**Mike’s Notes:**

* No reverse causality
* Product focus on altering engine displacement

1. **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]](#footnote-0). This has impacted the budgets of everyone across the country and it is now more important than ever to have a fuel efficient vehicle.

To this end, we model factors that contribute to a car’s miles per gallon (MPG) as a measure of fuel efficiency. 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). In this study we show that \_\_ is the best predictor for a car's MPG.

1. **A Brief Description of the Data**

* 398 data points, each with 8 features

This dataset was taken from the Statlib library, maintained at Carnegie Mellon University, used in the 1983 American Statistical Association (ASA) Exposition[[2]](#footnote-1), and was collected in 1982. Documentation for these data is limited and consequently we’ve relied on a letter from the ASA[[3]](#footnote-2). This donated dataset from July 1993 includes 398 observations which represent models of cars from the years 1970 - 1982. Car models can appear in the data multiple times and we believe these data are observational (can’t confirm given lack of documentation).

**3: A Discussion of How Key Concepts are Operationalized**

Fuel efficiency = MPG

Engine size = displacement

Engine power = horsepower

**4: An Explanation of Key Modeling Decisions**

The model will be built to predict the changes that could be made to increase the mpg for a car model. MPG will be modeled by the five continuous variables and three discrete variables. The first model will look at the interaction between weight and horsepower. The second model will add acceleration. Third adding displacement and cylinders since they would be covariates as the more cylinders the car has the more displacement (volume of the gas chamber). We’ve restricted our analysis to the most recent year for each model to account for temporality, removing 92 observations, and our modeling decisions have been made using a training set of 100 observations (33% of the data). Six values were excluded as there were no values for horsepower.

**Dependent Variable**: MPG

**Independent Variables**: cylinders, displacement, weight, horsepower, acceleration, and origin

Next Step Ideas:

* Log horsepower, maybe IQR for the outliers
* Cylinders / mile
* Looking at the values from the VIFs, it looks like the least correlated is horsepower.
* We can drop displacement for sure, since the value from the VIF is 20
* Maybe to correct for this, we can combine cylinders and weight together and then use horsepower as well?
* Can we add anything for the assumptions:
  + Large Sample Assumptions:
    - I.I.ID: Could accept with caution, noting that each model year could have an impact on the metrics of the car.
    - Unique BLP Exists: Finite variance could exist for each of the attributes, for example you can’t have an infinite number of displacement or horsepower.
* When evaluating the values from the VIF test, it was clear that the majority of the attributes had a significant level of multicollinearity, with VIF values well over 4. To correct this, the weight and cylinders attributes were combined by dividing weight by cylinders to create a new variable, ‘weight\_cylinder’.

**5: A table or visualization**

**6: A well formatted regression table**

**7: A Discussion of Results**

Zach:

* In the first model, the displacement is statistically significant, with weight being very statistically significant in the 2nd and third models. Maybe we can use it as a parameter for the second model.
* The R squared for the third model is the highest. But because it has the most features, it also could have the most multicollinearity.
* Reverse Causality: There isn’t any with mpg and any of the attributes

Outcome of model:

* Average displacement

**8: A Discussion of Limitations**

* A model of a car influencing the metrics of the next year’s model. (Not satisfying the I.I.D assumption).
  + *Similar to the release of iPhones each year, the previous year’s model often influences the next year’s, in terms of metrics and build. Therefore, it can be concluded that the assumption of I.I.D is not met.*
* We were not able to use the variables such as ‘model year’, ‘origin’, and ‘car name’, because they are not able to be modified in a way that will impact the miles per gallon target variable.
  + *As the project aims to discover modifiable attributes to optimize performance, there are several attributes that are unable to be used, such as model year, origin, and car name.*

- Omitted Variable bias for aerodynamics.

- *As aerodynamics could be both negatively and positively associated with miles per gallon, we would need more data to understand the true omitted variable bias. This can be represented by the following scenario; a car such as a prius is both aerodynamic, and has a high miles per gallon, but a car such as a mustang is aerodynamic but has a low miles per gallon.*

**8a: Statistical Limitations of your model**

* We don’t have much data, which may have limited the performance of the model when splitting into train and test sets.
* Omitted variable bias and it’s direction
  + Gearing/transmission
    - The basic system that turns engine revolutions into tire rotations is transmission. Series of gears can multiply the work the engine does to make it more efficient.
    - **Direction**: Number of gears likely has a positive relationship (π) with MPG (more gears = more efficient) and a negative relationship (γ) with our main variable displacement (more gears = less displacement needed). And our models show a negative relationship between displacement and MPG. Therefore the omitted variable bias for greeting is negative (π \* γ), pushing the already negative relationship between displacement and MPG away from zero and making our model underpredictive of MPG.
  + Aerodynamics
    - Vehicles with more vertical surface area at the front have to work harder to move through the air, making them less efficient. Automakers design car exteriors to minimize resistance/drag to increase fuel efficiency.
    - **Direction**: Aerodynamics likely has a positive relationship (π) with MPG (more aerodynamic = higher MPG) but the relationship (γ) it has with our main variable displacement could go in either direction. More data would help us determine if omitting this variable from our models results in over or underprediction of MPG.
  + Engine type:
    - Different mechanisms of action, such as V
  + Type of gas used
    - Vehicles optimzied for types of gasoline and will perform best under these conditions.
  + Summary:
    - The specific variables pertaining a car’s engine mechanisms are not taken into consideration in this dataset, a non-exhuastive list includes the fuel injection, engine type, cooling system, method of governing and valve arrangement. These factors contribute to the overall performance of the car, but may be captured by the
* The majority of the variables had some kind of multicollinerarity

**9: A conclusion:**

* *According to the results of the models shown in table 1, we recommend altering the displacement of the car. Although there are features with a greater impact, and are more statistically significant, such as acceleration and model year, these are not modifiable parameters.*

Using the results from model 4, we can infer that a one unit increase in a car’s displacement is associated with a 0.09 reduction in MPG. Alternatively, a 10% decrease in the displacement for the average car 19.4 (194 CIs) is associated with a 1.8 MPG reduction.

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.

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>. [↑](#footnote-ref-0)
2. 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>) [↑](#footnote-ref-1)
3. Donoho, David and Ramos, Ernesto (1982), ``PRIMDATA: Data Sets for Use With PRIM-H'' <http://lib.stat.cmu.edu/datasets/cars.desc> [↑](#footnote-ref-2)