A Model-Based Statistical Investigation of Factors Influencing Vehicle Fuel Efficiency

Group members: Feng Gu (T00751197), Jiahui Yan (T00752485),
Emmanuella Onyejiaka (T00748938), Yuzhuo Ye (T00751492),
Taiwo Ogunkeye (T00751495)

1 Introduction and Research Question

Fuel efficiency remains one of the most important performance indicators in the automotive industry, influencing both environmental impact and consumer cost. Previous studies [1, 2] have shown that some mechanical factors, like engine size and horsepower can affect fuel consumption.

With the right combination of these factors, improvement on fuel efficiency has been proven achievable, which promises the potential for developing more fuel-efficient vehicles by exploring optimal configurations on these factors and their interactions. Motivated by this background, this project aims to examine how a series of car factors (controllable factors) affect the fuel efficiency (target variable).

2 Research Objectives

This project aims to systematically investigate the factors influencing automotive fuel efficiency through statistically controls for major confounding factors. The primary objective is to examine how key automotive characteristics individually and collectively affect miles per gallon (MPG) performance. Specifically, we seek to:

1. Determine the main effects of each factor on fuel efficiency
2. Identify potential interaction effects among these factors to understand how combinations of characteristics influence fuel performance
3. Apply appropriate statistical analysis including assumption testing and effect size estimation

3 Data Description

The data are from the Auto MPG Dataset [1] from UCI Machine Learning Repository [4], it contains 398 automobiles manufactured between 1970-1982.

The dataset includes nine variables: the dependent variable **mpg** and eight predictors including cylinders, displacement, horsepower, weight, acceleration, model year, origin,

[1] Data source: `https://www.kaggle.com/datasets/uciml/autompg-dataset`

and car name, [2] the variable meanings are provided in Table 1 and their distributions are shown in Figure 2 in the Appendix.

Table 1: Summary of Variables in the Auto MPG Dataset

Variable	**Description**	**Units / Type**
mpg*(target)	Miles per gallon (fuel efficiency)	miles/gallon
cylinders	Number of engine cylinders	$\mathbb{N}$
displacement	Engine displacement	in^3(cubic inches)
horsepower	Engine power output	hp
weight	Vehicle weight	lb
acceleration	Time to accelerate 0–60 mph	s
model year	Year of manufacture (two-digit)	$\mathbb{N}$
origin	Country/region of origin	$\{1, 2, 3\}$
car name	Car make and model	string

The variable **car name** is a string and does not support distribution plots. We split its strings and extract the major car brands for all 398 samples, and the car brand distribution is shown in Figure 3 in the Appendix.

4 Literature Review

Contemporary automotive research employs statistical experimental design methodologies to investigate fuel efficiency factors. Jankovic and Magner [6] used factorial design approaches including full-factorial and central composite designs to optimize automotive engine fuel economy. Najafi et al. [7] applied response surface methodology to investigate multiple engine parameters including f uel blends and operating conditions to optimize both performance and emissions simultaneously. Win et al. [8] implemented Taguchi methods to evaluate diesel engine operating parameters for minimizing noise, emissions, and fuel consumption.

Leveraging the findings and methods of previous studies in developing fuel-efficient vehicles, Ahmad et al. [1] found 10% weight reduction improves fuel economy by 8%. Greene and Welch [2] showed higher cylinder counts reduce MPG due to increased energy requirements. Li and Zhao [3] observed newer models achieve better efficiency through improved engine management.

Compared to one-way effects, fewer studies examine comprehensive interaction effects between multiple mechanical, temporal, and geographical factors. This project addresses this gap using factorial design and multiple regression analysis to quantify both individual and combined influences on fuel efficiency.

5 Overview of Experimental Design

5.1 Experimental Units and Design Considerations

Experimental Units: This study employs an observational design using existing automotive data rather than a controlled experiment. Each of the 398 automobiles in the

[2] Here the car name is a long string with brand and model information.

Auto MPG dataset represents an independent observational unit.

