Internal Combustion Engine Vehicles Fuel Efficiency Regression Analysis with Physical and Engineering Perspective

Bosen Yang

College of Letters and Science, University of California Los Angeles, Los Angeles, USA

bosen@ucla.edu

Abstract. Internal combustion engine vehicles (ICEVs) have undergone various transformations over the 20th and 21st centuries. A revolutionary in enhancing ICEVs' fuel efficiency happened particularly around the 1970s under the energy crisis and tightening emission policies. The paper conducted regression analysis on AUTO MPG dataset from UCI Machine Learning Repository to study the underlying relationships between fuel efficiency and various vehicle variables. This helps to make inferences about how adaptive engineering adjustments on those variables enhance fuel efficiency in the 1970s and 1980s under the impact of the energy crisis. The paper introduces physical and engineering connections among given predictor variables from the dataset and derives a general relationship between mpg and its predictors. This substantiates the statistical models' validity via a rigorous deduced expression. Then, three statistical modelling approaches, namely simple logged factor interaction multilinear regression (MLR), composite-variable MLR, and partial least square (PLS) MLR, were applied for comprehensive analysis and interpretation. Three models have complimentary advantages and drawbacks in interpretability, robustness, and generalizability, but collectively reflect how fuel efficiency evolves and correlated with other variables during 1970 to 1982. Finally, the paper commented on future prospect of ICEV market under the impact of new energy vehicles and explained how ICEVs' unique advantages made them still a preferable choice in the future.

Keywords: ICEV; fuel efficiency; energy crisis; emission policies.

1. Introduction

From the birth of the first practical modern automobile in 1885 in Germany to wide adoption of modern vehicles in the 21st century, the automobile industry has undergone numerous transformations. In 1913, Ford's assembly line introduced the idea of production cars and made automobiles affordable to the public [1]. After World War II, expansion of city paved roads and freeway networks further surges private car ownership. One of the most revolutionary transformations occurred in the late 20th century across the world with large-scale implementation of emission-reduction technologies in engine designs. This movement was driven not only by the oil crisis of the 1970s, which raised concerns about overreliance on gasoline and petroleum resources, but also by growing awareness of environmental issues caused by vehicle emissions, such as air pollution and acid rain [2]. In this context, to gain better understanding of how vehicle manufacturers specifically make improvements in engine designs, this paper will analyze the Auto MPG dataset from UCI Machine Learning repository that encompasses information of vehicle performance from 1970 to 1982. This dataset, as a classic benchmark dataset, reflects a transformation of the automotive industry amid pressures of the 1970s energy crisis and emissions regulations. Thus, studying this dataset provides understanding of the relationship between engine parameters and fuel efficiency of vehicles during the transformation era. The analysis encompasses both physical and mechanical considerations, such as effects of horsepower, displacement, weight, and acceleration on fuel consumption. These insights will inform construction and interpretation of statistical models, providing a rigorous framework for analyzing relationships among engine parameters.

2. Vehicle Variables and Designs

In the study of vehicle fuel efficiency, it is crucial to understand the underlying physical or mechanical principles of vehicle engines. While fuel consumption rate is mainly determined by driving behavior and environmental conditions in short-term variations of MPG, mechanical factors like engine displacement, number of cylinders, and vehicle weight impact the long-term average performance of each model. If driving behavior and environment conditions are treated as short-term random effects for this study, the analysis should therefore primarily focus on engine mechanical factors across different models, and this will help to construct regression models with AUTO MPG dataset [3].

2.1. Displacement Analysis

Displacement is the total volume of air and fuel mixture absorbed into the combustion engine in one complete cycle, alternatively, the total volume swept by gas-pistons in all cylinders [4]. By using cylinder volume formula, the total displacement can be expressed as

$$D = \pi r^2 \cdot h \cdot N_{cyl} = \left(\pi\left(\frac{B}{2}\right)^2 \cdot L_{str}\right) \cdot N_{cyl} = \frac{\pi}{4} \cdot L_{str} \cdot B^2 \cdot N_{cyl} \quad (1)$$

where $D =$ total displacement, $L_{str} =$ stroke length, $B =$ bore diameter, $N_{cyl} =$ number of cylinders. For total displacement determination of an engine, vehicle engineers often choose to fix B and L_{str} to target a desired individual cylinder displacement (ICD), then multiply N_{cyl} in the end to determine total displacement due to the discrete nature of N_{cyl}. The design approach also reflects the fact that N_{cyl} is constrained by multiple factors, such as engine balance, smoothness and packaging considerations. Then for a fixed ICD given as:

