# About Urbanization & Gas Price

A MODS206 project

by Illan KNAFOU
Nghia Danh NGUYEN
Théo NGUYEN NGOC
and Victor PEREZ

Supervised by Laurie CIARAMELLA

# Table of contents

# Introduction of the subject

Oil has been fuelling the economic growth of the XXth century, but is now largely looked down upon because of the consequence its use has over global warming. Yet, oil is still pumped in vast quantities, and it is still referred to as the black gold. Its actual price, once refined, varies a lot from one country to another, meaning that people all over the globe don't have the same ability to buy petrol, and therefore to drive their car.

Indeed, the cheaper the petrol, the cheaper the cost of transportation by car. In Venezuela on the one hand, a liter of gasoline costs 0.020 U.S. Dollar, while on the other hand, in Hong Kong, this same liter costs 2.442 U.S. Dollars. These were the two extremal prices on March, 29th 2021, yet they shed light on the considerable disparity that exists between countries at a worldwide scale.

Another thing that tends to vary drastically from a country to another is the part of the population living in urban areas. The human population worldwide has been steeply growing since the dawn of humanity, and since 2007, more than half of the world population is living in cities. It is predicted that by 2050 about 64% of the developing world and 86% of the developed world will be urbanized.

Therefore, urbanization is at the core of the social development of the upcoming years. It is relatively obvious that people living in more sparsely populated areas might tend to drive around the lot more than the ones living in urban areas, where one can find anything he might need to live in a 10-kilometer radius. This is the reason why in our project we wanted to investigate the possible correlation existing between the oil price in a country, and the urbanization in the same country.

In order to conduct our project, we first had to find suitable datasets, which wasn't as easy as expected, yet we ended up finding enough data to lead our study. The countries we focused on are the United States of America, Canada, the United Kingdom and France. Although all these countries are relatively close when it comes to economic development, they were the ones with the most data available, and we believe that a panel of data, ranging from 1992 to 2019, is appropriate for us to try to shed a light on an hypothetical correlation, at least in the aforementioned countries, between the price of petrol and the level of urbanization.

# 1. State of the art

In order to begin the realization of our project, it was first necessary to research all existing information concerning our subject, to make a synthesis that could be used in the following. This was done through bibliographic work and an analysis of formal and informal publications concerning our study.

We were therefore able to observe that many publications agree on the fact that the price of petrol has an influence on the location of populations in general. What is interesting is that in our research we did not find any articles, publications or studies that actually show that there is a causality between urbanization and petrol prices.
Nevertheless, by focusing on the various impacts that the evolution of petrol prices can have, we were able to merge results that finally allow us to reach a causal link. For example, we learned that petrol prices affect household location [1], the development of low-density housing [2], and the value of suburban houses [3]. We have further developed the subject to deepen our study and to have a critical view on the subject, for example the lack of a significant effect of petrol prices on house prices under certain circumstances [4].

After a thorough literature search, we critically analysed the references to get an overview. We realized that most of the publications and studies conducted are indirectly based on official government sources. In order to start our study, we therefore collected data from these original sources to verify and corroborate the arguments mentioned in the articles selected for the state-of-the-art synthesis.

In a second step, we researched to find relevant control variables for our study. The main finding here is that the rate of urbanisation is undoubtedly related to the economy of the country. Many studies show that economic growth is a cause of urbanisation [5], [6]. Focusing this time on transport, we noted that in the literature the number of motorised vehicles and fuel consumption are factors influencing the number of people living in rural areas [7].

This in-depth research allowed us to select variables that are corroborated by the literature and that we will discuss in detail in the next section.

# 2. Data description

As a first step, we felt it was essential to start documenting our data at the beginning of our research project, even before we started collecting the data. This makes it easier to document the data and reduces the likelihood of forgetting aspects of our data later in the research project.

Once the data for the variables of interest: petrol prices and urbanization rates were found, we were able to collect them easily. Urbanization rates were collected from BP Statistical Review of World Energy in the form of an XLS file that needed to be cleaned up.

As for the price of petrol, it was more complicated. These data were only available from official sources in each country. So we had to search for our data on U.S. Energy Information Administration, Insee, Statistics Canada, etc.

| A | C | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ountry Nam | Indicator Name | 1960 | 1961 | 1962 | 1963 | 1964 | 1965 | 1966 | 1967 | 1968 | 1969 | 197 |
| Afghanistan | Urban population (% c | 8,401 | 8,684 | 8,976 | 9,276 | 9,586 | 9,904 | 10,233 | 10,57 | 10,918 | 11,275 | 11, |
| Albania | Urban population (% c | 30,705 | 30,943 | 31,015 | 31,086 | 31,158 | 31,23 | 31,302 | 31,374 | 31,446 | 31,548 | 3 |
| Algeria | Urban population (% c | 30,51 | 31,797 | 33,214 | 34,662 | 36,141 | 37,643 | 38,84 | 39,004 | 39,169 | 39,334 | |
| American Sa | Urban population (% c | 66,211 | 66,641 | 67,068 | 67,493 | 67,916 | 68,334 | 68,75 | 69,163 | 69,574 | 69,98 | 70, |
| Andorra | Urban population (% c | 58,45 | 60,983 | 63,462 | 65,872 | 68,205 | 70,445 | 72,593 | 74,641 | 76,588 | 78,424 | 80, |
| Angola | Urban population (% c | 10,435 | 10,798 | 11,204 | 11,624 | 12,058 | 12,504 | 12,965 | 13,441 | 13,932 | 14,436 | 14, |
| Antigua and | Urban population (% c | 39,656 | 39,04 | 38,427 | 37,817 | 37,211 | 36,61 | 36,012 | 35,418 | 34,829 | 34,245 | 33, |
| Arab World | Urban population (% c | 31,2341425 | 31,9799192 | 32,7291285 | 33,5015971 | 34,3035974 | 35,1476995 | 35,9497898 | 36,6410603 | 37,3439600 | 38,0562095 | 38,780 |
| Argentina | Urban population (% c | 73,611 | 74,217 | 74,767 | 75,309 | 75,844 | 76,369 | 76,888 | 77,398 | 77,901 | 78,394 | 7 |
| Armenia | Urban population (% c | 51,275 | 52,147 | 53,019 | 53,889 | 54,758 | 55,622 | 56,483 | 57,341 | 58,195 | 59,042 | 59, |
| Aruba | Urban population (% c | 50,776 | 50,761 | 50,746 | 50,73 | 50,715 | 50,7 | 50,685 | 50,67 | 50,654 | 50,639 | 50, |
| Australia | Urban population (% c | 81,529 | 81,941 | 82,228 | 82,511 | 82,792 | 83,068 | 83,34 | 83,507 | 83,672 | 83,836 | 83, |
| Austria | Urban population (% c | 64,72 | 64,814 | 64,863 | 64,913 | 64,962 | 65,011 | 65,061 | 65,11 | 65,159 | 65,208 | 65, |
| Azerbaijan | Urban population (% c | 52,663 | 52,364 | 52,064 | 51,764 | 51,464 | 51,164 | 50,864 | 50,564 | 50,263 | 49,963 | 50, |
| Bahamas, T | Urban population (% c | 59,712 | 60,454 | 61,193 | 61,926 | 62,642 | 63,343 | 64,04 | 64,73 | 65,416 | 66,093 | 66, |
| Bahrain | Urban population (% c | 82,32 | 82,337 | 82,355 | 82,372 | 82,389 | 82,5 | 82,761 | 83,019 | 83,275 | 83,526 | 83, |
| Bangladesh | Urban population (% c | 5,135 | 5,278 | 5,498 | 5,727 | 5,964 | 6,211 | 6,467 | 6,733 | 7,009 | 7,296 | 7, |
| Barbados | Urban population (% c | 36,777 | 36,849 | 36,922 | 36,995 | 37,068 | 37,141 | 37,214 | 37,287 | 37,36 | 37,433 | 3 |
| Belarus | Urban population (% c | 32,401 | 33,522 | 34,663 | 35,822 | 37 | 38,127 | 39,269 | 40,422 | 41,588 | 42,76 | 4 |
| Belgium | Urban population (% c | 92,46 | 92,554 | 92,679 | 92,835 | 92,988 | 93,137 | 93,284 | 93,428 | 93,569 | 93,707 | 93, |
| Belize | Urban population (% c | 54,028 | 53,72 | 53,41 | 53,101 | 52,79 | 52,481 | 52,17 | 51,86 | 51,549 | 51,238 | 50, |
| Benin | Urban population (% c | 9,275 | 9,856 | 10,47 | 11,118 | 11,801 | 12,519 | 13,275 | 14,069 | 14,903 | 15,776 | 16, |
| Bermuda | Urban population (% c | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |
| Bhutan | Urban population (% c | 3,596 | 3,792 | 3,999 | 4,217 | 4,446 | 4,687 | 4,94 | 5,207 | 5,487 | 5,781 | 6, |

