaggle.[10] Table 1 shows a part of the original dataset. Table 2 presents the data statistics: total counts, mean, standard deviation, minimum, and maximum values in the object at 25, 50, and 70 percentiles.

In the dataset, the vehicle class is divided along with the interior volume of the engine. The engine size exhibits the total volume of all the cylinders in the engine in liters. The cylinders are the power unit of an engine where fuel is burned. Transmission is defined as automatic (A), automated manual (AM), automatic with select shift (AS), continuously variable (AV), and manual (M) with the number of gears and speed. Fuel types are regular gasoline (X), premium gasoline (Z), diesel (D), and electricity (B). Fuel consumption is measured in liters per kilometer and describes the car's fuel per kilometer. This fuel consumption is based on 5-cycle test works, which monitor a vehicle's performance using a chassis dynamometer machine.[10] Highway fuel consumption (Hwy) presents how much fuel the car uses on the highway. The Comb (L/100 km) and Comb (mpg) provide an average consumption at a city in liters per 100km and mile per gallon (282.42 / (L/100km)), respectively. CO_2 emissions are defined as the amount of carbon dioxide that the vehicle produces in grams per kilometer. CO_2 and smog ratings are estimated by examining the vehicle's tailpipe emission of CO_2 and smog-forming pollutants rated on a scale from 1 (worst) to 10 (best).

Table 1: A part of the dataset applied in this study.[9,10] It consists of 13 parameters related to the vehicle's performance. Unit is provided. Provides raw data obtained from Natural Resources Canada.[9]

Year	Make	Model	Vehicle Class	Engine Size (L)	Cylinders	Transmission	Fuel Type	Fuel Consumption (L/100Km)	Hwy (L/100 km)	Comb (L/100 km)	Comb (mpg)	CO_2 Emissions (g/km)	CO_2 Rating	Smog Rating
2023	Acura	Integra	Full-size	1.5	4	AV7	Z	7.9	6.3	7.2	39	167	6	7
2023	Acura	Integra A-SPEC	Full-size	1.5	4	AV7	Z	8.1	6.5	7.4	38	172	6	7
2023	Acura	Integra A-SPEC	Full-size	1.5	4	M6	Z	8.9	6.5	7.8	36	181	6	6
2023	Acura	MDX SH-AWD	SUV: Small	3.5	6	AS10	Z	12.6	9.4	11.2	25	263	4	5
2023	Acura	MDX SH-AWD Type S	SUV: Standard	3	6	AS10	Z	13.8	11.2	12.4	23	291	4	5

Table 2: Statistics of the dataset applied in this study. Counts are the total number; mean is the average; std is the standard deviation; min is the minimum value; max is the maximum value; and 25%, 50%, and 75% are the percentile. Rates range from 0 (worst) to 10 (best). Data is processed.

	Engine Size (L)	Cylinders	Fuel Consumption (L/100Km)	Hwy (L/100 km)	Comb (L/100 km)	Comb (mpg)	CO_2 Emissions (g/km)	CO_2 Rating	Smog Rating
count	833	833	833	833	833	833	833	833	833
mean	3.15	5.63	12.43	9.35	11.05	27.38	257.47	4.52	5.24
std	1.35	1.97	3.46	2.30	2.88	7.56	64.26	1.28	1.67
min	1.20	3.00	4.40	4.40	4.40	11.00	104.00	1.00	1.00
25%	2.00	4.00	10.10	7.70	9.00	22.00	211.00	4.00	5.00
50%	3.00	6.00	12.10	9.10	10.70	26.00	254.00	5.00	5.00
75%	3.60	6.00	14.60	10.70	12.90	31.00	299.00	5.00	7.00
max	8.00	16.00	30.30	20.90	26.10	64.00	608.00	9.00]	8.00

Analysis:

We used the Google Colab notebook[11] to find the linear regression, which allowed us to write and execute Python code conveniently. First, we removed missing or misspelled values in the data for cleaning. After that, we analyzed all variables to see if anything abnormal was in the data. Since we aim to predict CO_2 emission, we speculated that linear regression could be straightforward, considering the relationship between the dependent variables, CO_2 emission, and the independent variables (other parameters: attributes). In addition, we plotted the residuals to see if they are typically distributed to clarify if there is a low correlation between independent variables. We chose Comb (L/100 km) rather than Comb (mpg) because both are the same quantity. CO_2 Rating and Smog Rating are highly correlated or directly related to CO_2 Emissions. So, we will drop those in the model training to avoid redundancy. The final variables we selected are Engine Size (L), Cylinders, Fuel Consumption (L/100 km), Hwy (L/100 km), Comb (L/100 km), and CO_2 Emissions (g/km). Figure 1 shows the heatmap of the relationship between the variables.