Design Considerations: While randomization is not applicable to this observational study, the dataset's comprehensive coverage across multiple manufacturers, model years, and geographic regions provides natural variation that supports robust statistical inference. Potential confounding variables including transmission type, engine displacement, and manufacturing differences will be acknowledged as limitations.

5.2 Hypotheses and Statistical Models

The following hypotheses will be tested to examine the relationships between automotive characteristics and fuel efficiency at $\alpha = 0.05$ significance level:

5.2.1 ANOVA Model for Categorical Factors:

$$\text{MPG}_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \tag{1}$$

where i represents cylinder groups, j represents origin groups, and $(\alpha\beta)_{ij}$ represents interaction effects. The error term is tentatively assumed to follow $\epsilon_{ijk} \sim N(0, \sigma^2)$ for observations within each group.

Table 2: Main Effect Hypotheses for Automotive Fuel Efficiency Analysis

Variable	**Null Hypothesis (H_0)**	**Alternative Hypothesis (H_1)**
Cylinders	Number of cylinders has no significant effect on MPG	Number of cylinders has a significant effect on MPG
Origin	Country/region of origin has no significant effect on MPG	Country/region of origin has a significant effect on MPG
Cylinders × Origin	There is no significant interaction effect between cylinders and origin on MPG	There is a significant interaction effect between cylinders and origin on MPG

5.2.2 Multiple Linear Regression Model for Continuous Factors:

$$\begin{aligned}\text{MPG} = \beta_0 &+ \beta_1(\text{Cylinders}) + \beta_2(\text{Displacement}) + \beta_3(\text{Horsepower}) \\ &+ \beta_4(\text{Weight}) + \beta_5(\text{Acceleration}) + \beta_6(\text{Model Year}) \\ &+ \beta_7(\text{Origin}_2) + \beta_8(\text{Origin}_3) + \epsilon \end{aligned} \tag{2}$$

where $\epsilon \sim N(0, \sigma^2)$ and Origin is dummy-coded with USA as reference category.

6 Preliminary Analysis: One-Way ANOVA on Car Origin and Fuel Efficiency

As a preliminary investigation, we conducted a one-way ANOVA to examine whether car origin (USA, Europe, Japan) significantly affects fuel efficiency, and this analysis serves as a foundational step to more complex multi-factor designs.

Table 3: Specific Directional Hypotheses

Relationship	**Null Hypothesis**(H_0)	**Alternative Hypothesis**(H_1)
Cylinders-MPG	$\beta_{\text{cylinders}} = 0$	$\beta_{\text{cylinders}} < 0$ (More cylinders decrease MPG)
Weight-MPG	$\beta_{\text{weight}} = 0$	$\beta_{\text{weight}} < 0$ (Higher weight decreases MPG)
Horsepower-MPG	$\beta_{\text{horsepower}} = 0$	$\beta_{\text{horsepower}} < 0$ (Higher horsepower decreases MPG)
Model Year-MPG	$\beta_{\text{year}} = 0$	$\beta_{\text{year}} > 0$ (Newer cars have higher MPG)
Origin Effect	$\mu_{\text{USA}} = \mu_{\text{Europe}} = \mu_{\text{Japan}}$	At least one origin mean differs significantly

Experimental Design: The one-way ANOVA employed a completely randomized design with car origin as the single factor at three levels: Origin 1 (USA, n=249), Origin 2 (Europe, n=70), and Origin 3 (Japan, n=79). Each car represents an independent experimental unit, with MPG as the continuous response variable.

Hypotheses Tested:

- H_0: $\mu_{\text{USA}} = \mu_{\text{Europe}} = \mu_{\text{Japan}}$ (all origin groups have equal mean MPG)
- H_1: At least one origin group has a significantly different mean MPG