*Figure 2.1. Extract of the urbanization rate per country and per year since 1960*

We then proceeded in the same way to collect the data for the control variables. Using the World Bank national accounts data, we were able to obtain CSV files containing GDP per capita and percentage GDP of industry and service sectors. We also collected data on petrol consumption from the BP Statistical Review of World Energy.

Finally, the most difficult data to collect was the number of motorized vehicles per country. The range of vehicles and their uses was so varied that the figures collected did not seem to us to be really justified and reliable.

In the end, we decided to opt for the number of cars per country. The difficulty here was that for each country the data had to be collected from different sources to obtain the most reliable results. Once the data was collected for Canada, the USA, France and the UK, we were able to create a single file by merging our data to facilitate the processing of the data frame.

As the data was country-specific, the dates for which we had the number of cars were also different. We therefore filled in some empty fields so that the regression with and without the added values had the same coefficients.

In the end, we had seven cleaned DataFrames to deal with in our study:

- The two variables of interest: the percentage of population in urban areas and the price of petrol
- The five control variables: number of cars per country, daily consumption of oil in barrel, GDP per capita, the percentage of the industry and service sector in the total GDP.

# 3. Descriptive Analysis Data

The data we collected is summed up in the following table, visualizing average, standard deviation and extremal values for our different variables.

*Table 4.1. Descriptive statistics about our data*

| | **Mean** | **Maximum** | **Minimum** | **Standard deviation** |
|---|---|---|---|---|
| Urbanization (%) | 81,17 | 88,59 | 74,40 | 4,00 |
| Gas price ($/Litre) | 0,76 | 1,78 | 0,18 | 0,36 |
| GDP per capita ($) | 37 033,25 | 65 297,52 | 18 389,02 | 11 156,54 |
| Nb of cars (in millions) | 10 364,49 | 44 469,00 | 23,43 | 18 753,04 |
| Oil consumption (barrel/day) | 6 114,67 | 20 532,00 | 1 505,00 | 7 376,58 |
| Industry (% of GDP) | 22,20 | 29,88 | 17,07 | 3,71 |
| Services (% of GDP) | 67,30 | 77,51 | 28,01 | 9,45 |

All the values are computed over the whole period and among the four countries. More detailed statistics can be found in the all_variables file.

As we can see in Table 4.1, all the countries we sourced our data from were largely urbanized at the end of the XXth century, yet the GDP per capita does vary quite a lot, mostly over time, since the countries got richer over the time span that we study. When it comes to the price of petrol, although the movement is very erratic, yet upon closer inspection, it rises steadily over time, with a mean of about one dollar in 2019 compared to a mean of 40 cents for a litre in 1992. It is also important to note that there is a much more significant part of the GDP coming from services than industry in 2019 than in 1992. Concerning oil consumption, it is very interesting to note that the United Kingdom and France have reduced their consumption over the last thirty years, while the United States and Canada have augmented it. It is quite revealing of the concern about climate change in these countries, Europeans seeming to be ahead of the rest of the world.

# 4. Empirical Strategy

## 4.1 Data collection

### 4.1.1 Collect data

We had to collect the following data: the percentage of population in urban areas (= Y), the price of gasoline (variable of interest) *gas_price*, then the GDP per capita *gdp*, the number of cars *num_car*, the daily consumption of oil in barrels *oil_consumption*, the percentage of the industry and service sector in the total GDP *industry; services* (control variables). These data are collected according to year (1992 to 2019) and country (USA, UK, France and Canada).

### 4.1.2 Process data

Some data are available for all countries and all years in a single CSV file, for others it is up to us to gather all the data found on different sources to be able to compare them.

In the second case, problems of units can arise. For example, for the *gas_price* we have to harmonize all the data so that they are expressed in $/litre. Or, some data may be missing, which is why we chose to focus our study on 1992 - 2019, as all the data was available during this period.

Finally, after processing the different data sets separately to put them in the same format, we put them together in a single table to make them easier to use.

| | dates | country | urbanization | gas_price | gdp | num_car | oil_consumption | industry | services | intercept |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1992 | US | 82.789 | 0.287155 | 25418.990776 | 194.42735 | 16969.0 | 23.132100 | 71.809000 | 1.0 |
| **1** | 1993 | US | 83.654 | 0.281872 | 26387.293734 | 198.04134 | 17161.0 | 23.132100 | 71.809000 | 1.0 |
| **2** | 1994 | US | 84.485 | 0.283985 | 27694.853416 | 201.80192 | 17635.0 | 23.132100 | 71.809000 | 1.0 |
| **3** | 1995 | US | 85.280 | 0.293495 | 28690.875701 | 205.42721 | 17635.0 | 23.132100 | 71.809000 | 1.0 |
| **4** | 1996 | US | 86.043 | 0.316742 | 29967.712718 | 210.44125 | 18245.0 | 23.132100 | 71.809000 | 1.0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **107** | 2015 | United Kingdom | 82.626 | 0.708271 | 44974.831877 | 31.20000 | 1552.0 | 18.141621 | 70.408410 | 1.0 |
| **108** | 2016 | United Kingdom | 82.886 | 0.476838 | 41064.133432 | 31.80000 | 1597.0 | 17.581142 | 70.922420 | 1.0 |
| **109** | 2017 | United Kingdom | 83.143 | 0.669715 | 40361.417383 | 32.20000 | 1610.0 | 17.562620 | 70.934180 | 1.0 |
| **110** | 2018 | United Kingdom | 83.398 | 0.741141 | 43043.227816 | 32.50000 | 1584.0 | 17.518831 | 71.043378 | 1.0 |
| **111** | 2019 | United Kingdom | 83.652 | 0.723782 | 42330.117537 | 32.90000 | 1545.0 | 17.420687 | 71.276739 | 1.0 |