$$ICD = \frac{\pi}{4} \cdot L_{str} \cdot B^2 \quad (2)$$

Engineers select appropriate N_{cyl} (usually 3, 4, 6, 8) to scale ICD to desired total displacement to meet performance or fuel economy goals. (2) shows B has a quadratic effect and thus is more deterministic on ICD than L_{str}. Differences of the logged form yields

$$\Delta ln(ICD) = \Delta ln(L_{str}) + 2\Delta ln(B) \quad (3)$$

Which provides a quantitative measure of how design adjustments impact ICD. Logarithms naturally approximate small percentage changes such that:

$$ln(after) - ln(before) \approx \frac{after - before}{before} \quad (4)$$

Thus, percentage impact on ICD of B is twice as much as L_{str}. From a mechanical standpoint, engineers typically adjust B and L_{str} to balance combustion efficiency and thermal management. B sets valve sizing, which directly determines the rate of air mixture intake and exhaust. Wider B allows larger valve size and allows higher breathing rate under high RPM. On the other hand, narrower B restricts valve size but generates higher torque by taking advantage of ram-air effect under low RPM. In each complete cycle of a gas-piston, L_{str} decides distance and speed of the piston travels. Long L_{str} provides a larger swept volume given a fixed B, which increases torque due to greater leverage on the crankshaft under low RPM. However, high piston speed could also result in higher frictional losses and thermal stress. Conversely, Short L_{str} reduces piston speed, enabling safer spinning conditions under high RPM and cutting down frictions, but gives less torque per cycle. Therefore, the interplay of B and L_{str} does not solely determine ICD, but also affects engine's efficiency, thermal factors, and mechanical durability. To further evaluate how D and N_{cyl} affect real-world performance, researchers applied simulated models to study cylinder deactivation (CDA) strategies [5]. The study constructed simulated models for gasoline engines of 3, 4, 6, and 8 cylinders

and 1.0 to 5.0L displacement. By comparing fixed-type and variable-type CDA conditions, researchers were able to make quantitative improvements on fuel efficiency while maintaining net indicated mean effective pressure (NMEP). The experiment shows fuel efficiency improves from 2.2% to 10.0% fixed type and 2.2% to 12.9% for variable type. Among those, engines with large displacements and more cylinders performed better fuel efficiency after conducting CDA.

2.2. Horsepower Analysis

Horsepower (HP) is strongly associated with total displacement D since piston swept volume decides amount of air-fuel mixture per cycle that powers the vehicle. Start with the power formula of the rotating system

$$P = \tau \cdot \omega \tag{5}$$

HP, as a unit of power, is directly proportional to power:

$$HP \propto \tau \cdot \omega,$$
$$\tau = \frac{W}{\theta} = \frac{W_{cycle}}{4\pi} = \frac{\int P\, dV}{4\pi} = \frac{P_{avg} \cdot D}{4\pi} \tag{6}$$

And

$$HP \propto \tau \cdot \omega \Rightarrow HP \propto P_{avg} \cdot D \cdot \omega \Rightarrow HP = k \cdot P_{avg} \cdot D \cdot \omega \tag{7}$$

Where $k =$ positive constant, $P_{avg} =$ mean effective pressure, $\tau =$ torque and $\omega =$ rotational velocity. The proportionality relationship of HP and P_{avg}, D, ω indicates HP increases linearly with those three factors. Furthermore, none of those are deterministic to HP, as the output depends on their linearly combined contribution. For instance, a large D may yield, modest HP if P_{avg} is low and ω is low, whereas small D may still achieve high HP when P_{avg} and ω are high. A recent study shows how precise control and cylinder pressure can raise P_{avg}, illustrating how higher piston pressure contributes significantly to increased HP given D and ω to being fixed. Hence, HP is a joint outcome of pressure, volume, and speed rather than being attributed to one of the dominant factors [6].

