An Empirical Analysis of Vehicle Fuel Efficiency

Nicholas Lee, Mike Varner, Zachary Galante, and Mark Herrera

8/2/2022

Introduction

With recent world events, such as the war in Ukraine and supply chain disruptions, consumers have seen record high gas prices. Average gas prices in the US are "...on average at levels rarely seen in the last 50 years..." even when adjusting for inflation.[1] This has impacted consumer budgets across the country, and it is now more important than ever to have a fuel efficient vehicle.

While electric vehicles offer a way to reduce our collective dependence on gas, they remain niche products that comprise a small percentage of vehicle sales (4.6%)[2] and which many consumers cannot afford. Therefore, it is important to understand factors that contribute to gas car's mileage per gallon (MPG) as a measure of fuel efficiency.

To this end, we model factors that contribute to a car's MPG using data from the UCI Machine Learning Repository. We are keenly interested in the impact that a car's engine displacement has on fuel efficiency as manufacturers have direct control over engine design. Common wisdom suggests that vehicles with high displacement are less fuel efficient than lower displacement vehicles (ex. trucks vs sedans). We find evidence of a statistically significant negative relationship, that is robust to alternative specifications, between engine displacement and MPG on the order of ~(0.09) MPG/cubic inch of engine displacement (10% decrease in the average car's displacement yielding a 1.19 to 2.73 MPG increase).

Data and Methodology

Our study uses a dataset donated to the Statlib library at Carnegie Mellon University, collected in 1982 and used in the 1983 American Statistical Association (ASA) Exposition[3]. Documentation for these data is limited and consequently we've relied on a letter from the ASA[4]. Each row represents a car model sold in the years 1970 - 1982, with a total of 398 observations. Given the lack of sufficient documentation we can not confirm these data are observational.

We cleaned the data by removing 6 observations with missing values for horsepower. Additionally, we removed cars with multiple entries from different model years, keeping only the most recent year's model. This was done to address potential temporal effects of having the same car model repeated in our sampling. Ultimately, these efforts reduced our observations from 398 to 302. Given our data has over 100 points, we feel it appropriate to use the large sample assumptions for OLS as opposed to the classic linear model assumptions.

We then performed all exploration and model building on a 33% sample of the data. The remaining 67% was used to generate the statistics in this report.

[1] Koeze, Ella, and Clifford Krauss. "Why Gas Prices Are so High." The New York Times, The New York Times, 14 June 2022, https://www.nytimes.com/interactive/2022/06/14/business/gas-prices.html

[2] Blanco. "Electric Cars' Turning Point May Be Happening as U.S. Sales Numbers Start Climb." Car and Driver, 14 May 2022, https://www.caranddriver.com/news/a39998609/ev-sales-turning-point/

[3] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. (https://archive.ics.uci.edu/ml/datasets/Auto+MPG)

[4] Donoho, David and Ramos, Ernesto (1982), "PRIMDATA: Data Sets for Use With PRIM-H" http://lib.stat.cmu.edu/datasets/cars.desc

We focus on displacement as our variable of influence because it is a physical attribute of cars that manufactures could directly alter to change fuel efficiency, in the hopes of driving consumer demand. Initial exploratory plots for displacement suggest a negative relationship between displacement and MPG. This leads us to build regressions in the general form

$$\widehat{MPG} = \beta_0 - \beta_1 \cdot displacement - \mathbf{Z}\gamma$$

where β_1 represents the decrease in MPG per unit of displacement, $\mathbf{Z}$ is a row vector of additional covariates, and γ is a column vector of coefficients.

Exploratory plots show similar negative relationship between cylinders and MPG, however as displacement is typically a calculation that includes number of cylinders as a variable, we exclude cylinders from models assuming its collinearity with displacement will be problematically high. Thus we can use displacement as an encompassing variable for number of cylinders.