112 rows × 10 columns

*Figure 4.1 Final dataset after processing all the datasets*

## 4.2 Simple regressions

First, we did a very simple regression with the explained variable *urbanization* and the variable of interest *gas_price*. This model can be biased because we can omit many other factors that might have an impact on ‘urbanization’ such as the

GDP, the oil consumption... Therefore, we added into our model several control variables that might help us include the omitted factors. The order of adding control variables is illustrated in Table 4.1.

*Table 4.1. Simple regression models*

| Variables | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 |
|---|---|---|---|---|---|---|
| *gas_price* | *x* | *x* | *x* | *x* | *x* | *x* |
| *gdp* | | *x* | *x* | *x* | *x* | *x* |
| *num_car* | | | *x* | *x* | *x* | *x* |
| *oil_consumption* | | | | *x* | *x* | *x* |
| *industry* | | | | | *x* | *x* |
| *services* | | | | | | *x* |

## 4.3 Include entity and time fixed effects

Because we are dealing with the panel data, we also tried to include the entity and time fixed effects into our model. In our project, time is measured by years and the entity is the country.

### 4.3.1 Include entity fixed effects

In our data, we just have four countries (for entities), thus, we used the method of 'n-1 binary regressors' in order to include entity fixed effects into our model. To implement this, we introduced three binary variables: *is_Canada*, *is_France*, and *is_UK*. *is_Canada* is equal to 1 if this observation's country is Canada and *is_Canada* is zero otherwise, similarly for *is_France* and *is_UK*. We have not added the binary variable *is_USA* to avoid the dummy variable trap.
In this part, we added two models **Model 7** and **Model 8**. We add three above dummy variables into **Model 6** to create **Model 7**, while **Model 8** just included the variable of interest (*gas_price*) and three above binary variables.

### 4.3.2 Include time fixed effects

To include time fixed effects in our model, we have two different methods: 't-1 binary regressors' or year-demeaned. In our data, we have 28 years (from 1992 to 2019), therefore, 't-1 binary regressors' can be very complicated. Thus, we did the year-demeaned method first, and this will be presented in our **Model 9**. Then, because the year-demeaned method is very sensitive to human mistakes, we also executed 't-1 binary regressors' in **Model 14**. In **Model 14**, we introduce three new additional binary variables named *1999_2005, 2006_2012,* and *2013_2019*. *1999_2005* is equal to 1 if the year of the observation is between 1999 and 2005, and it will be zero otherwise. We used the similar ideas for *2006_2012,* and *2013_2019* variables.

Then, after adding entity and time fixed effects into our model, we made some joint hypothesis tests to see whether these additional variables are statistically different from 0 or not.

## 4.4 Non-linear regressions

By plotting the data we have, we can see whether there is any non-linear effect of *gas_price* on *urbanization* or not. We firstly tried to fit the model with the polynomials of *gas_price*. The **Model 112** represents the quadratic model (including power 2 of *gas_price* - *gas_price^2*), and the **Model 11** represents the cubic model (including both power 3 and 2 of *gas_price* - *gas_price^3, gas_price^2*).

After this, we also considered the interaction effects between a binary and a continuous variable, and two continuous variables. A new binary variable we introduce here is *is_USA* - whether this observation is of the USA or not. The idea is to find out whether the impacts of *gas_price* on *urbanization* in the USA are different from these remaining countries. Therefore, we added three new cross variables *is_USA X gas_price, is_USA X gas_price^2,* and *is_USA X gas_price^3* (shown in **Model 13**). The similar idea was applied for *oil_consumption*, we introduced three additional cross variables: *oil_consumption X gas_price, oil_consumption X gas_price^2,* and *oil_consumption X gas_price^3*. These variables are presented on the **Model 19**.

By adding these new variables, we also tried to take some hypothesis tests to see the significance of these variables. And, based on these tests, we decided which variables should be kept in our model and which should not.

In the next step, we included some logarithm forms into our model. First, we wonder if the percentage change in GDP can lead to different impacts of *gas_price* on *urbanization*, so we changed *gdp* into *log(gdp)* to observe it (represented in **Model 15**). Then, we are also done for adding log-linear, linear-log, and log-log models (represented by **Model 16, Model 17,** and **Model 18**, respectively) to interpret the percentage changes.

At the end, we implemented draws to see what model fits best with our data before doing some casual interpretation.

# 5. Our Results and Tests

## 5.1 Simple regressions

The principle is simple, compute the regression of the urbanization as a function of *gas_price*, adding to each new regression a new control variable to see if it has a particular impact.

A) *urbanization = beta0 + beta1 * gas_price*

*Table 5.1.4 Simple regression results*

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| intercept | 81.8584 | 0.886 | 92.433 | 0.000 | 80.103 | 83.613 |
| gas_price | -0.9004 | 1.047 | -0.860 | 0.392 | -2.976 | 1.175 |

In this case, as the P-Value of the *gas_price* is high and the confidence interval includes 0, we can conclude that, in this configuration, *gas_price* has no influence on urbanization. The estimation is biased because we're not taking into account some important omitted variables.

B) *urbanization = beta0 + beta1 * gas_price + beta2 * gdp*

*Table 5.1.2 Simple regression results adding gdp*

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| intercept | 74.3860 | 0.984 | 75.628 | 0.000 | 72.437 | 76.335 |
| gas_price | -4.1667 | 0.826 | -5.044 | 0.000 | -5.804 | -2.530 |
| gdp | 0.0003 | 2.69e-05 | 10.025 | 0.000 | 0.000 | 0.000 |

The wealth of a country is a determinant to know if the population in a country would afford to buy high price gas or not. Adding the GDP per capita finally improves our regression as the P-Value of the *gas_price* is near 0 and because it reveals the negative impact of *gas_price* on the urbanization rate.

Indeed, if the *gas_price* decreases more people can afford to buy some to drive to work. Then the urbanization rate increases as urban areas are spreading around the centre of towns.

