Cars fuel consumption in the 1970s

by Dávid Szabados (ID: 2302806)

Introduction

Currently, fossil fuel consumptions (e.g.: coal, oil) are still high in some countries and have a problem to adapt to newer resources. This can be seen in Germany[1], where the fear of nuclear power plants lead to a negative trend. Instead of changing the fossil power plants to renewable ones, they are changing nuclear power plants instead. As fossil fuels, especially oil/petrol, are still one of the main resources that makes many economies run, it should be interesting to look at the cars of the past, and check, what influenced the fuel consumption for them. Another goal of this project is to get a general assumption about the car market of the USA during the 1970s.

Data

This research will use the "Automotive Fuel Economy" dataset from Maven Analytics. This dataset includes the fuel economy in mpg (miles per gallon) for almost 400 cars, that have been sold in the US during the 70s and 80s, but we will only focus on one decade, the 70s' sales. With this, we hopefully exclude time-series effects of the 80s, from the analysis.

The dataset includes one table with the following variables:

- mpg: The fuel economy of the car in terms of miles travelled per gallon of gasoline | *<u>Dependent Variable</u>*
- horsepower: Horsepower is a measure of power the engine produces | *<u>Main Independent/Explanatory Variable</u>*
- cylinders: The number of cylinders in the car's engine
- displacement: The volume of air displaced by all the pistons of a piston engine
- weight: The total weight of the car
- acceleration: The time in seconds it takes for the car to reach 60 miles per hour
- model year: The year (in the 20th century) the car model was released. For example: 75 means the car was released in 1975.
- origin: The region where the car was manufactured. 1 - USA. 2 - Europe. 3 - Japan
- car name: The name of the car model.

The analysis will use mpg as the dependent variable and mainly horsepower as the explanatory variable to see, how they are related to each other, based on the given dataset. Missing values will be excluded from the analysis, as it would bias it, while possible extreme values will be included, as these kinds of sales can happen now-and then and should be dealt as an actual observation, rather than a mistake of the buyer.

The initial intuition is that mpg and horsepower are negatively associated, which means that a car's horsepower in the 70s is positively associated with its fuel consumption (mpg) (in other words, it can travel less distance with the same amount of fuel). In order to achieve more power, more resources are needed. On the contrary, an engine which has more horsepower suggests

[1] https://www.voanews.com/a/german-finance-minister-casts-doubt-on-2030-coal-exit/7337035.html [Accessed: 12/18/2023]

that it uses the fuel more efficiently. Thus, for the same amount of fuel, it can use more power. In any case, this research project hypothesizes a relationship between them and correlation of 0 (no relationship) seems very unlikely.

In **App. 1a**, the mean of the mpg variable (21.084) is higher than the 50^{th} percentile, the median (20), due to high values. Thus, the distribution is right skewed, as it is also seen in the **App. 1b** histogram. This variable underwent logarithmic transformation in order to analyze the relative change of it, which a seemingly normal distribution can be seen in **App. 1c.**

Model

Three models will be used to see certain aspects of the dataset. The first will be a simple regression, to get a basic idea of the unconditional association between the main dependent and explanatory variables. Next, further variables will be added to the regression, to see how other variables are associated with the mpg. At the end, the cars' origin will be investigated to acquire more information about U.S. made cars. Every model will use a HC1 covariance, to ensure robustness.

Unconditional Linear Regression

The first model was a simple **log-level** regression between the ln(mpg) and horsepower variables, without any further conditionals, to analyze their unconditional association. This and further charts are made using ggplot2 in python. Looking at the **App. 2** scatter plot, it can be seen, that there is a negative association between them. It means, that on average, a car with more horsepower is associated with travelling **less** miles, with fuel quantity remaining constant. This suggests the first null hypothesis should be rejected. On the right side of the graph, the points are sparser, as it seems during the 70s, not many cars have been sold in the USA with a horsepower of more than 190-200. On **App. SUM**, the horsepower correlation coefficient is significant, as it has a p-value of less than 1%, **for every 10 more horsepower** a car has, **it is associated with traveling 6% less miles** with the same amount of fuel. The equation looks like the following:

$$(\ln mpg)^E = 3.713 + (-0.006)horsepower$$

Conditional Multiple Linear Regression

Of course, the horsepower of the engine is not the only attribute a car has. It also has a weight (as it is the American market and the fuel consumption is also in mpg, weight is measured in pounds (1pound ~ 0.45kg)) and the engine can have different numbers of cylinders and air displacement values. Air displacement is based on the numbers of cylinders in the engine, so only the former one is used to avoid biases between the conditional variables. The next multiple linear regression will include horsepower (like previously), displacement and weight. On **App. SUM** the **R^2 for this regression is 0.824**, which means that this regression approximates 82% of the variance in the dependent variable. This model is stronger than the previous simple linear regression. There is a large difference, but still a significant value for horsepower at a -0.002 for every additional horsepower (**for every 10 more horsepower** a car has, **it can travel 2% less miles** with the same amount of fuel). While displacement is not that significant, weight is. **Every 100 additional pounds** of weight is associated with **the car travelling 2% less miles**.

This is logical, as more power is needed to move heavier objects. This number can be seen more precisely on **App. 2b.** The equation looks like the following:

$$(\ln mpg)^E = 3.926 + (-0.002)horsepower + (-0.000)displacement + (-0.0002)weight$$

Conditional Regression Based on Origin

The U.S. has a free market, so their cars are not solely from their own country; the vehicles can be imported from Europe and from Japan. Taking a look at **App. 2c**, the previously shown scatterplot is now updated, where the origin of the vehicle can be seen as well. The non-USA cars are mostly grouped on the left side of the plot, which suggests these cars tend to have a better fuel consumption. Using U.S. origin as the reference category, two new columns are made, an *IsEU*, which denotes a binary value whether the car's origin is Europe or not, and an *IsJP*, which is the same, but for Japan. On **App. SUM**, if a car is originated from outside of the USA, it tends to be associated with a better fuel consumption rate. European and Japanese import cars tend to be associated with **travelling 11.5% and 17.5% more miles** respectively, with the same amount of fuel, than American cars. The equation looks like the following:

$$(\ln mpg)^E = 3.926 + (-0.002)horsepower + (0.115)IsEU + (0.175)IsJP$$

External Validity and Causal interpretation

To keep the external validity in mind, the interpretations are derived from the cars sold in the U.S. during the 1970s. While the general car attribute causations may be used for other car markets in the world, but only for the same time-period, the 70s. Technological advancements are made in various parts of the world, where qualitative outputs may holder greater importance than quantitative (regarding the effective usage of input resources, rather than just looking at the attributes of the end-product). The interpretations also shouldn't be used for

From the regressions, it can be seen, that fuel consumption heavily depends on the power of the engine (measured in horsepower), which sounds logical. As stated in the initial intuition: achieving higher power yields the need for more resources. While air displacement does not directly influence the fuel consumption of the car significantly, it is somewhat related to horsepower, as part of the air displacement is about the petrol-mixture outtake, which runs the engine. (The reason why displacement is included, while cylinders are not, is because cylinders are directly related to the displacement variable. Weight is also logical; if there would be a 100kg cement block, it is harder to push than a 50kg block.

The last interpretation is that U.S. cars tend to have a higher fuel consumption in the dataset. The 70s in the U.S. was the end of the golden era, where sustainability was not a popular subject in the world, and the consumer mindset took its root during the post-WWII economic boom. It did not matter how much petrol a car needed, as long as it is appealing and fast, the infamous attribute of capitalism in the U.S. The U.S. had a kind of car culture, which is vastly different from Europe or Japan, where for many cities had robust public transportation systems. This is still the main method of transportation in many European and Japanese cities, and thus more powerful cars with higher consumption rate were not as demanded, while in the U.S., for many people, public transportation is not an option and are heavily dependent to use car as the mean of transportation.