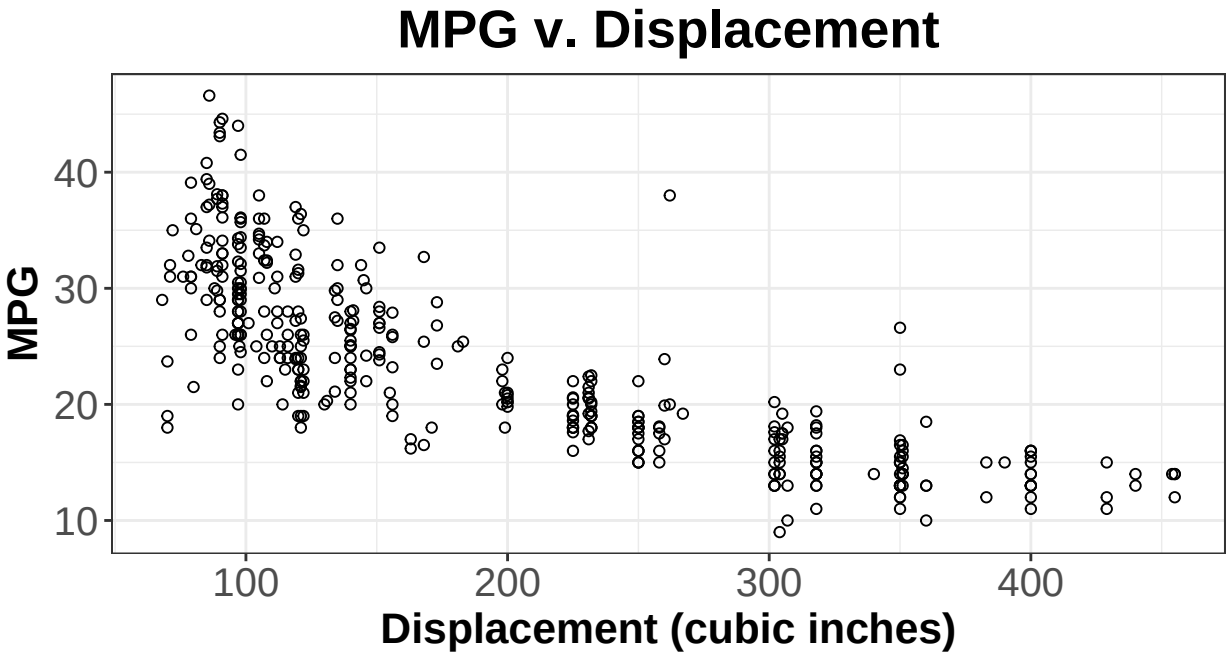

Figure 1: MPG vs. Displacement

We note that the scatterplot of MPG and displacement shows a slight curve, akin to a wide upward parabola, in the negative relationship, which we address by incorporating a squared term for displacement in some of our models.

Results

The Table 1 below shows the results of four regression models. Across all models, the key coefficients for *displacement* were highly statistically significant. Point estimates ranged from -0.14 to -0.06. To put this into context, this estimates that for a vehicle with a displacement equal to the average for the dataset (195.87 cubic inches), a 10% decrease in displacement would increase fuel efficiency by 1.19 to 2.72 MPG or 5% to 11% of the mean MPG (23.72) holding everything else constant. From our perspective, this is a practically significant effect, as such an increase in fuel efficiency would result in significant cost savings on gas over the lifetime of car ownership.

Table 1: Summary Statistics of Models

	Dependent variable:			
	Miles Per Gallon (MPG)			
	(1)	(2)	(3)	(4)
Displacement (cubic inches)	-0.06^{***}	-0.14^{***}	-0.07^{***}	-0.10^{***}
	(0.003)	(0.02)	(0.02)	(0.02)
Displacement (cubic inches) squared		0.0002^{***}	0.0001^{***}	0.0002^{***}
		(0.0000)	(0.0000)	(0.0000)
Horsepower			-0.01^{***}	-0.01^{***}
			(0.001)	(0.001)
Weight				0.23
				(0.14)
Acceleration				0.81^{***}
				(0.07)
Model Year	35.65^{***}	42.55^{***}	49.28^{***}	-15.77^{**}
	(0.85)	(1.93)	(2.03)	(5.44)
Observations	202	202	202	202
R^2	0.64	0.68	0.72	0.83
Residual Std. Error	4.73 (df = 200)	4.50 (df = 199)	4.21 (df = 198)	3.26 (df = 196)
Note:	*p<0.05; **p<0.01; ***p<0.001			
	HC_1 robust standard errors in parentheses.			