2.3. Weight and Acceleration

Give Newton's second law: $F = ma$ and air/rolling friction formulas as

$$F_{aero} = \frac{1}{2}\rho C_d A v^2$$
$$F_{roll} = mgf_r \tag{8}$$

Where $f_r \approx 0.01 \sim 0.02$ for passenger vehicles, engine power required to maintain a constant cruising velocity v (counteracting friction) and to accelerate with instantaneous velocity v are given by:

$$P_{cruise} = (F_{aero} + F_{roll})v = \left(\frac{1}{2}\rho C_d A v^2 + mgf_r\right)v$$
$$P_{accel} = F_{accel}v = mav \tag{9}$$

Respectively, given other factors to be fixed, these two expressions directly reflect a positive correlation between vehicle weight and required power to cruise/accelerate [7]. Vehicle weight is also positively associated with horsepower needed to achieve same cruising speed and acceleration. In other words, heavier cars require more power to achieve the same acceleration. Statistically, this helps explain why cars with high performance tend to have lower mpg since more fuel is burned to provide the necessary acceleration energy. However, vehicle weight is not the only factor affecting power. Since f_r is weakly dependent on v (near constant low speed and negligible under high speed), F_{roll} is primarily dependent on vehicle weight and road conditions. On the other hand, F_{aero} is

associated further with vehicle frontal area A and drag coefficient C_d depending on the model. In the 1970s and 1980s, passenger vehicles (mostly box shaped) typically exhibited higher drag coefficients due to less refined wind tunnel testing, and lower emphasis on fuel efficiency compared to modern vehicles [8]. Poor aerodynamic optimization compounds the resistive load on the engine, resulting in less fuel efficiency. Generally, F_{aero} contributes more to energy demand under high cruise speed since it grows quadratically with speed.

2.4. General Form

To fully understand the relationship between fuel efficiency indicator mpg and its predictors, making connections of above formulas and ideas to the response variable mpg is necessary. First, the general expressions of fuel consumption per distance (gpm) and fuel efficiency (mpg) are given by:

$$gpm = \frac{fuel\ consumed}{distance\ travelled} = \frac{V_{fuel\ flow}}{v}$$
$$mpg = \frac{distance\ travelled}{fuel\ comsumed} \tag{10}$$

So, mpg is inversely proportional to gpm. Also know that rate of fuel flow is proportional to power needed at the wheels, then:

$$mpg = \frac{1}{gpm} = \frac{v}{V_{fuel\ flow}} \propto \frac{v}{P} = \frac{v}{P_{cruise} + P_{accel}} \tag{11}$$

Plug in (9) to get:

$$mpg \propto \frac{v}{\left(\frac{1}{2}\rho C_d A v^2 + mgf_r\right)v + mav} = \frac{1}{\frac{1}{2}\rho C_d A v^2 + mgf_r + ma} = \frac{1}{m(gf_r + a) + \frac{1}{2}\rho C_d A v^2} \tag{12}$$

Since $A \sim D$ and f_r has a week positive dependence on speed, finally, mpg follows an inverse relationship with its predictors such that:

$$mpg \sim \frac{1}{m(a + v) + Dv^2} \tag{13}$$

Also, given that D is positively correlated with N_{cyl} and HP by (1) and (7), then mpg I negatively correlated with five main predictors: vehicle mass, acceleration, displacement, horsepower, and cylinders. This provides a general framework for statistical modelling that not only helps to validate these overall relationships but also offers insight into some underlying philosophies of vehicle designs in the 1970s and 1980s. However, engineers cannot simply minimize these variables to solely maximize mpg because fuel economy is a trade-off problem balancing efficiency against performance and other factors. Also, other engineering considerations such as emissions standards, durability, and comfort also constrain those predictors maximizing mpg. While a statistical framework helps to highlight the negative correlations between mpg and its major predictors, vehicle design requires balancing all considerations to achieve an optimal compromise aligning with market demands.

2.5. Engine Structures

Along with 7 quantified variables from AUTO MPG dataset, car_name comprises additional information determining implicitly. Cylinder arrangement of the internal combustion engine is an underlying factor that varies across different models. The dataset records models mostly with 3,4, and 6 cylindered inline engines (I3, I4, and I6), and V-type 8 cylindered engines commonly used in American muscle cars, which need high instantaneous outburst power to accelerate and move their heavy weights. In the 1970s and 1980s, Japanese and European automakers favored I4 and I6 engines that prioritized space efficiency and fuel economy. While inline engines are generally simpler and

