# 

**CSE523 : Machine Learning**

Winter 2021 - 2022

**Weekly Report - 4**

Dt : 23-02-2022

**Group Name : Discover Decipher**

**Group Members**

Nimisha Patel - AU1940146 - Btech Computer Science and Engineering.

Sakshi Shah - AU1940213 - Btech Computer Science and Engineering.

Astha Patel - AU1940312 - Btech Computer Science and Engineering.

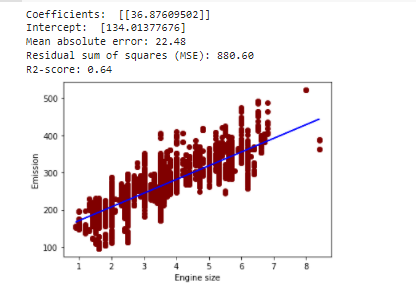
Kareena Matwani - AU1940314 - Btech Computer Science and Engineering.

**Task Performed this week**

**Modeling and Analyzing the Single Variable Linear Regression**

Regression is a "Supervised Learning" technique that predicts a continuous value as a function of a set of independent variables. It has a number of advantages, including quick training and no need for parameter adjustment. In its most basic form, linear regression is the process of determining the "best fit" line for a set of data. In this scenario, we plot the output variable vs. independent variable scatter plot to get the best fit line.

In the below graph, a regression line is plotted for the correlation on the train dataset.



**Goal :** To predict CO2 emissions given engine size using linear regression.

**Formula :** Y = A\*X + B

Where Y (to predict) = CO2 emissions and X (train parameter) = Engine size

**Output:** Values of A (regression coefficient - slope) and B (regression intercept) are calculated using LinearRegression() function of Skilearn library.

Over here, the value of A= 36.87609502 which describes that the model predicts the response 36.87 when the train parameter X is 0. The output (that is CO2 emission) increases by 134.01377676 (intercept - B) when X (engine size) is increased by 1.

The feature variables like Engine Size (L), Fuel Consumption City, Fuel Consumption Hwy, Fuel Consumption Comb and number of cylinders are strongly correlated to CO2 emission. We performed single linear regression for all the strong correlated feature variables.

| **Feature variable and Output variable** | **Coefficient (A)** | **Intercept (B)** | **Error detection** | **Interpretation** |
| --- | --- | --- | --- | --- |
| Cylinders Vs CO2 Emission | 26.6321904 | 100.84300209 | Mean absolute error: 24.52  Residual sum of squares (MSE): 1056.50  R2-score: 0.54 | Linear regression line fits up to 54%  Poor fit |
| Fuel Consumption City (L/100 Km) Vs. CO2 Emission | 15.33686324 | 57.89522376 | Mean absolute error: 14.46  Residual sum of squares (MSE): 523.70  R2-score: 0.81 | Linear regression line fits up to 81%  good fit |
| Fuel Consumption Comb (L/100 Km) Vs. CO2 Emission | 18.53986085 | 47.01600518 | Mean absolute error: 13.89  Residual sum of squares (MSE): 516.18  R2-score: 0.81 | Linear regression line fits up to 81%  good fit |
| Fuel Consumption Hwy (L/100 Km) Vs. CO2 Emission | 23.20954427 | 40.69657404 | Mean absolute error: 17.68  Residual sum of squares (MSE): 688.72  R2-score: 0.74 | Linear regression line fits up to 74%  good fit but is not as other fuel consumption feature variables |
| Engine Size (L) Vs. CO2 Emission | 36.69788783 | 134.41496153 | Mean absolute error: 22.88  Residual sum of squares (MSE): 906.38  R2-score: 0.62 | Linear regression line fits up to 62%  Not so good fit |