Conclusion

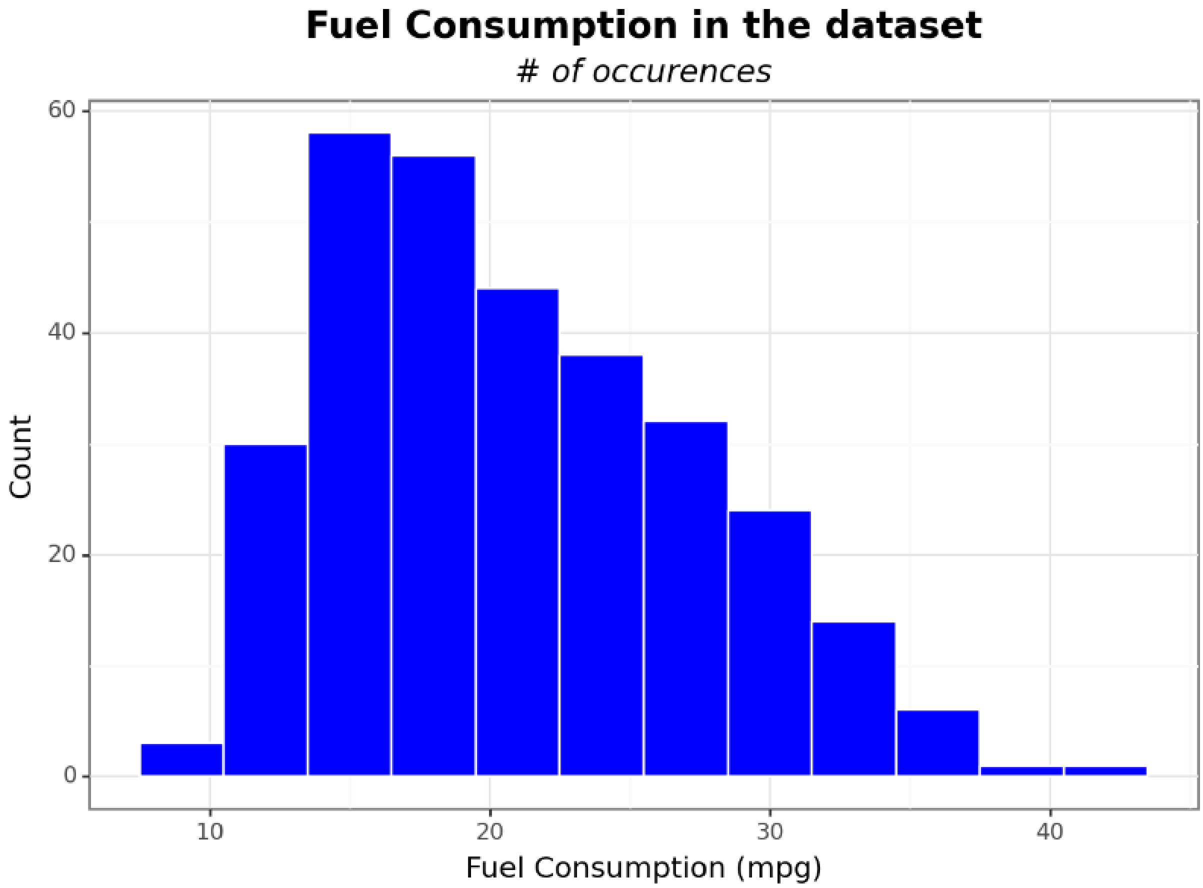
During the 70s, a car's fuel consumption was heavily dependent on the power of the engine and the weight of the car. A more powerful engine can transport more weight, but also will use up more resources as a result. U.S. models tend to have a much higher fuel consumption as compared to European or Japanese cars. The reason of this conclusion may come from the fact, that the dataset only includes lower horsepower imported cars, thus this conclusion should only be looked at for the U.S. market, as the data did not include values for imported cars with higher engine power. These conclusions and the previous causal interpretations should lead towards further car developments in the U.S. for the present time. Instead of consumerism, the market and firms should represent a more qualitative use of the resources used, to make a car and run it in an eco-friendlier manner, building higher quality engines, or using lighter and more durable materials for cars. Fuel consumption can be reduced in many ways, but for this, looking outside of the country and getting new ideas is necessary.