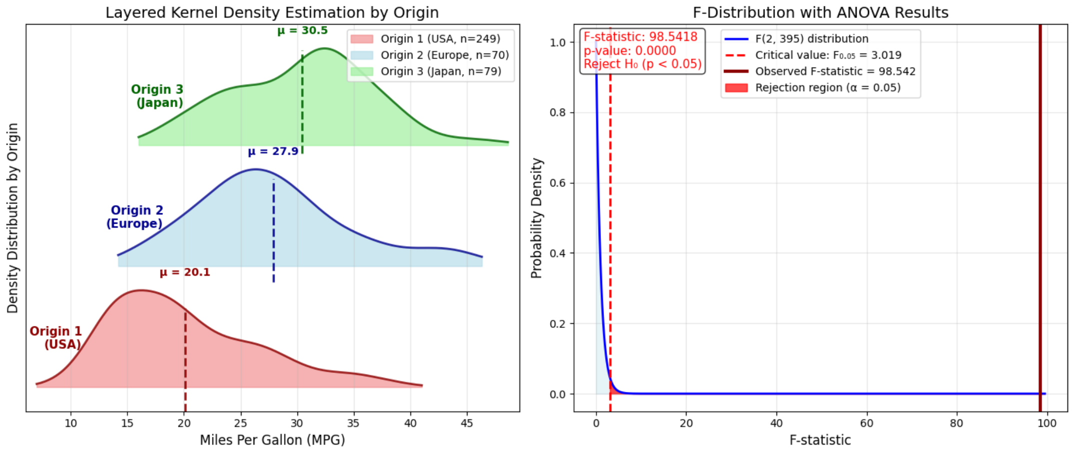

Figure 1: Preliminary ANOVA Results: Effect of Car Origin on Fuel Efficiency (MPG)

The ANOVA revealed a highly significant effect of car origin on fuel efficiency ($F_{(2,395)} = 98.54$, $p < 0.001$, $\eta^2 = 0.333$). Japanese cars showed highest efficiency (M = 30.45), followed by European (M = 27.89) and American cars (M = 20.08), leading to rejection of the null hypothesis.

This preliminary analysis confirms significant regional differences in automotive fuel efficiency, supporting the inclusion of origin as a key factor in subsequent multi-factorial designs.

Table 4: One-Way ANOVA Results: Effect of Car Origin on Fuel Efficiency (MPG)

ANOVA Summary	**Value**
Null Hypothesis (H_0)	All group mpg means are equal
Alternative Hypothesis (H_1)	At least one group mpg mean differs
Significance level (α)	0.050
Total sample size (N)	398
Number of groups (k)	3
Degrees of freedom (Between groups)	2
Degrees of freedom (Within groups)	395
Critical F-value	3.0186
Observed F-statistic	98.5418
p-value	< 0.001
Statistical Decision	**Reject** H_0
Effect Size (η^2)	0.3329

7 References

References

[1] Ahmad, N., et al. (2020). *Weight reduction and fuel efficiency in automotive design: An integrated assessment*. Transportation Research Part D: Transport and Environment, 86, 102446.

[2] Greene, D. L., & Welch, T. (2017). *Impact of engine size and vehicle weight on fuel economy: Policy implications*. Energy Policy, 108, 273–282.

[3] Li, F., & Zhao, J. (2022). *Advances in automotive efficiency: The role of model year innovation*. Applied Energy, 325, 119823.

[4] U.S. Department of Energy. (2021). *Auto MPG Dataset*. UCI Machine Learning Repository. Retrieved from `https://www.kaggle.com/datasets/uciml/autompg-dataset`

[5] Montgomery, D. C. (2019). *Design and Analysis of Experiments* (9th ed.). John Wiley & Sons.

[6] Jankovic, M., & Magner, S. (2006). Fuel economy optimization in automotive engines. *2006 American Control Conference*. IEEE. pp. 3334-3340.

[7] Najafi, G., Ghobadian, B., Yusaf, T., Ardebili, S. M. S., & Mamat, R. (2015). Optimization of performance and exhaust emission parameters of a SI (spark ignition) engine with gasoline–ethanol blended fuels using response surface methodology. *Energy*, 90, 1815-1829.

[8] Win, Z., Gakkhar, R. P., Jain, S. C., & Bhattacharya, M. (2005). Investigation of diesel engine operating and injection system parameters for low noise, emissions, and fuel consumption using Taguchi methods. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 219(10), 1237-1251.

