Analysis of CO2 Emission of New Passenger Cars in Europe Union from 2011 to 2016

Authors:

Qiuyao Liu, Jiawen Liang, Hanxing Li, Minqi Lu, Guanjia Wang

12/11/2018

Abstract

Since 2009, the European Union has required for each member state to provide detailed information on each new passenger car that registered in its territory, including specific emission of CO2, engine power and size and so on. We are interested in analyzing the data and quantifying the relationship between the EU new passenger cars' factors and the CO2 emission through the statistics model. To achieve our goal, we performed the Principle Component Analysis for dimensionality reduction, used LASSO for feature selection and created the Linear Regression Model upon that. Additionally, we included the market share of each manufacturer group in order to have a better understanding of affection on the CO2 emission. We can determine that the CO2 emission of the new passenger car is dependent on the physical factors including mass, engine power, and fuel type. Using the data from 2011 to 2016, we found that the CO2 emission is in the decreasing trend; however, there are some adverse effects involved which we will discuss in detail in our conclusion part.

1 Prior Work

1.1 Background

Greenhouse effect and climate change are the serious concerns we have to face and try to control immediately. The carbon dioxide (CO2) is the most significant part that made up greenhouse gases, which is a massive contributor to climate change. Moreover, vehicles' CO2 emission is considered as the main portion of CO2 emission in general. According to the International Organization of Motor Vehicle Manufacturers, 15.9% of the total CO2 emission comes from road transportation (cars, trucks, buses, etc.). The European Commission set the emission target for each manufacturer maker based on their amount of vehicles sold. In 2008, The European Commission presented the definitive package called 2020 climate & energy package, which focuses on alleviating the emission, achieving the energy renewables, and improving energy efficiency. Moreover, the EC released their 2020 target for new passenger cars' CO2 emission, which is 95 g CO2/km.

1.2 Data Source

The data source comes from the European Environment Agency and is under regulation 443/2009 monitoring by the European Commission. The following details of each new passenger car are all required to submit to the Commission: specific emission of CO2, mass of vehicle, wheel base, engine capacity, fuel type and manufacturer name and so on. There are 23 columns in the dataset and we will extract the necessary columns for our research.

2 Data Preprocessing

After looking through the dataset, we found out there are several columns that are irrelevant to our project, such as “Commercial Name” and “Manufacturer Harmonised”. Based on that, we took only ten columns that are related to our project out of 23 columns.

Before we started the data cleaning, we combined the data from 2011 to 2016 into one csv file and added the year column for each yearly dataset then combined them into one final dataset for further analysis. There are two types of variables: numeric variables, and categorical variables. For numeric variables, we decided to fill the missing values with the mean. For categorical variables, we noticed that there are some differences in terms of case sensitivity; therefore, we unified all the names into uppercase and replaced space with underscore. In order to increase the accuracy of our model, we removed duplicates and extracted the numerical information for standardization.

We also included the market share of each group into the data. The final preprocessing step was to estimate the ‘market share’ for each manufacturers group. This is because manufacturers have total emissions targets, which may incentivize them to keep emissions low for popular models. The market share for each manufacturers group was defined as its contribution to the total number of registered vehicles over the entire data set, expressed as a percentage. After preprocessing the data, we normalized it. There are some outliers detected that occurred under unusual circumstances might be caused by typos during the data entry process. We kept the outliers as we are uncertain about its reason behind.

3 Context and exploratory analysis

We calculated the correlations among our seven numeric variables: CO2 emission, the Wheelbase, Engine power, Mass, Engine capacity, Axle width steering axle, Axle width other axles. Among which, Mass, Engine power, and Engine capacity show the significant correlations with CO2 emissions. Besides, other numeric variables also show the noticeable correlations with CO2 emissions. However, since all the numeric variables show the moderated high correlation with each other, there is a big chance that the collinearity exists. It is therefore worth trying to find a smaller number of uncorrelated covariates that retain most of the information on vehicle characteristics. To do this, the principal components (PCs) were calculated: the first, second and third components accounted for 50%, 80%, and 90% respectively of the total variance.