Appendix

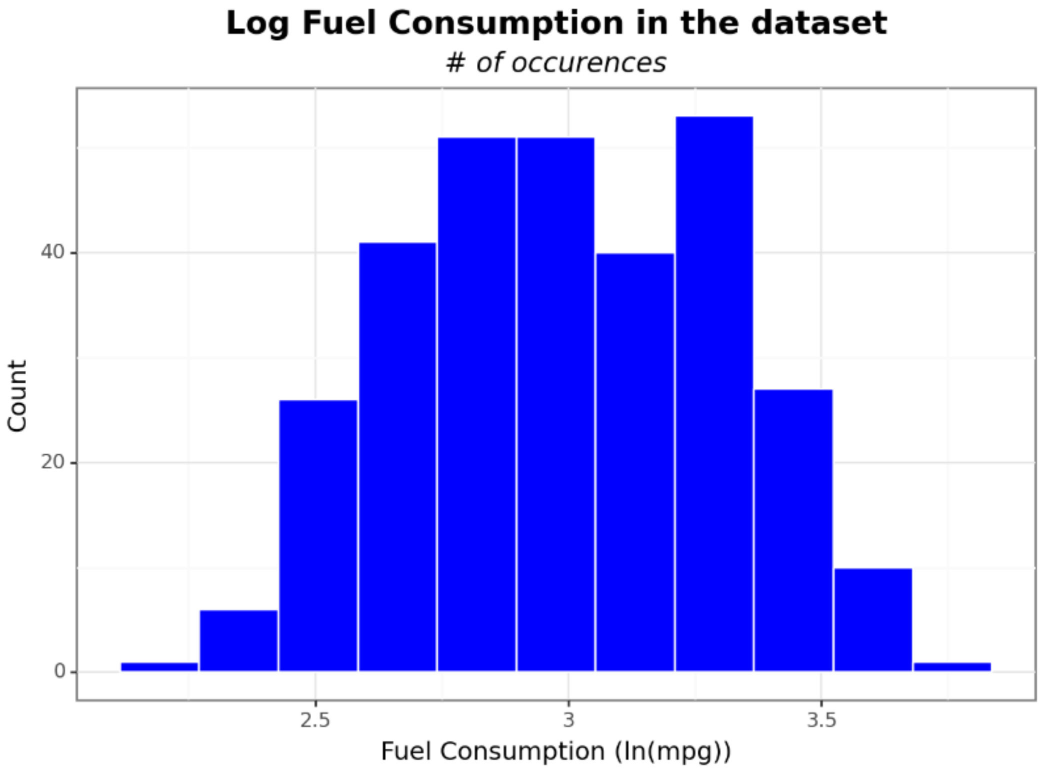
App. 1a

count	307.000
mean	21.084
std	6.482
min	9.000
25%	16.000
50%	20.000
75%	26.000
max	43.100