*C) urbanization = beta0 + beta1 * gas_price + beta2 * gdp + beta3 * num_car*

*Table 5.1.3 Simple regression results adding num_car*

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| intercept | 74.6572 | 0.585 | 127.671 | 0.000 | 73.498 | 75.816 |
| gas_price | 0.4244 | 0.588 | 0.722 | 0.472 | -0.741 | 1.590 |
| gdp | 9.196e-05 | 2.03e-05 | 4.535 | 0.000 | 5.18e-05 | 0.000 |
| num_car | 0.0335 | 0.002 | 14.167 | 0.000 | 0.029 | 0.038 |

After adding *num_car*, the influence of *gas_price* becomes negligible. This might be because *gas_price* has a high correlation with the *num_car*, and the *num_car* has a higher impact on *urbanization*. The model is also biased because other omitted independent variables are not taken into account in this model and are not negligible like the influence of the country or the year. Indeed, for example, an oil producing country can have a low *gas_price* and few cars while another country can have a low *gas_price* and many cars.

D) *urbanization = beta0 + beta1 * gas_price + beta2 * gdp + beta3 * num_car + beta4 * oil_consuption + beta5 * industry + beta6 * services*

*Tables 5.1.4 Simple regression results adding the tree last control variables*

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| intercept | 73.3594 | 0.532 | 137.778 | 0.000 | 72.304 | 74.415 |
| gas_price | 0.0972 | 0.500 | 0.194 | 0.846 | -0.895 | 1.089 |
| gdp | 0.0002 | 1.97e-05 | 7.928 | 0.000 | 0.000 | 0.000 |
| num_car | -0.0556 | 0.014 | -4.085 | 0.000 | -0.083 | -0.029 |
| oil_consumption | 0.0011 | 0.000 | 6.614 | 0.000 | 0.001 | 0.001 |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| intercept | 73.8619 | 2.285 | 32.319 | 0.000 | 69.330 | 78.394 |
| gas_price | 0.2957 | 0.556 | 0.532 | 0.596 | -0.807 | 1.398 |
| gdp | 0.0001 | 2.22e-05 | 6.679 | 0.000 | 0.000 | 0.000 |
| num_car | -0.0601 | 0.016 | -3.848 | 0.000 | -0.091 | -0.029 |
| oil_consumption | 0.0011 | 0.000 | 5.943 | 0.000 | 0.001 | 0.002 |
| industry | -0.0342 | 0.059 | -0.579 | 0.564 | -0.152 | 0.083 |
| services | 0.0064 | 0.021 | 0.305 | 0.761 | -0.035 | 0.048 |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| intercept | 74.4084 | 1.416 | 52.546 | 0.000 | 71.601 | 77.216 |
| gas_price | 0.2815 | 0.552 | 0.510 | 0.611 | -0.812 | 1.375 |
| gdp | 0.0002 | 2.12e-05 | 7.114 | 0.000 | 0.000 | 0.000 |
| num_car | -0.0611 | 0.015 | -4.004 | 0.000 | -0.091 | -0.031 |
| oil_consumption | 0.0011 | 0.000 | 6.200 | 0.000 | 0.001 | 0.002 |
| industry | -0.0422 | 0.053 | -0.800 | 0.426 | -0.147 | 0.062 |

*Table 5.1.5 Simple regression models*

| Variables | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 |
|---|---|---|---|---|---|---|
| *intercept* | *81.85** (0.88)* | *74.39** (0.98)* | *74.66** (0.58)* | *73.36** (0.53)* | *74.41** (1.42)* | *73.86** (2.29)* |
| *gas_price* | *-0.90 (1.04))* | *-4.17** (0.83)* | *0.42 (0.59)* | *0.10 (0.50)* | *0.28 (0.55)* | *0.30 (0.56)* |
| *gdp* | | *0.01e-1** (2.68e-05)* | *9.20e-05** (2.03e-05)* | *0.01e-1** (1.97e-05)* | *0.00** (2.12e-05)* | *0.01e-2** (2.22e-05)* |
| *num_car* | | | *0.03** (0.00)* | *-0.06** (0.01)* | *-0.06** (0.02)* | *-0.06** (0.02)* |
| *oil_consumption* | | | | *0.01e-1** (0.01e-1)* | *0.01e-1** (0.01e-1)* | *0.01e-2** (0.01e-2)* |
| *industry* | | | | | *-0.04 (0.05)* | *-0.03 (0.06)* |
| *services* | | | | | | *0.06e-1 (0.02)* |

## 5.2 Regressions with entity and time fixed effects

When we include the fixed effects into our model by mixing ‘n-1 binary regressors’ and ‘t-1 binary regressors’, we receive the regression results shown in Table 5.2, **Model 7** and **Model 14**. We also added in Table 5.2 three previous models to make some comparisons.

*Table 5.2. Results of including fixed effects*

| Variables | Model 2 | Model 3 | Model 6 | Model 7 | Model 14 |
|---|---|---|---|---|---|
| *intercept* | *74.39** (0.98)* | *4.66** (0.58)* | *73.86** (2.29)* | *79.62** (3.15)* | *78.53** (2.94)* |
| *gas_price* | *-4.17** (0.83)* | *0.42 (0.59)* | *0.30 (0.56)* | *2.22** (0.49)* | *1.37** (0.45)* |
| *gdp* | *0.01e-1** (2.68e-05)* | *9.20e-05** (2.03e-05)* | *0.01e-2** (2.22e-05)* | *1.35e-05 (1.59e-05)* | *-3.39e-05 (1.74e-05)* |
| *num_car* | | *0.03** (0.00)* | *-0.06** (0.02)* | *-0.03** (0.01)* | *-0.04** (0.00)* |
| *oil_consumption* | | | *0.01e-2** (0.01e-2)* | *0.001** (0.000)* | *0.001** (0.0001)* |
| *industry* | | | *-0.03 (0.06)* | *-0.40** (0.06)* | *-0.21** (0.06)* |
| *services* | | | *0.06e-1 (0.02)* | *0.03* (0.01)* | *0.01 (0.01)* |
| *is_Canada* | | | | *5.16 (2.80)* | *3.14 (2.71)* |

| | | | | | |
|---|---|---|---|---|---|
| *is_France* | | | | *0.63*<br>*(2.72)* | *-0.003*<br>*(2.60)* |
| *is_UK* | | | | *4.70*<br>*(2.80)* | *3.64*<br>*(2.67)* |
| *1999_2005* | | | | | *1.14***<br>*(0.27)* |
| *2006_2012* | | | | | *2.44***<br>*( 0.45)* |
| *2013_2019* | | | | | *3.28***<br>*(0.54)* |
| ** implies that this coefficient is statistically significantly different from 0 at 5% of significant level.*<br>*** implies that this coefficient is statistically significantly different from 0 at 1% of significant level.* | | | | | |

Compared to the previous model **Model 6**, **Model 7** includes three additional binary variables that represent the unchanged factors over the years. By doing this, in the **Model 7** the coefficient of *gas_price* has become statistically significant. This proved that there are some omitted variables that do not change over time. In Figure 6.1, we can see that the model with the entity time effects fits to our data better than every previous model.