operate smoother due to reduced frictional loss even firing orders, they usually deliver less power and torque compared to V8 engines. Typically, V8 engines fit better in small engine compartments than I8 engines because V8 engine's angled arrangement shortens its length and provides more compactness [9]. The distribution of cylinder arrangement reflects regional manufacturing practice and trade-offs between performance and efficiency that underlies the variation of mpg. Moreover, Internal combustion engines are also broadly categorized by their air intake method: self-aspiration and turbocharge. Self-aspiration engines have simpler structures and solely rely on atmospheric pressure to draw air into cylinders, yielding smooth throttle response. In contrast, turbocharged engines use turbines to compress air, enabling higher combustion efficiency under high oxygen levels, which yields greater power [10]. In the 1970s and 1980s, self-aspiration type dominates popular models while turbocharging still stayed in a niche position. In general, these designs reflect varying market demands and trade-offs between performance and efficiency.

3. Regression Modelling

3.1. Full MLR model

To further investigate how the vehicle variables influence fuel efficiency (mpg) in the AUTO MPG dataset, a multiple linear regression (MLR) full model is fitted with predictors cylinders, displacement, horsepower, weight, acceleration, model_year, and origin in RStudio. car_name is temporarily held for consideration since it serves as a cardinality label rather than a meaningful predictor. The results indicate that displacement, weight, model_year, and origin are the only significant predictors of mpg with significant level $\alpha = 0.01$. From the model summary, origin and model year both show a strong positive association with mpg, with slopes $\beta_{origin} = 1.426$ and $\beta_{year} = 0.751$ respectively, whereas weight is negatively associated with mpg with slopes $\beta_{weight} = -0.006$. Nevertheless, the positive slope for displacement appears to be counterintuitive when compared to the real-world expectations. Correlation analysis reveals that displacement is highly correlated with weight, model_year, and origin with VIF function, indicating multicollinearity among those predictors. In addition, given the theoretical importance of other engine variables, partial models that exclude predictors such as cylinders or horsepower are not considered appropriate. Instead, some alternative modeling strategies are performed below to better examine the underlying correlations among those variables.

3.2. Power Transformed MLR Model with Factor Interactions

To further examine the effects of predictors on model performance while assessing potential transformations to improve normality, a Box-Cox power transformation is applied to the set of variables. This approach stabilized skewed continuous variables through log transformation while preserving interpretability. However, the analysis produces inconsistent power variables λ_i for each variable, indicating that no single transformation can simultaneously normalize all variables. The likelihood ratio test yields $p < 2.22\text{e}-16$ for both hypotheses, resulting in rejecting both no transformation ($\lambda = 1$) and uniform log transformation ($\lambda = 0$). Thus, partial power transformation is considered given the distinct powers λ_i. Nevertheless, the resulting MLR model with power transforms reveals that important vehicle variables such as cylinders and displacement lose significance with $\alpha = 0.05$, suggesting that partial power transformation does not necessarily improve the model's validity.

The subsequent analysis restricted the Box-Cox procedure to continuous predictors only by excluding all discrete variables, including cylinders, model_year, and origin. Another important observation is that these three discrete variables are the only ones for which the estimated Box-Cox power variables are different from zero. Under this specification, the likelihood ratio test yields $p = 0.326$ for $\lambda = 0$, suggesting that log-transformations adequately stabilized variances and

approximate normality for these continuous variables. The partial log-transformed MLR model is given by:

$$log(mpg) = 9.348 - 0.162\,log(disp) - 0.516\,log(hp) - 0.291\,log(wt) - 0.269\,log(accel) \quad (14)$$

Some variables from the model are abbreviated for brevity: $cyl =$ cylinders, $disp =$ displacement, $hp =$ horsepower, $wt =$ weight, $accel =$ acceleration, $my =$ model_year with the intercept and all slopes being significant with $\alpha = 0.01$. Those slopes also align with theoretical expectations of negative associations between mpg, and these four predictors. These findings contrast with earlier models that inappropriately included discrete predictors leading to inconsistent λ estimates and significance loss of some predictors.