Having calculated the PCs, plots such as those in Figure 2 were produced to visualize the relationships between potential covariates and emissions rates for different vehicle groups. Overall correlations are shown for each plot, along with group-specific nonparametric regression curves. Figure 2 shows that physical variables mass and engine size seem to increase with emissions and that, of the potential numeric covariates, PC1 and Mass have the clearest relationships with CO2: the regression curves for these variables seem roughly a straight line for the most of fuel types. The highest correlations are with EngineSize(ec.cm3.), but that for EngineSize is probably inflated by the electric vehicles with zero engine size; moreover, the regression curves for these variables are not entirely plausible. Figure 2 also shows outliers in some variables.

4 Methodology & Model

We consider using the linear regression model, which is based on the increased variability between higher-emissions vehicles in Figure 2, together with the linear relationships with EnginePower and Mass.

The first model incorporated only the most essential covariates (Mass, EnginePower, and

FuelType). The LASSO test suggested that mass and EnginePower are the most predominant numeric variables. This first model had an adjusted R-squared of 0.6652, which is not a bad result. Even though the standard diagnostics revealed no problems, the residual distribution had a larger tail than expected.

Sequently, the remaining variables were considered, including manufacturer groups and EU member state. As all member states within EU have similar regulations between each state, it will not be an essential covariate. Also examined via residual boxplots, as in Figure 3, the distributions seem fairly similar in all member states, but for manufacturer groups, variation is substantial and hence should be accounted for in the model. Moreover, we added the Market Share into the model along with manufacturer groups. After adding two more variables into to model, adjusted R-squared increased to 0.6948. The result of F test shows the p-value is significant, so we believed the new model is better.

The exploratory analysis suggested for the choices on pcs, and we replaced Mass and EnginePower by PC1, PC2, and PC3 to get a new model. The improvement on the Adjusted R-squared of the new model to 0.7042 shows a slightly better fit than the previous model and a lower residual standard error. Standard diagnostics indicate that the model does an excellent job of capturing the systematic structure in both the mean and variance of the response, although, the residual distribution still has a substantial heavy tail. We also tried the log to see if there is the improvement on the heavy tail, which did not work. Hence, we kept this model.

5 Conclusion/ Finding

CO2 emissions for new passenger cars are represented by a linear regression model, containing 34 coefficients and representing the potential CO2 emissions as a linear combination of four continuous variables (PCs 1, 2, 3 and market share) and two categorical factors (fuel group and manufacturer group).

The modeled effect of market share is straightforward, as we have discovered that the manufacturer groups with larger market share tend to have a lower impact on the CO2 emission volume, although the effect of this is very weak — this suggests that the definition of market share could be improved. In fact, the manufacturer group with the largest market share, the VW group, has a negative coefficient. We believe that this is due to the fact that the larger companies have more revenues and thus can invest more money into hybrid fuel types and other methods for CO2 emission reduction. The effects of other variables are more complicated. On average, emissions increase with vehicle size as expected, but at different rates for different manufacturers and fuel types. After studying the coefficients of the categorical variables, we found that the coefficients for diesel-electric, petrol-electric, and electric are negative, which means that these three fuel types can reduce the volume of CO2 emissions.

6 Further analysis

To further analyze in which factors directly influence the CO2 emission, we could add more terms. Firstly, emission-reduction technology will cause a significant reduction in the CO2 emission and whether this technology will be affected by different vehicle models is worth discussing. Moreover, we used six-year data but it was not reflected in our model. In fact, every year the average CO2 emissions decrease and it is worth to discuss the influence of the year to the emissions. In addition, we can add the interaction of the covariates to improve our model, for example, physical factors and FuelType, which may have a better interpretation of statistical models.