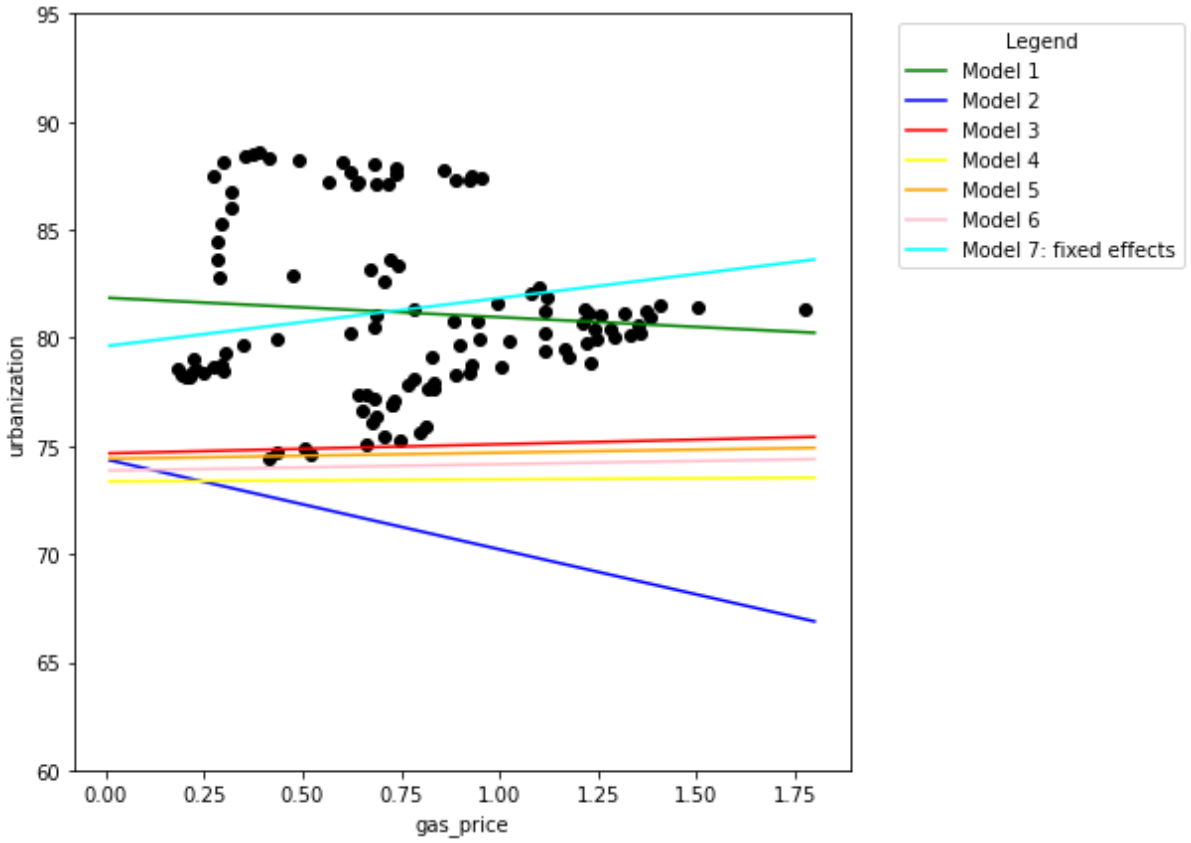


*Figure 5.1. Regression lines of model from 1 to 7*

In terms of time fixed effects, we added 3 binary variables to receive **Model 14**. In this model, the entity fixed effects are still included. By looking at the coefficients of these variables because they are statistically significant, we can conclude that there are some factors that do not change over the countries. To make sure whether we

should keep the entity and time fixed effects, we took some hypothesis tests, which are shown in Table 5.3.

*Table 5.3 Tests on entity and time fixed effects*

| Tests | F-statistic | p-value |
|---|---|---|
| *is_Canada = 0, is_France = 0, is_UK = 0* | 107.64 | 4.78e-31 |
| *1999_2005 = 0, 2006_2012 = 0, 2013_2019 = 0* | 18.04 | 2.02e-9 |
| *is_Canada = 0, is_France = 0, is_UK = 0, 1999_2005 = 0, 2006_2012 = 0, 2013_2019 = 0* | 178.08 | 9.84e-51 |
| *All tests are performed based on **Model 14**.* | | |

By looking at the p-values of our tests, we can see that all p-values are very small (compared to 0). Therefore, we can accept that we do need to include these variables in our model to perform the fixed effects.

## 5.3 Non-linear regressions

Before doing any non-linear regressions, we tried to draw the data we have to see the relation between the explained variable and the variable of interest. Figure 5.2 shows the data we have. From this figure, we can see that linear regression might be not the best choice to fit the data. Therefore, we continue by considering the non-linear regression models. First we tried on polynomial forms then we also did the regressions on logarithm forms.

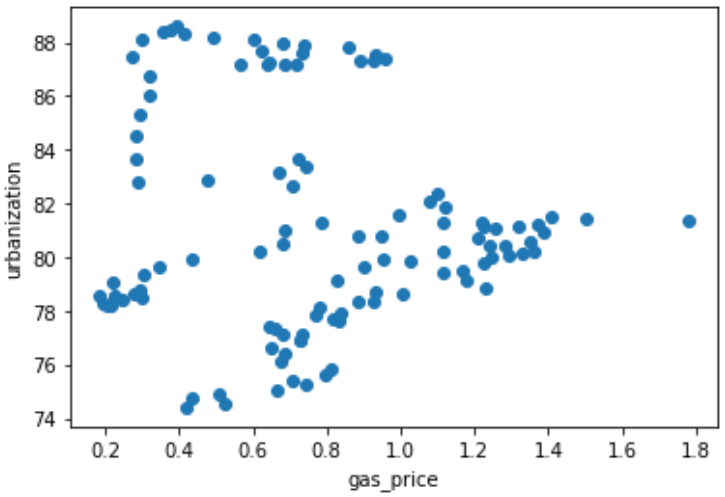


*Figure 5.2. The distribution of data*

### 5.3.1 Polynomials

The results of polynomials regressions are shown in Table 5.3. Firstly, we make the quadratic regression by adding a new variable, the gas price squared

(*gas_price^2*). **Model 112** shows us that by adding *gas_price^2* the model does not change much. In particular, the coefficients of all variables changed a little bit, while the coefficient of *gas_price^2* is not statistically significant. Therefore, we can conclude that the quadratic regression does not fit well our data.

We continued to try with the cubic regression, the results are shown in **Model 11** in Table 5.4. Now, by including *gas_price^3* into the model, thet model changed a lot. The coefficients of non-linear parts now are statistically significant. These results imply that the cubic regression fits well with the data. To make sure one of the coefficients of *gas_price^2* and *gas_price^3* is statistically different from zero, we took the joint hypothesis tests, shown in Table 6.5. The p-value of the hypothesis test on the null hypothesis 'the coefficient of *gas_price^2 and gas_price^3 = 0*' is equal to 0.001 (smaller than 1%). Then, we can reject this hypothesis at 1% of significant level. Therefore, we kept these variables in our model.