Since discrete variables cylinders, model_year, and origin are not included in model (14) but still heavily related to fuel efficiency, the preferred strategy is to retain the log transformed continuous predictors while treating the discrete variables as trend effects. In other words, these discrete variables are treated as factors of the regression model. The factor-augmented regression model is given as

$$log(mpg) = 8.372 - 0.085\,log(disp) - 0.290\,log(hp) - 0.421\,log(wt) - 0.203\,log(accel) + \{factors\} \quad (15)$$

Where the factor effects expression is represented as

$$\{factors\} = \gamma_{cyl} \cdot cyl + \delta_{my} \cdot my + \zeta_{origin} \cdot origin \quad (16)$$

With all slopes of predictors and indicator coefficients of factors being significant with $\alpha = 0.001$ from ANOVA results, and levels $\gamma_{cyl,4} = 0.300$, $\gamma_{cyl,5} = 0.327$, $\gamma_{cyl,6} = 0.257$, $\gamma_{cyl,8} = 0.248$, $\delta_{my,76} = 0.065$, $\delta_{my,77} = 0.132$, $\delta_{my,78} = 0.135$, $\delta_{my,79} = 0.226$, $\delta_{my,80} = 0.319$, $\delta_{my,81} = 0.243$, $\delta_{my,82} = 0.297$, $\zeta_{origin,3} = 0.040$ being significant with $\alpha = 0.05$. Alternatively, 6 interaction models are constructed considering all factors, all continuous predictors, 3 pairs of factor-predictor, and all interactions. Those models are compared with model (15) with ANOVA function, and the only significant model under $\alpha = 0.01\ or\ 0.05$ is given as

$$log(mpg) = 8.813 - 0.295\,log(hp) - 0.506\,log(wt) - 0.274\,log(accel) + \{factors^*\} + \{interaction\} \quad (17)$$

Where the factor effects expression is represented as

$$\{factors^*\} = \gamma^*_{cyl} \cdot cyl + \delta^*_{my} \cdot my + \zeta^*_{origin} \cdot origin \quad (18)$$

With levels $\gamma^*_{cyl,4} = 0.522$, $\gamma^*_{cyl,5} = 0.531$, $\gamma^*_{cyl,6} = 0.478$, $\gamma^*_{cyl,8} = 0.463$, $\delta^*_{my,75} = 0.067$, $\delta^*_{my,76} = 0.087$, $\delta^*_{my,77} = 0.155$, $\delta^*_{my,78} = 0.155$, $\delta^*_{my,79} = 0.241$, $\delta^*_{my,80} = 0.336$, $\delta^*_{my,81} = 0.262$, $\delta^*_{my,82} = 0.32$, $\zeta^*_{origin,2} = -1.61$, and interactions expression is specified as

$$\{interaction\} = -0.432\,log(disp) \cdot origin_3 + 0.294\,log(accel) \cdot origin_2 \quad (19)$$

With all slopes, indicator coefficients, and interaction coefficients being significant with $\alpha = 0.05$. The F-test on model (15) and model (17) yields a p-value less than 0.006. They both reflect a general increasing trend of mpg with respect to year, and a gentle decreasing trend of mpg with respect to cylinders, and proportional slopes of mpg over other continuous variables. Model (17) omitted the slope for displacement but indicated some interaction results capturing regional design philosophies and providing a more realistic representation of fuel efficiency determination than model (15). In expression (19), efficiency penalty of increasing displacement is stronger for Japanese vehicles compared to US vehicles ($\beta = -0.047$); and acceleration improvements contribute more positively to mpg for European vehicles relative to US vehicles ($\beta = 0.294$). These observations align with historical perspective that Japanese manufacturers emphasized more on strict efficiency trade-offs, whereas European designers favored balanced performance with lighter vehicles.

3.3. Composite Variable MLR Model

Even though model (4) captured most of the relationships between mpg and other vehicle variables, the multicollinearity issue among displacement, horsepower, weight, and acceleration is not yet

addressed in above modelling methods. To resolve this, composite variable MLR models are constructed, aiming to take account of relationships among predictors. First, consider two composite variables denoted as $ESI = Z[\log(disp)] + Z[\log(hp)] VMI = Z[\log(wt)] + Z\,[\log(accel)]$. Representing engine size index and vehicle mass index, respectively. All predictor variables are standardized with $Z[X] = \frac{X-\bar{X}}{\sigma}$.