8 Appendix

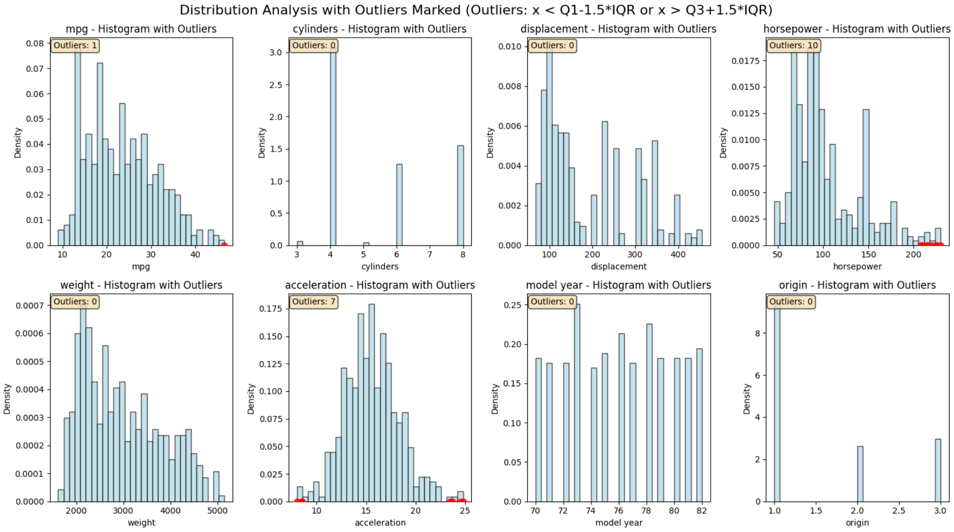

Figure 2: Distributions of Numerical Variables in the Auto MPG Dataset

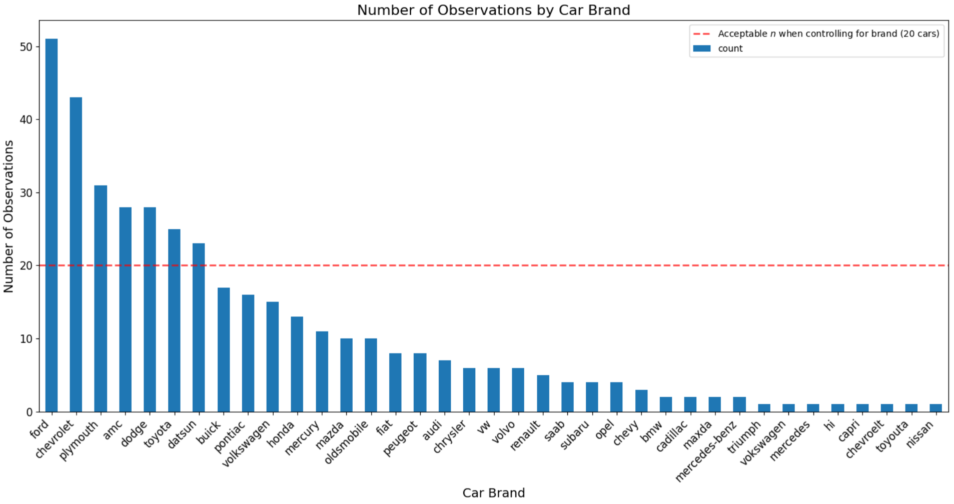

Figure 3: Distribution of samples by Car Brand

Table 5: Descriptive Statistics of Fuel Efficiency (MPG) by Car Origin

Origin	**Sample Size (n)**	**Mean MPG**	**Standard Deviation**
Origin 1 (USA)	249	20.08	6.40
Origin 2 (Europe)	70	27.89	6.72
Origin 3 (Japan)	79	30.45	6.09
Total	**398**	**23.51**	**7.82**

8.1 Ethical Considerations

Although the dataset is non-human and publicly available, ethical integrity is maintained by:

- **Data Source:** `https://www.kaggle.com/datasets/uciml/autompg-dataset`
- **Privacy and Compliance:** Ensuring no personally identifiable information is used.
- **Fair Representation:** Avoiding selective reporting or p-hacking.
- **Academic Integrity:** Properly citing all sources and acknowledging dataset creators.