**Advantages of Linear Regression**

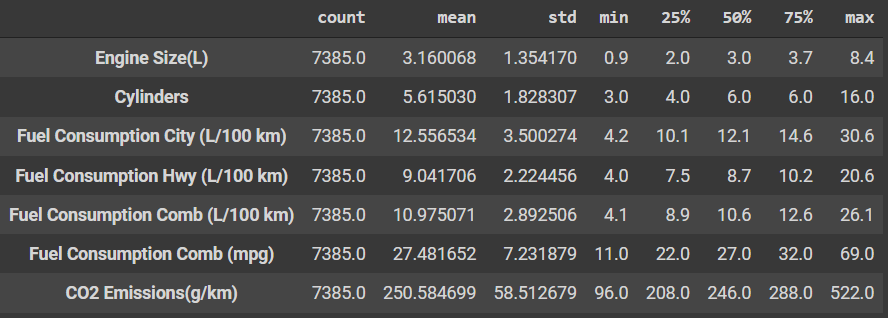
* Simple implementation.
* Performance on linearly separable datasets
* Overfitting can be reduced by regularization.

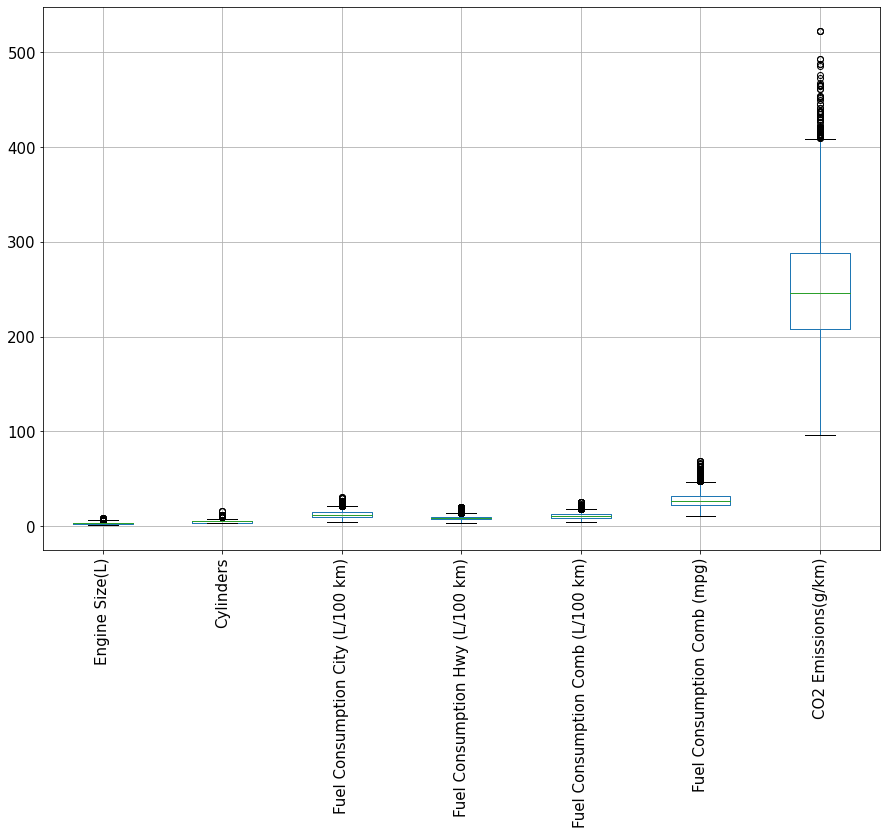
**Disadvantages of Linear Regression**

* Prone to underfitting.
* Sensitive to outliers.
* Linear Regression assumes that the data is independent.

**Check for outliers in the dataset as the linear regression is sensitive to outliers**

Higher standard deviation for CO2 Emission column suggests presence of outliers





Further we tried to remove outliers and perform the linear regression to note the r2 scores.

**References**

(PDF) the carbon emissions of Chinese cities. (n.d.). Retrieved March 1, 2022, from [https://www.researchgate.net/publication/258564432\_The\_carbon\_emissions\_of\_Chinese\_cities.](https://www.researchgate.net/publication/258564432_The_carbon_emissions_of_Chinese_cities)

Real Python. (2021, March 19). *Linear regression in python*. Real Python. Retrieved March 2, 2022, from <https://realpython.com/linear-regression-in-python/>

**Tasks to be Performed Next Week**

Analyze multivariable linear regression, understand the model precision for the data set.

Try different efficient algorithms to predict CO2 emission.