The Python package called sklearn provides a convenient way to train and evaluate the linear regression model.[12] It also provides a function called train_test_split,[13] which splits the input data into training and testing samples with a 75-25 ratio. We use training samples to train the model and optimize the parameters to obtain the best possible model based on the training dataset. Then, we evaluate our model using the testing samples to minimize the training bias. We utilized R^2, coefficient of determination,[14] to assess the model's fit to the training data. The typical range of R^2 (R-squared: a goodness-of-fit measure) values is between 0 and 1. The perfect fit gives R^2 equals 1, and R^2 for poor fit goes to 0. Here are the fit parameters:

Table 3: The optimized parameters after the model training. The second last row shows the intercept and the last row R^2 (R-squared: a goodness-of-fit measure) of training data. The R^2 value, a goodness-of-fit measure, is shown to be close to 1, indicating a good fit. The coefficient is varied by the input parameter.

Parameter (Column) name	coefficient
Engine Size (L)	-0.855
Cylinders	3.566
Fuel Consumption (L/100 km)	13.061
Hwy (L/100 km)	11.399
Comb (L/100 km)	-4.204
Analysis fit parameters	
Intercept (bias)	17.931
R^2 (goodness of fit)	0.939

Using the numeric columns, the data was represented as the heatmap in Figure 1 to evaluate the correlation for all cases. A color code describes each correlation with the range of -1 (negative relation) to 1 (positive relation); 0 is for zero related. We found a close dependency between engine size, cylinders, Fuel consumption (L/100km), Hwy (L/100km), and Comb (L/100km). Afterward, we divided the dataset into training and testing samples, which were 75% to 25%. The testing samples, made prior to this step, are used to evaluate the performance of the trained model. The predicted values from the testing samples are obtained from the trained model and then compared with the ground truth values from the testing samples. To make the linear relationship of the model, the training samples were used to train, and the testing samples were applied to confirm the accuracy. We used the train_test_split function, provided by sklearn,[10] to prepare the two datasets. The linear regression function, also provided by sklearn,[12, 13] was used to fit the linear model and calculate the R2 (coefficient of determination) of the determined model.

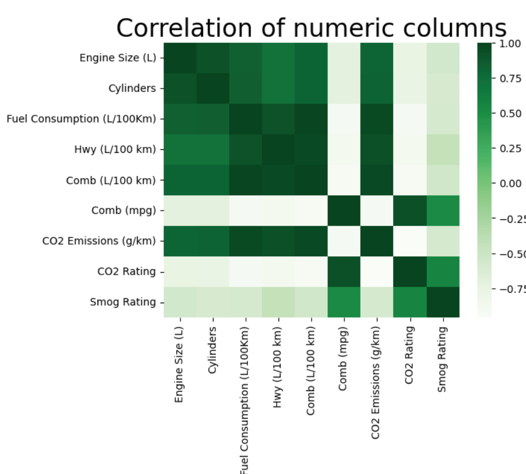

Figure 1: Heatmap for the correlation of numeric columns. The relation is presented by a scale from -1 (no relation, white) to 1 (100% dependent, darkest green). The scale bar is presented at the right. The green highlighted scale bar shows the close relationship.

We looked over each distribution histogram for each column (attributes): vehicle maker, vehicle class, engine size, number of cylinders, transmission, fuel type, fuel consumption (L/100km), highway fuel consumption (L/100km), combined fuel consumption (L/100km), combined fuel consumption (mile/gallon), carbon emission in grams per kilometer, CO_2 ratings, and smog ratings. Figure 2 represents the distribution histogram of fuel consumption for Fuel, Highway (L/100km), and Combined city and highway (L/100km). It was to understand the overall data structure before analysis. All other histograms are presented in our Collab note.[15]

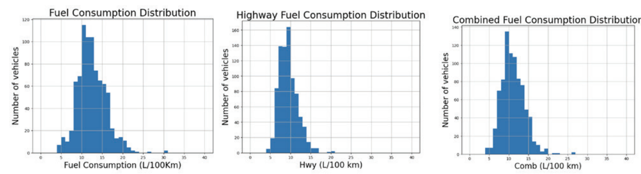

Figure 2: Distribution histogram of fuel consumption, including fuel, highway, and combined city and highway. The histograms are used to understand the data before conducting the analysis.

Results and Discussion

Figure 3 shows the comparison between the predicted values and the ground truth values. The predicted CO_2 emissions are based on factors that show a strong correlation, as shown in Figure 1, including engine size, number of cylinders, fuel consumption, and highway, city, and combined fuel consumption. The red line is plotted to indicate the perfect match. While most data points are very close to the red line, only a few are off. To quantify the extent of the deviation by the residuals, we plot the residual distribution by fitting it with a Gaussian function, as shown in Figure 4. The Gaussian function has advantages in evaluating the interval between the predicted values.[16] After that, we estimated the mu and sigma (the normal distribution's mean shift and standard deviation, respectively). The fitting shows that our prediction is slightly shifted to the left with an error of 5.657 [g/km]. Since our mean CO_2 emission is about 257.5 g/km, described in Table 2, the simple percentage ratio indicates the predicted error is about 2%. Since R2 is a degree of fitting the expected values in terms of the statistical model, the obtained R2 ~ 0.94 could be considered to be a good fit.