7 Bibliography

1 “Climate Change & CO2.” *OICA*. www.oica.net/category/climate-change-and-co2/

2. Source code for all of the analytical work: https://github.com/Williamburgson/ads_co2_eu

Appendix 1. Contribution

In this project, I am responsible for the model analysis and result interpreting, which include model selection, model improvement, and result visualization, etc. I also co-wrote the final paper with other team members.

Appendix 2. Figures

Figure 1:
Projected 2D plot of the six numerical variables, labeled by manufacturer group

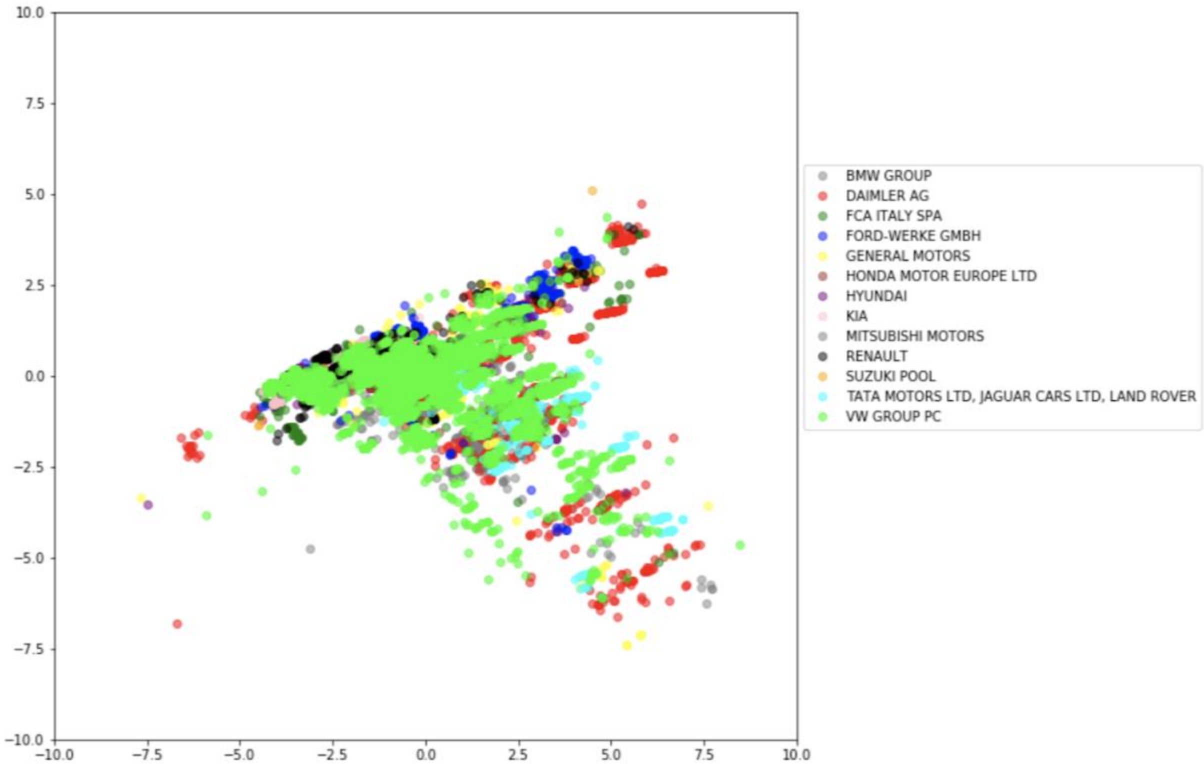


Figure 2:

Relationships between CO2 emissions rates and potential continuous covariates, including principal components. Covariate names are given beneath each plot, and the overall correlation is given above. Colors correspond to vehicles with different fuel types, with nonparametric regression curves obtained using the lowess() command in R with default settings.