*Table 5.4 Polynomials regression results*

| Variables | Model 7 | Model 14 | Model 112 | Model 11 | Model 13 | Model 19 |
|---|---|---|---|---|---|---|
| *intercept* | *79.62** (3.15)* | *78.53** (2.94)* | *77.91** (3.01)* | *79.64** (2.87)* | *83.45** (2.24)* | *71.19** (2.91)* |
| *gas_price* | *2.22** (0.49)* | *1.37** (0.45)* | *2.34* (1.14)* | *-7.55** (2.88)* | *-5.17 (2.90)* | *40.81** (13.64)* |
| *gas_price^2* | | | *-0.56 (0.61)* | *11.63** (3.35)* | *9.32** (3.30)* | *-45.56* (20.59)* |
| *gas_price^3* | | | | *-4.40** (1.19)* | *-3.70** (1.16)* | *18.41 (9.33)* |
| *gdp* | *1.35e-05 (1.59e-05)* | *-3.39e-05 (1.74e-05)* | *-3.14e-05 (1.77e-05)* | *-3.42e-5* (1.66e-5)* | *-4.21e-5* (1.65e-5)* | *-1.69e-5 (1.49e-05)* |
| *num_car* | *-0.03** (0.01)* | *-0.04** (0.00)* | *-0.04** (0.01)* | *-0.03** (0.01)* | *-0.001 (0.02)* | *-0.06** (0.02)* |
| *oil_consumption* | *0.001** (0.000)* | *0.001** (0.0001)* | *0.001** (0.00)* | *0.001** (0.00)* | *0.001** (0.00)* | *0.007** (0.001)* |
| *industry* | *-0.40** (0.06)* | *-0.21** (0.06)* | *-0.19** (0.06)* | *-0.31** (0.06)* | *-0.23** (0.06)* | *-0.23** (0.06)* |
| *services* | *0.03* (0.01)* | *0.01 (0.01)* | *0.01 (0.01)* | *0.007 (0.01)* | *0.008 (0.01)* | *0.02** (0.01)* |
| *is_Canada* | *5.16 (2.80)* | *3.14 (2.71)* | *2.78 (2.74)* | *6.61* (2.78)* | *-0.38 (0.58)* | *0.53 (0.51)* |
| *is_France* | *0.63 (2.72)* | *-0.003 (2.60)* | *-0.29 (2.62)* | *2.85 (2.61)* | *-3.89** (0.28)* | *-3.30** (0.25)* |
| *is_UK* | *4.70 (2.80)* | *3.64 (2.67)* | *3.38 (2.69)* | *6.52* (2.67)* | | |
| *is_USA* | | | | | *-18.14** (4.16)* | *-103.36** (21.58)* |
| *1999_2005* | | *1.14** (0.27)* | *1.11** (0.27)* | *0.93** (0.26)* | *0.87** (0.24)* | *0.79** (0.20)* |
| *2006_2012* | | *2.44** ( 0.45)* | *2.37** (0.46)* | *2.11** (0.44)* | *2.30** (0.42)* | *2.10** (0.35)* |

| | | | | | | |
|---|---|---|---|---|---|---|
| *2013_2019* | | *3.28** (0.54)* | *3.27** (0.54)* | *2.83** (0.53)* | *3.00** (0.50)* | *2.95** (0.41)* |
| *is_USA X gas_price* | | | | | *52.59* (24.96)* | *383.66** (111.47)* |
| *is_USA X gas_price^2* | | | | | *-106.39* (40.76)* | *-414.82* (167.66)* |
| *is_USA X gas_price^3* | | | | | *62.04** (21.27)* | *154.29* (76.74)* |
| *oil_consumption X gas_price* | | | | | | *-0.02** (0.006)* |
| *oil_consumption X gas_price^2* | | | | | | *0.02** (0.01)* |
| *oil_consumption X gas_price^3* | | | | | | *-0.009* (0.00)* |
| ** implies that this coefficient is statistically significantly different from 0 at 5% of significant level.*<br>*** implies that this coefficient is statistically significantly different from 0 at 1% of significant level.* | | | | | | |

We can see the above conclusion more clearly in Figure 5.3, while the fixed effects line and the quadratic line are almost the same, the cubic line is different and fits better with our data.

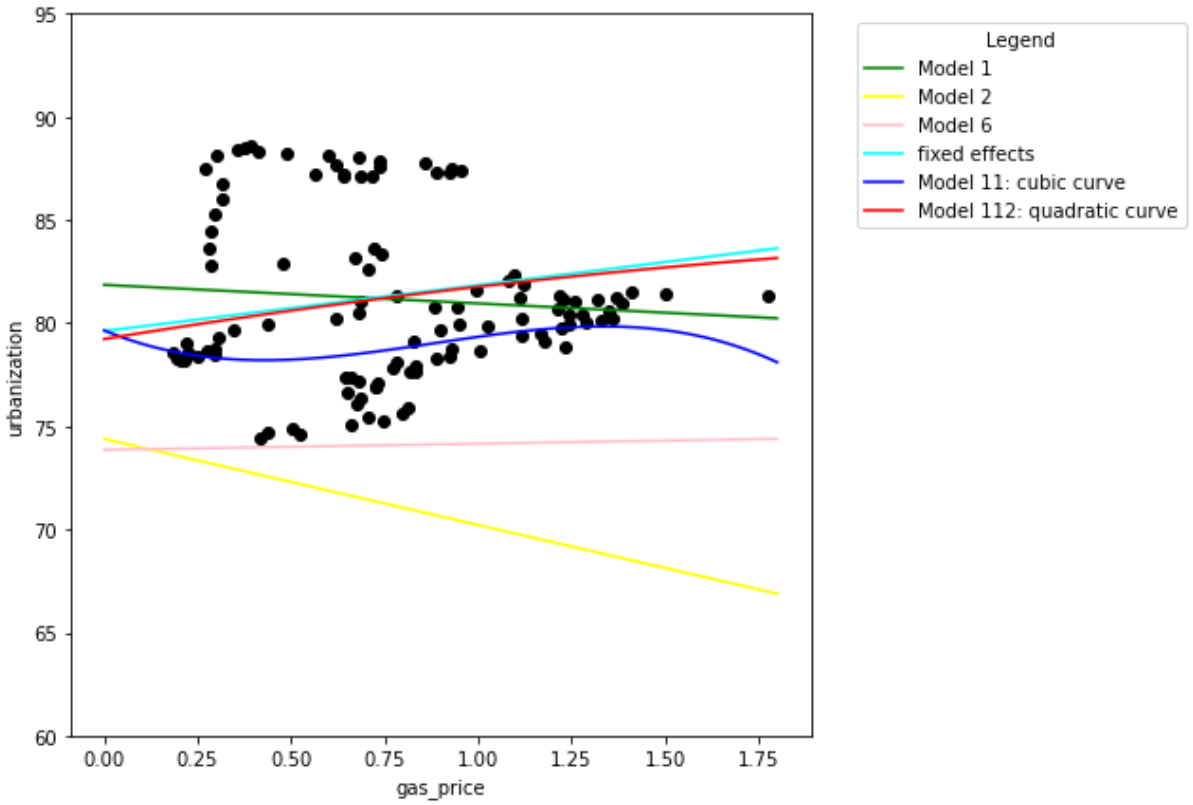


*Figure 5.3 Simple, quadratic, cubic regression lines*

As mentioned above, we included some cross variables between gas price and the variable *is_USA* and *oil_consumption*, the results are shown in **Model 13** and **Model 19**. First, regarding **Model 13** when we added three cross variables of *is_USA* and *gas_price,* we can see that the coefficients of these three variables are not statistically significant. Or in other words, they are not statistically different from 0. Therefore, there might be no interaction between *is_USA* and *gas_price*, but we can not be sure until now. We continue to add some other cross variables to see more results.