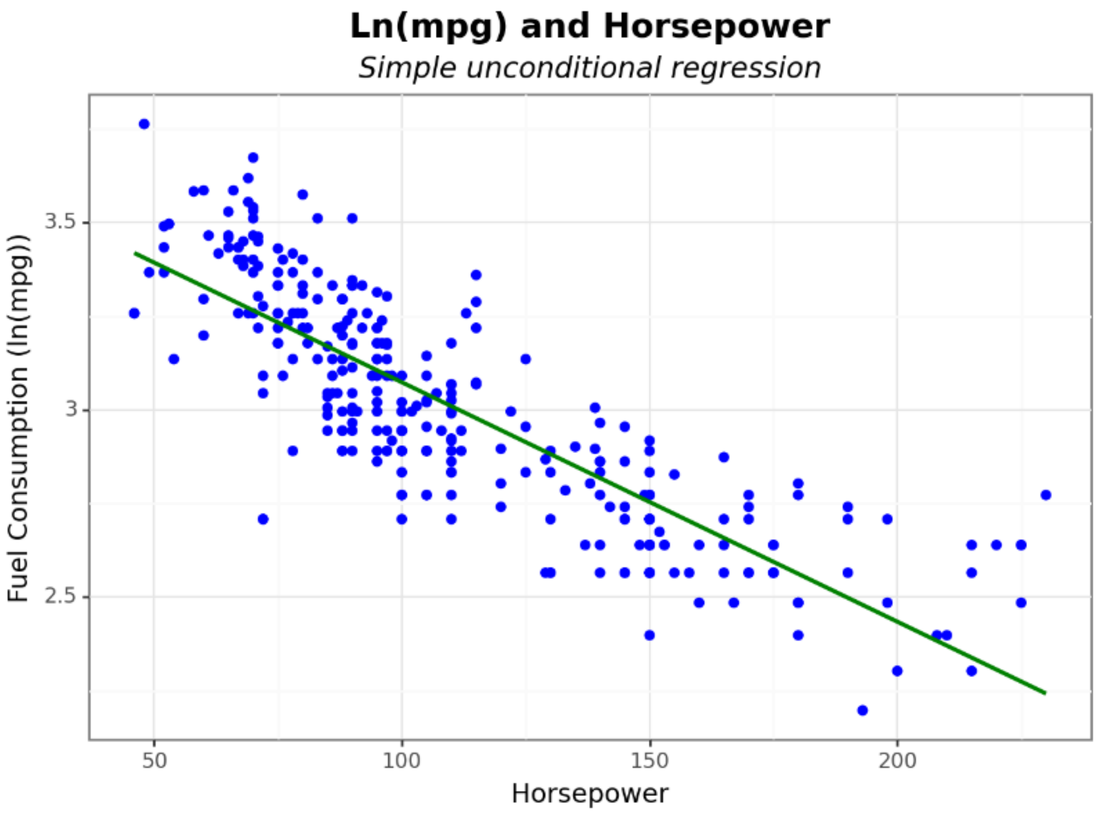
App. 1b



App. 1c


Log Fuel Consumption in the dataset
# of occurences
Count
40
20
0
2.5
3
3.5
Fuel Consumption (ln(mpg))

App. 2a


Ln(mpg) and Horsepower
Simple unconditional regression
Fuel Consumption (ln(mpg))
3.5
3
2.5
50
100
150
200
Horsepower