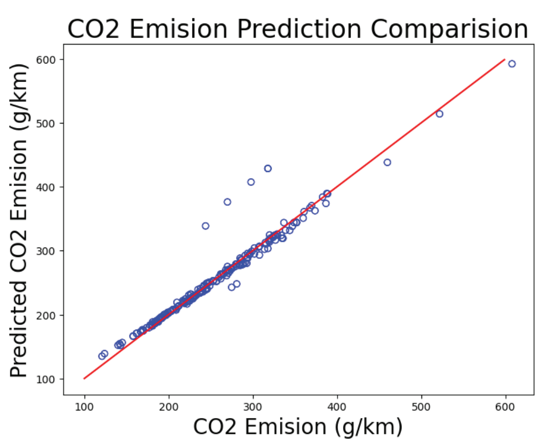

Figure 3: CO_2 emission prediction comparison between predicted values and ground truth values. The linear red line indicates the perfect match between the two values.

DOI: 10.36838/v7i4.98

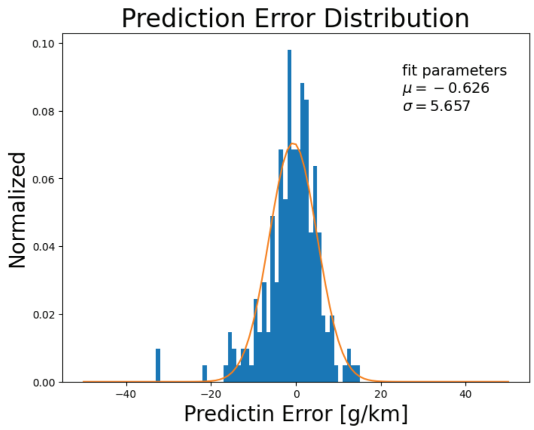

Figure 4: The residual distribution of prediction errors. The x-axis shows the difference between predicted emission and ground truth in the [g/km] unit. The orange line is the fitted Gaussian function where the mu (the mean of the normal distribution) is shifted slightly to the left, and the sigma (the standard deviation of the normal distribution) is about 5.657 g/km.

CO_2 gas emissions from passenger vehicles contribute almost 30% of the greenhouse effect. Without proper control, the average Earth temperature will rapidly increase, and the greenhouse effect will accelerate. To reduce CO_2 emission, vehicle performance parameters should be evaluated in terms of the emission rate. Thus, in this study, we tried to predict the amounts of carbon dioxide (CO_2) emitted by various registered automobiles based on their classified regulated performance. Since linear regression has a simple statistical way of exhibiting the relations between the independent variables, we applied this method to accurately estimate the relation of the automobiles' CO_2 emission associated with fuel consumption. The vehicles in our dataset were officially registered in Canada Natural Resources,[10] and their performance related to CO_2 emission was well-defined. Our statistical analysis confirmed that the distribution of each automobile highly depends on the classification of each year's fuel consumption guidance. Evaluation of the correlation of numerical values from the fuel consumption classification showed that engine size, cylinders, fuel consumption, highway, city, and highway combined fuel consumption are highly related. Linear regression coefficients (Table 3) indicate that the fuel consumption (liter per 100km) is the most decisive parameter, and highway fuel consumption (liter per 100 km) follows the next. City and highway consumption and engine size are likely significant parameters. Typically, one should determine how well the applied model can fit the data after fitting with a linear regression. This measure is R2 (R-squared: a goodness-of-fit), representing the residuals deviating from the perfect fit. Our obtained R2 is 0.939, suggesting a generally good fit, as the R2 value is acceptable when the R2 value is between 0.50 and 0.99. However, the R2 value may not be considered a criterion to determine whether the model is adequate. This is because the acquired value describes variant relatives, considering each model and the analysis environment. It is practically difficult to say that a value above 0.9 is always highly desirable if all other well-defined fittings are present.[17] Similarly, the predicted error can also be a source to indicate the accuracy of our prediction. Figure 4 shows that mu (the mean of the normal distribution) is -0.63, and the sigma (the standard deviation of the normal distribution) is about 5.7 g/km. This suggests that the error distribution is very small, and its value is significantly smaller compared to 257.5, the mean value of the predicted carbon emitted, as shown in Table 2. It is calculated to be only 2% when converted to a percentage.