By looking at **Model 19**, we can see that the coefficient of all cross variables and of the variable of interest now become statistically significant. We should include these cross variables in the model. To be more sure, we also compute the F-statistic of some joint hypothesis tests and the results are shown in Table 5.5. The p-values of all tests are very small compared to 0, thus, we can reject all of these hypotheses. In other words, we should include these variables in our model.

*Table 5.5 Tests on non-linear model and cross variables*

| Tests | F-statistic | p-value |
|---|---|---|
| *gas_price^2 = 0, gas_price^3 = 0 [1]* | 7.30 | 0.001 |
| *is_USA X gas_price = 0, is_USA X gas_price^2 = 0, is_USA X gas_price^3 = 0 [2]* | 27.52 | 9.27e-13 |
| *oil_consumption X gas_price = 0, oil_consumption X gas_price^2 = 0, oil_consumption X gas_price^3 = 0 [2]* | 24.81 | 8.04e-12 |
| *All cross variables are equal to 0 [2]* | 35.24 | 1.01e-21 |
| *[1] Tests are performed based on **Model 11**.*<br>*[2] Tests are performed based on **Model 14**.* | | |

### 5.3.2 Logarithms

Until now, **Model 19** is the best model we have built. However, we also tried to make some logarithm regressions to see whether we can find the better model or not. The regression results of some logarithm regressions are shown in Table 5.6. The **Model 16** gives us that the coefficient of *gas_price* is not statistically significant. Therefore, we can not interpret the impact of the change in one unit of gas price on the percentage change in *urbanization*. Meanwhile, **Model 17** and **Model 18** show us the better result when the coefficients of variable of interest are statistically significant.

*Table 5.6. Results of logarithm regressions*

| Variables | Model 14 | Model 11 | Model 16 *ln(urbanization)* | Model 17 | Model 18 *ln(urbanization)* |
|---|---|---|---|---|---|
| *intercept* | *78.53** (2.94)* | *79.64** (2.87)* | *4.37** (0.03)* | *79.53** (2.95)* | *4.39** (0.03)* |
| *gas_price* | *1.37** (0.45)* | *-7.55** (2.88)* | *0.016 (0.00)* | | |
| *gas_price^2* | | *11.63** (3.35)* | | | |
| *gas_price^3* | | *-4.40** (1.19)* | | | |
| *ln(gas_price)* | | | | *0.76* (0.29)* | *0.009* (0.004)* |
| *gdp* | *-3.39e-05 (1.74e-05)* | *-3.42e-5* (1.66e-5)* | *-4.00e-7 (2.15e-7)* | *-3.12e-05 (1.76e-05)* | *-3.66e-07 (2.18e-07)* |
| *num_car* | *-0.04** (0.00)* | *-0.03** (0.01)* | *-0.0005** (0.00)* | *-0.04** (0.01)* | *-0.0006** (0.00)* |
| *oil_consumption* | *0.001** (0.0001)* | *0.001** (0.00)* | *1.31e-05** (1.98e-6)* | *0.001** (0.000)* | *1.29-e5** (2.01e-6)* |
| *industry* | *-0.21** (0.06)* | *-0.31** (0.06)* | *-0.002** (0.00)* | *-0.21** (0.06)* | *-0.002** (0.001)* |
| *services* | *0.01 (0.01)* | *0.007 (0.01)* | *0.0001 (0.00)* | *0.01 (0.01)* | *0.0002 (0.00)* |
| *is_Canada* | *3.14 (2.71)* | *6.61* (2.78)* | *0.02 (0.03)* | *2.35 (2.71)* | *0.01 (0.03)* |
| *is_France* | *-0.003 (2.60)* | *2.85 (2.61)* | *-0.01 ( 0.03)* | *-0.76 (2.63)* | *-0.02 (0.03)* |
| *is_UK* | *3.64 (2.67)* | *6.52* (2.67)* | *0.02 (0.03)* | *2.84 (2.69)* | *0.01 (0.03)* |
| *1999_2005* | *1.14** (0.27)* | *0.93** (0.26)* | *0.014** (0.00)* | *1.18** (0.27)* | *0.01** (0.00)* |
| *2006_2012* | *2.44** ( 0.45)* | *2.11** (0.44)* | *0.03** (0.00)* | *2.53** (0.46)* | *0.03** ( 0.00)* |
| *2013_2019* | *3.28** (0.54)* | *2.83** (0.53)* | *0.04** (0.00)* | *3.48** (0.54)* | *0.04** (0.00)* |

** implies that this coefficient is statistically significantly different from 0 at 5% of significant level.*
*** implies that this coefficient is statistically significantly different from 0 at 1% of significant level.*

When looking at the regression lines in Figure 5.4, we can see that all models (16, 18, 19) are fitted with our data. However, **Model 16** has the coefficient of variable of interest that is not statistically significant. Therefore, we keep only **Model 18** and **Model 19** for causal interpretations.