App. 2b

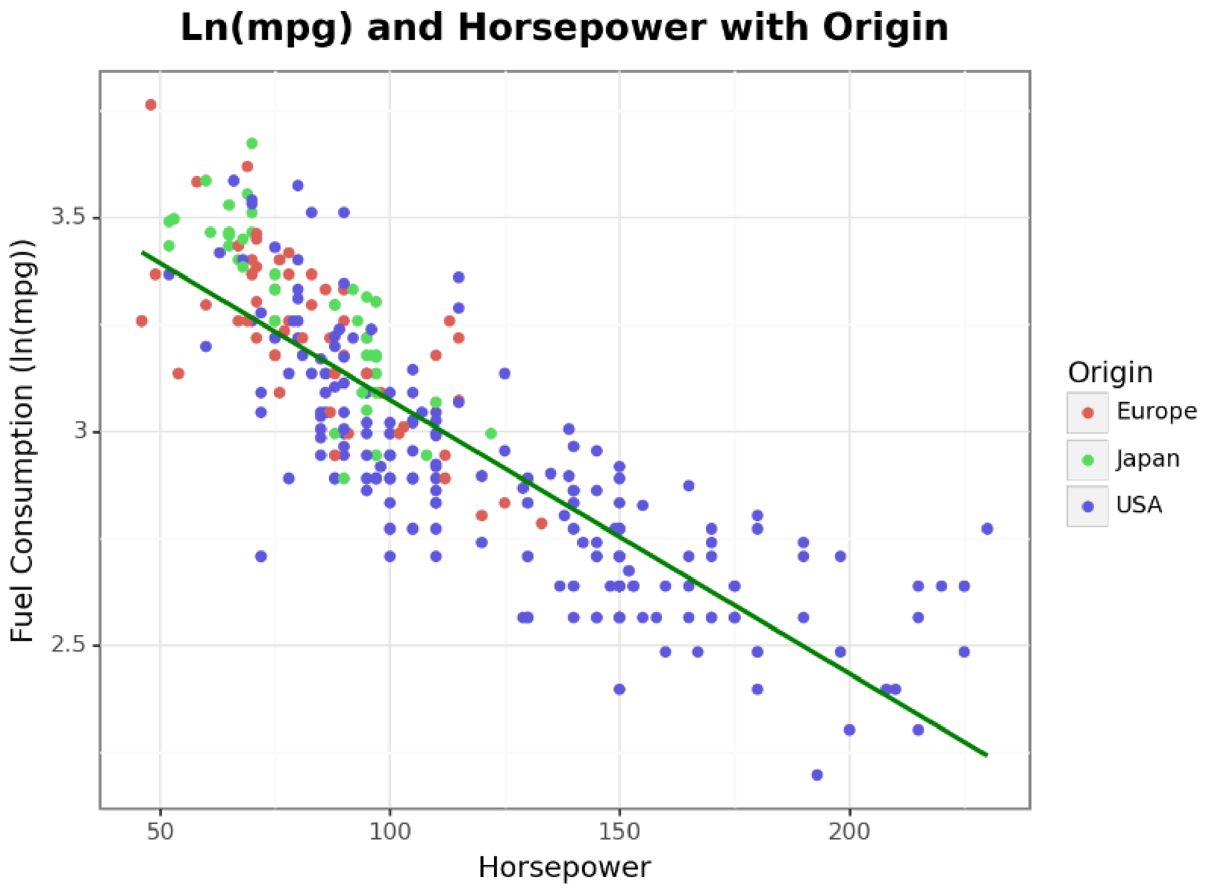
OLS Regression Results

Dep. Variable:	lnmpg	R-squared:	0.824
Model:	OLS	Adj. R-squared:	0.822
Method:	Least Squares	F-statistic:	472.0
Date:	Tue, 19 Dec 2023	Prob (F-statistic):	7.86e-114
Time:	17:14:56	Log-Likelihood:	193.98
No. Observations:	307	AIC:	-380.0
Df Residuals:	303	BIC:	-365.1
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.9264	0.040	97.981	0.000	3.848	4.005
horsepower	-0.0016	0.000	-3.834	0.000	-0.002	-0.001
displacement	-0.0002	0.000	-1.087	0.278	-0.001	0.000
weight	-0.0002	2.33e-05	-9.559	0.000	-0.000	-0.000

Omnibus:	2.357	Durbin-Watson:	1.003
Prob(Omnibus):	0.308	Jarque-Bera (JB):	2.280
Skew:	-0.076	Prob(JB):	0.320
Kurtosis:	3.394	Cond. No.	1.76e+04

App. 2c



App. SUM

		Dependent variable: lnmpg	
	(1)	(2)	(3)
Constant	3.713^{***}	3.926^{***}	3.575^{***}
	(0.034)	(0.045)	(0.044)
Horsepower	-0.006^{***}	-0.002^{***}	-0.006^{***}
	(0.000)	(0.001)	(0.000)
Air Displacement		-0.000	
		(0.000)	
Car Weight (pounds)		-0.000^{***}	
		(0.000)	
Europe Origin			0.115^{***}
			(0.026)
Japan Origin			0.175^{***}
			(0.027)
Observations	307	307	307
R^2	0.698	0.824	0.736
Adjusted R^2	0.697	0.822	0.734
Residual Std. Error	0.169 (df=305)	0.129 (df=303)	0.158 (df=303)
F Statistic	465.042^{***} (df=1; 305)	460.703^{***} (df=3; 303)	256.943^{***} (df=3; 303)
Note:			$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$