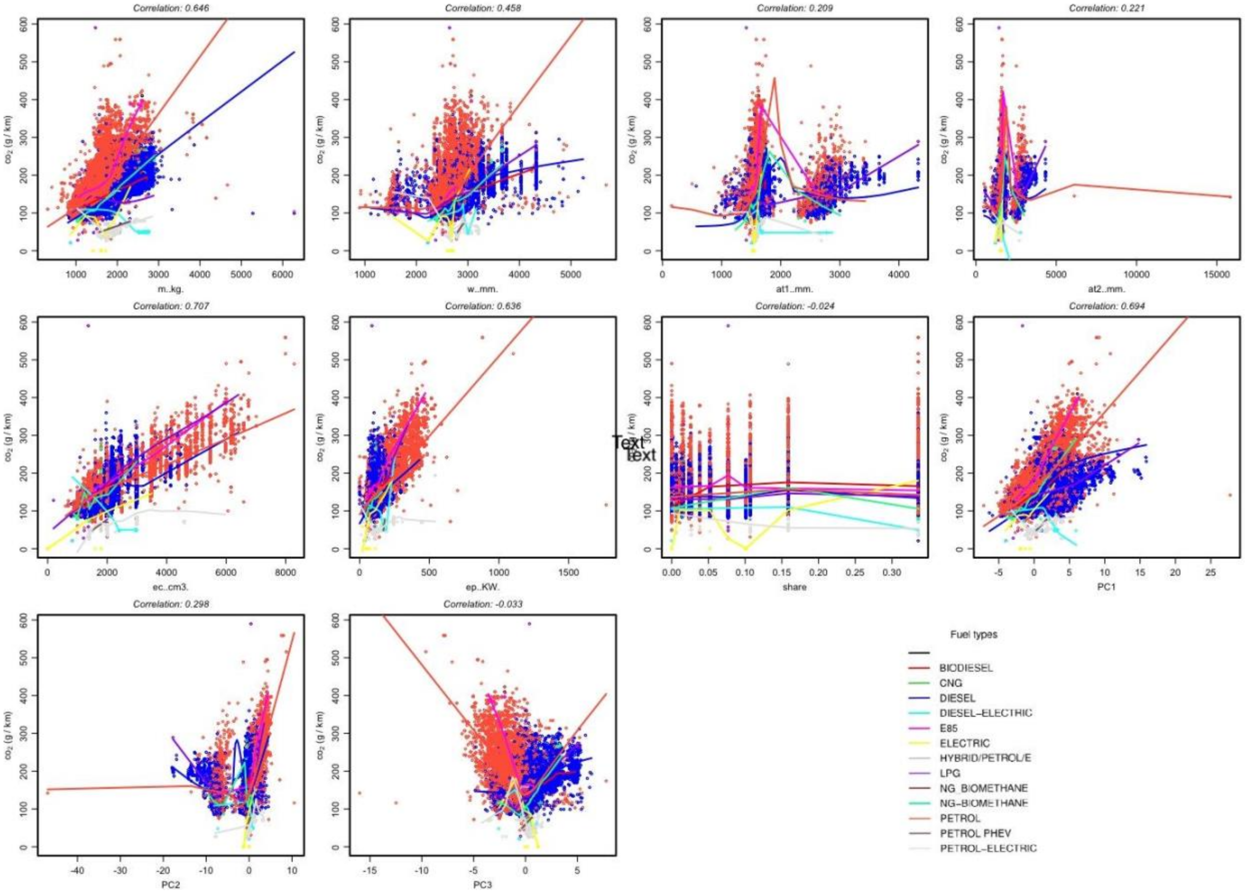


Figure 3:
Distributions of residuals from model 1 by EU member state and by manufacturer group