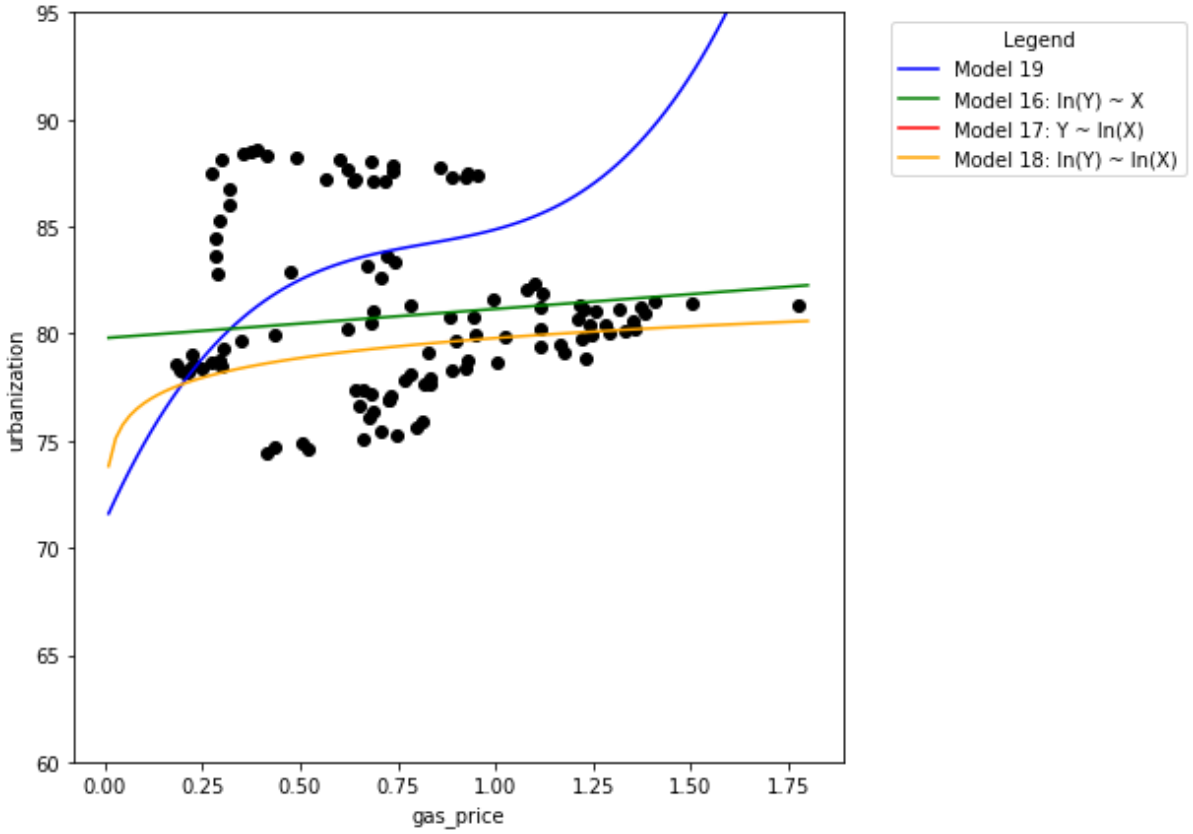


*Figure 5.4 Regression lines*

## 5.4 Final models and causal interpretation

In conclusion, we keep only two models for causal interpretation, **Model 19 and Model 18**. They are respectively represented below.

| | | | | |
|---|---|---|---|---|
| ***urbanization =*** | ***71.19 (2.91)*** | ***40.81*gas_price (13.64)*** | ***-45.56*gas_price^2 (20.59)*** | ***18.41*gas_price^3 (9.33)*** |
| | ***-1.69e-5*gdp (1.49e-05)*** | ***-0.06*num_car (0.02)*** | ***0.007* oil_consumption (0.001)*** | ***-0.23*industry (0.06)*** |
| | ***0.02*services (0.01)*** | ***0.53*is_Canada (0.51)*** | ***-3.30*is_France (0.25)*** | ***-103.36*is_USA (21.58)*** |
| | ***0.79*(1999_2005) (0.20)*** | ***2.10*(2006_2012) (0.35)*** | ***2.95*(2013_2019) (0.41)*** | ***383.66*(is_USA X gas_price) (111.47)*** |
| | ***-414.82* (is_USA X gas_price^2) (167.66)*** | ***154.29*(is_USA X gas_price^3) (76.74)*** | ***-0.02*(oil_consumption X gas_price) (0.006)*** | ***0.02*(oil_consumption X gas_price^2) (0.01)*** |
| | ***-0.009*(oil_consumption X gas_price^3) (0.00)*** | | | |

| | | | | |
|---|---|---|---|---|
| ***ln(urbanization) =*** | ***4.39 (0.03)*** | ***0.009*ln(gas_price) (0.004)*** | ***-3.66e-07*gdp (2.18e-07)*** | ***-0.0006*num_car (0.00)*** |
| | ***1.29-e5*oil_consumption (2.01e-6)*** | ***-0.002*industry (0.001)*** | ***0.0002*services (0.00)*** | ***0.01*is_Canada (0.03)*** |
| | ***-0.02*is_France (0.03)*** | ***0.01*is_UK (0.03)*** | ***0.01*(1999_2005) (0.00)*** | ***0.03*(2006_2012) ( 0.00)*** |
| | ***0.04*(2013_2019) (0.00)*** | | | |

Based on **Model 19,** we can interpret how *urbanization* changes when the *gas_price* increases by one unit. In particular, when the gas price increases from 0.1 dollar per liter to 0.2 dollar per liter, the urbanization will increase 2.84%. In addition, when the gas price increases from 1 dollar per liter to 1.1 dollar per liter, the urbanization will increase 0.63%. This difference is because of the non-linear model.

In **Model 18**, we can interpret how many percent of *urbanization* can be changed when the gas_price increases by one percent. To be more precise, when the *gas_price* increases one percent, this leads to a 0.009 percent increase in *urbanization*.

# Conclusion

Motivated by the considerable discrepancies that subsist in our world, when it comes to both the price of petrol and the percentage of urbanized population, from a country to another, our project led us to create more than twenty models, each new one improving from the previous one. In the end the two models detailed in part 5.4 of our report highlight the correlation between the price of petrol in a country and its level of urbanization, and the various variables they need. In both cases, we can tell how much the former would increase if we were to change the latter. In the end, we can conclude that the price of petrol does have an influence over the level of urbanization of a country. In the next few decades, it will be interesting to see how urbanization changes on a worldwide scale, since the use of oil should decline. When petrol natural reserves dwindle, mankind will have either to invent new means of transportation, using other sources of energy, or to rethink its localization of population. One thing that we would have included if we had considered less closely developed countries is the development of public transportation in the countries, since public transportation is a way for a lot of people to commute, without using a car, and therefore to use way less oil, although some kinds of public transportation, like buses do use petrol. To look at the big picture, it is quite obvious that the unhealthy entanglement petrol has in our society will eventually lead us to a very important choice, between changing the way we live, or keeping everything the way they are and wishing for technological improvements to save us.

# Bibliography


[1] Effect of High Gasoline Prices on Low-Density Housing Development | Leadership and Management in Engineering | Vol 13, No 3

[2] Are Gasoline Prices a Factor in Residential Relocation Decisions? Preliminary Findings from the American Housing Survey, 1996–2008 - Guangqing Chi, Jamie Boydstun, 2017

[3] Brody, J. E. (2014, December 11). Gasoline prices make suburban houses more valuable. The Washington Post.

[4] Do Gasoline Prices Affect Residential Property Values? Aeaweb

[5] Chen M, Zhang H, Liu W, Zhang W. The global pattern of urbanization and economic growth: evidence from the last three decades. PLoS One. 2014;9(8):e103799. Published 2014 Aug 6. doi:10.1371/journal.pone.0103799

[6] The more, the merrier? Urbanization and regional GDP growth in Europe over the 20th century. Kerstin Enflo, Anna Missiaia, Joan Rosés. University of Southern Denmark, Odense 25-26 April 2019

[7] Kaza N, Knaap GJ, Knaap I, Lewis R. Peak oil, urban form, and public health: exploring the connections. Am J Public Health. 2011;101(9):1598-1606. doi:10.2105/AJPH.2011.300192