We note that in Model 2, adding in the squared version of displacement (*displacement_sq*) increases both the magnitude of the *displacement* coefficient and the explanatory power of the model while also being highly statistically significant in its own right. In Model 3, adding *weight* gives the model a similar increase in explanatory power as well as another highly statistically significant variable. Lastly, in Model 4, we see our only statistically insignificant coefficient (*acceleration*) as well as our strongest effect (*model_year*) and our biggest increase in explanatory power. Finding a statistically significant positive coefficient for *model_year* supports the theory that cars have tended to become more efficient over time via technological advancement.

Limitations

Consistent regression estimates require that the data used are independent and identically distributed (IID). Car manufacturers that have multiple models in the dataset may introduce clustering effects, due to similarity in materials or production methods, which could influence our outcome variable (MPG).

The other assumption we need to demonstrate is that the best linear predictor (BLP) is unique and we have met this by verifying that our variables have finite variances: *MPG* (62.6), *displacement* (10893.7), *weight* (735363.6), and *acceleration* (7.0). We believe these variables do not have infinite variance as they are the result of physical processes and car manufactures are unlikely to manufacture models of car that are well beyond the norm. Cars can't have infinitely large displacements or MPG and acceleration is bounded at zero.

Additionally, while we attempted to account for temporal effects by eliminating repeat model entries from different years, we must note that newer cars appear to have a distinct advantage in fuel efficiency. As fuel efficiency is likely to improve over time across all vehicle manufacturing, we suggest future study which uses methods to account for this effect, such as a survey of cars at a single point in time.

We must also note omitted variables and their potential biases. One example is the number of gears in the

engine transmission - the basic system that turns engine revolutions into tire rotations, where gears can multiply the work the engine does to make it more efficient. This likely has a positive relationship with fuel efficiency, and a negative one with displacement, leading our model to be underpredictive of fuel efficiency. Another example is a car's aerodynamic rating. Aerodyndamics has a positive relationship with fuel efficiency, but an unknown relationship with displacement. We reason that cars with both low displacement (a small sedan) and high displacement (sports cars) might be designed to minimize air resistance, and therefore omitting it could make our models under or overpredictive of fuel efficiency.

Additionally, we lack information on a number of mechanisms such as fuel injection, engine type, type of gas used, cooling system, method of governing and valve arrangement, all of which likely contribute to a vehicle's fuel efficiency.

While exploring our data, we observed what appeared to be two subgroups of cars when looking at displacement and horsepower. We hypothesize that car manufacturers may specifically design cars for two groups of consumers - those more concerned with engine efficiency (i.e. sedan drivers) versus those more concerned with engine power (i.e. truck/SUV drivers). This is another limitation to our model as we do not have a way to control for these vehicle types.

To evaluate multicolinearity within our models, we conducted VIF tests and found high (>4) VIF values for *displacement*, *displacement_sq*, and *weight*. We've attempted to mitigate multicolinearity by removing cylinders and horsepower from the models, but this is still a deficiency. We also introduced some multicolineary by adding in a squared version of *displacement*, but we felt this was worthwhile given we observed a non-linear relationship with MPG, but we did consider using a logged version as well.

Conclusion

This study estimated the impact of a vehicle's engine displacement on its fuel efficiency (MPG). For every cubic inch reduction of displacement to a vehicle's engine, our models predict a 0.06 to 0.14 increase in MPG. Future research to refine these models could gather data on vehicle characteristics such as gear ratios/transmission types, aerodynamic ratings, and fuel and engine types. The aim of this work is to help car manufacturers determine which vehicle characteristics can be modified to best optimize a vehicle's MPG, given the importance consumer's place on fuel efficiency as they make purchasing decisions.