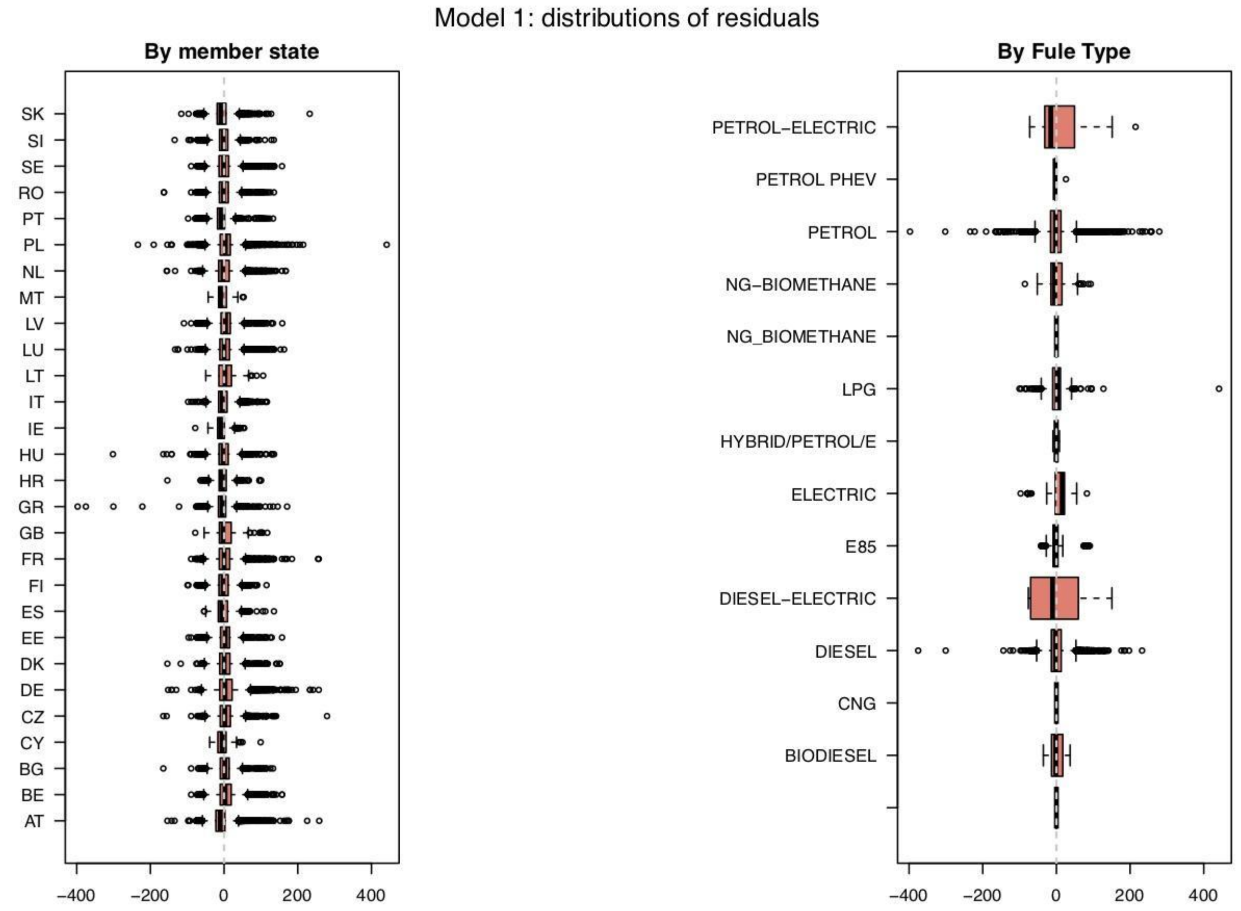
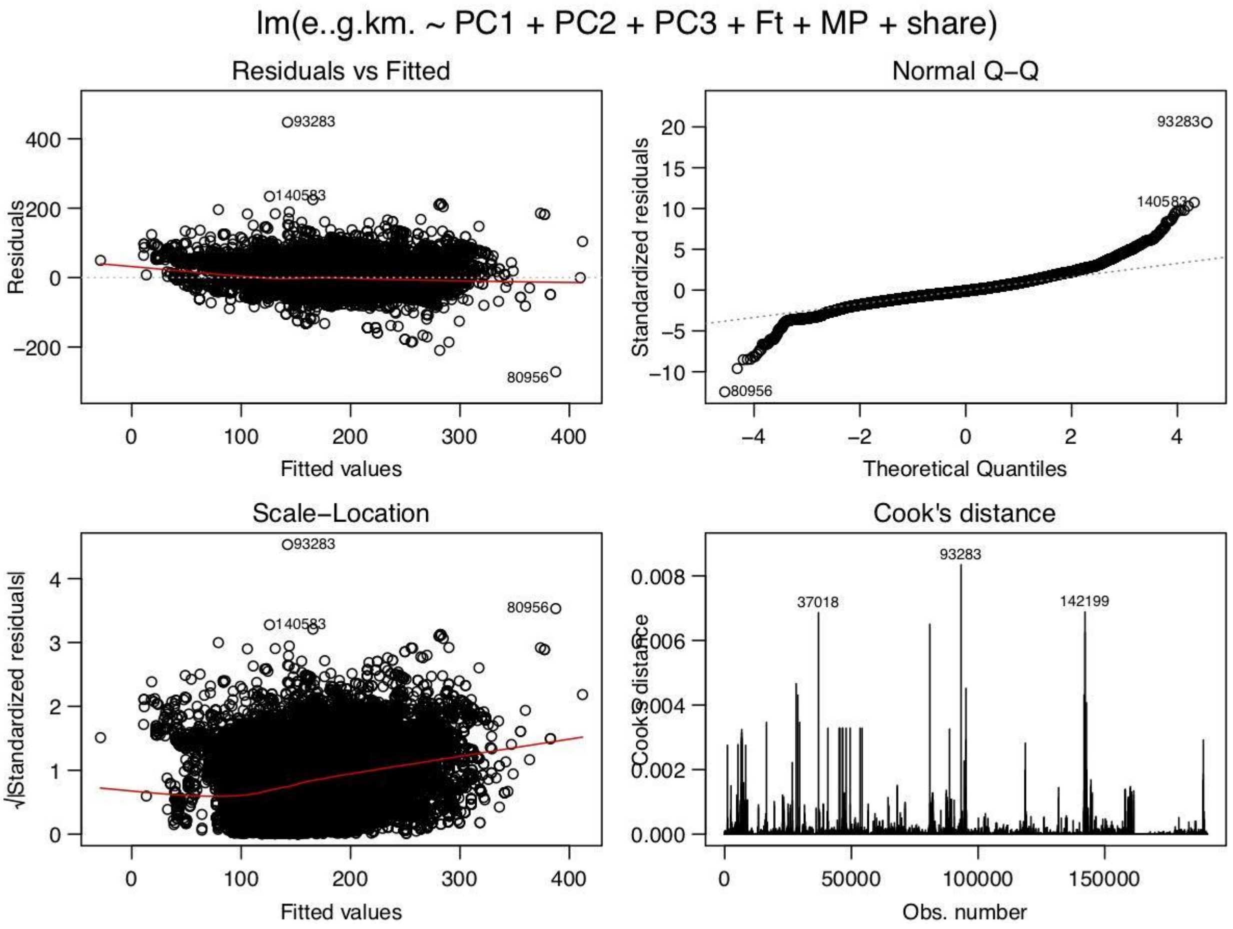


Figure 4:
Final model diagnosis plots


lm(e..g.km. ~ PC1 + PC2 + PC3 + Ft + MP + share)
Residuals vs Fitted
Residuals
400
200
0
−200
93283
140583
80956
0
100
200
300
400
Fitted values
Normal Q−Q
Standardized residuals
20
15
10
5
0
−5
−10
93283
140583
80956
−4
−2
0
2
4
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
4
3
2
1
0
93283
80956
140583
0
100
200
300
400
Fitted values
Cook's distance
Cook's distance
0.008
0.006
0.004
0.002
0.000
93283
37018
142199
0
50000
100000
150000
Obs. number