Linear regression has different model types, and its characteristics depend highly on the model process and the established variables.[18] The least square method is used to examine the best fit, as shown in our analysis. The review of the linear regression methodology suggests that a proper selection of the model with the choice of model variables is more important to acquiring high accuracy and valid statistical results. Although we applied a simple linear regression model in this study, we could compare the CO_2 emission rates of automobile performances using various regression models in future work.

Conclusion

We used linear regression to characterize the predictive automobile carbon emission after evaluating the prediction error contribution with statistical analysis. This data-driven regression analysis enabled a practical forecast of high-emission vehicles concerning the historical emission categories comprising the regulatory features: engine size, number of cylinders, highway fuel consumption, and combined fuel consumption. The coefficient of determination appeared to be 0.939, which is considered to be a good fit. We also showed that the residual distribution of predicted errors is generally distributed with a 2% prediction error. Among the features, fuel consumption shows the most substantial relation to CO2 emission. Since carbon emission is the most critical concern for the recent global climate change, our analytical approach would help understand the significance of passenger vehicles on global warming.

Acknowledgments

We especially thank Dr. Z. Sung (STEM Program Mentor) for advising us on the technical approach and helping us establish the central concept.

References

1. "What is Climate Change?" United Nations, https://www.un.org/en/climatechange
2. "Climate Change," NASA, https://science.nasa.gov/climate-change/
3. "Greenhouse Gas (GHS) emissions and removals," EPA United States Environmental Protection Agency, https://www.epa.gov/ghgemissions
4. "Road transport," European Environment Agency, https://www.eea.europa.eu/en/topics/in-depth/
5. J. B. Burkholder; J. P. D. Abbatt; I. Barnes; J. M. Roberts; M. L. Melamed; M. Ammann; A. K. Bertram; C. D. Cappa; A. G. Carlton; L. J. Carpenter; J. N. Crowley; Y. Dubowski; C. George; D/E. Heard; H. Herrmann; F. N. Keutsch; J. H. Kroll; V. F. McNeill; N. L. Ng; S. A. Nizkorodov; J. J. Orlando; C. J. Percival; B. Picquet-Varrault; Y. Rudich; R. W. Seakins; J. D. Surratt; H. Tanimoto; J. A. Thornton; Z. Tong; G. S. Tyndall; A. Wahner; C. J. Weschler;

K. R. Wilson; P. J. Ziemann, "The Essential Role for Laboratory Studies in Atmospheric Chemistry," *Environ. Sci. Technol.* **2017**, 51, pp 2519-2528

6. P. Guanna; X. Yuan; M. Jieqi, "The potential of CO2 satellite monitoring for climate governance: A review", Journal of Environmental Management, Vol 277, **2021**, 111423,
7. D. Tena-Gago; Q. Wang; J. M. Alcaraz-Calero, "Non-invasive, plug-and-play pollution detector for vehicle on-board instantaneous CO_2 emission monitoring," Internet of Things, Vol 22, **2023**, 100755,
8. "Greenhouse Gas Emissions from a Typical Passenger Vehicle," EPA United States Environmental Protection Agency, https://www.epa.gov/ghgemissions
9. EnerGuide for vehicles, Natural Resources Canada, Government of Canada, "https://natural-resources.canada.ca/energy-efficiency/energuide/energuide-vehicles/21010"
10. Fuel Consumption Ratings 2023, https://www.kaggle.com/datasets/imtkaggleteam/fuel-concumption-ratings-2023
11. Google Colaboratory, "https://colab.google/"
12. Scikit-learn, Machine learning in python, "https://scikit-learn.org/stable/index.html"
13. "Train_test_split," Scikit-learn, Machine learning in python, "https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html"
14. J. Gao, "R-squared(R2) - How much variation is explained?", Research Methods in Medicine & Health Sciences, vol 5, issue 4, pp 104-109,
15. Regresion_analysis, https://colab.research.google.com/drive/1-Qy7WjjDHiAOK0oTV0gstlMmpcAT2luC?usp=sharing
16. H Guo, "A Simple Algorithm for Fitting a Gaussian Function [DSP Tips and Tricks]," Sep 2011, IEEE Signal Processing Magazine, Vol 28, Issue 5, pp134-137.
17. What's a good value for R-squared, Linear regression model, "https://people.duke.edu/~rnau/rsquared.htm"
18. D Maulud, and A M. Abdulazee, "A review on linear regression comprehensive in machine learning," Dec 2020, JASTT, vol. 1, no. 2, pp. 140-147.

Author

Sihyun Lee is an Asia Pacific International School senior interested in machine learning, environmental science, and robotics. He hopes to build a career as a pioneer in robotics and economics using state-of-the-art IT technology, such as machine learning and artificial intelligence.

DOI: 10.36838/v7i4.98