To ensure comparability under the same distribution, then summed to form the indices. These indices incorporate physical and mechanical relationships among vehicle variables deduced in Section 2 while reducing multicollinearity in the MLR model. The factor-augmented composite-variable model is expressed as

$$\log(mpg) = \beta_0 + \beta_{ESI} \cdot ESI + \beta_{VMI} \cdot VMI + \gamma_{cyl}^{CV} \cdot cyl + \delta_{my}^{CV} \cdot my \tag{20}$$

And the most significant interaction model among 6 models through F-tests yields a p-value less than 0.0008 and is given as

$$\log(mpg) = 2.592 - 0.113ESI - 0.059VMI + \gamma_{cyl}^{CV*} \cdot cyl + \delta_{my}^{CV*} \cdot my + \{interaction_{CV}\} \tag{21}$$

Where:

$$\{interaction_{CV}\} = -0.010\ ESI \cdot VMI - 0.517\ cyl_4 \cdot my_{81} - 0.569\ cyl_6 \cdot my_{81} \tag{22}$$

With levels $\gamma_{cyl,4}^{CV*} = 0.438$, $\gamma_{cyl,5}^{CV*} = 0.627$, $\gamma_{cyl,6}^{CV*} = 0.430$, $\gamma_{cyl,8}^{CV*} = 0.333$, $\delta_{my,75}^{CV*} = 0.111$, $\delta_{my,77}^{CV*} = 0.376$, $\delta_{my,78}^{CV*} = 0.243$, $\delta_{my,79}^{CV*} = 0.299$, $\delta_{my,80}^{CV*} = 0.203$, $\delta_{my,81}^{CV*} = 0.692$, $\delta_{my,82}^{CV*} = 0.384$ under $\alpha = 0.05$. Model (21) highlights some important theoretical insights into fuel efficiency. Both ESI and VMI exert negative effects on log(mpg), indicating that heavier and high-performance vehicles usually have lower fuel efficiency. In addition, a negative coefficient for interaction between ESI and VMI reveals that vehicles with both high engine power and heavy mass tend to suffer less in lower mpg from the interaction than individual contributions from ESI or VMI. Cylinders factor effects reveal that 4 to 6 cylindered vehicles reached the optimal mpg, but the advantages of 4 and 6 cylindered vehicles are diminished in 1981 based on the interaction term. Model year term discloses that later models are generally associated with higher fuel efficiency and mpg peaks in 1981. Nonetheless, origin's effect is insignificant in this model. This suggests that regional differences do not largely impact fuel efficiency due to globally generalized basic design ideologies across the world in the 1970s and 1980s. Besides, generally higher mpg of Japanese vehicles and lower mpg of US vehicles are mostly due to engineering determinants such as displacement and cylinders, whereas attributing mpg differences to origin alone reflects stereotyping.

3.4. Modelling with PLS

Another way to address multicollinearity is to adopt partial least square (PLS) regression model. Let $Z[Y]$ and $Z[X_i]$ be the standardized logged response variables and continuous predictors such that $Y = \log(mpg)$ and $X = \{\log(disp), \log(hp), \log(wt), \log(accel)\}$. Then the first latent component that captures maximal joint variation between predictors and response is given by

$$t_1 = Z[X]w_1, \qquad w_1 = \frac{Z[X]^T Z[Y]}{||Z[X]^T Z[Y]||} \tag{23}$$

Where w_1 is the normalized weight vector pointing to the same direction of the maximal covariance between X and Y. Then, subsequent latent components $t_2, t_3, \ldots$ are extracted iteratively by the deflation process on the predictor and response matrices

$$Z[X]_{k+1} = Z[X]_k - t_k t_k^T Z[X], \qquad Z[Y]_{k+1} = Z[Y]_k - t_k t_k^T Z[Y] \tag{24}$$

After extracting components, the regression coefficients for standardized predictors are:

$$\hat{\beta} = W(V_X^{-1} W) V_Y \tag{25}$$

Where $W = \{w_1, \dots, w_n\}$, $V_X = \{Z[Y]t_1, \dots, Z[Y]t_n\}$, and $V_Y = \{Z[Y]t_1, \dots, Z[Y]t_n\}$. And the intercept $\hat{\beta}_0$ is obtained by:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}\bar{X} \tag{26}$$

Finally, the PLS regression model can be expressed as:

$$\log(mpg) = \hat{\beta}_0 + \hat{\beta}^T X_{latent} + \{factors_{PLS}\}, \qquad X_{latent} = t_1, \dots, t_n \tag{27}$$

Based on the validation plots from R in Fig. 1, augmented and interaction 3 model exhibits a faster drop in RMSEP (prediction error) when increasing the number of components.

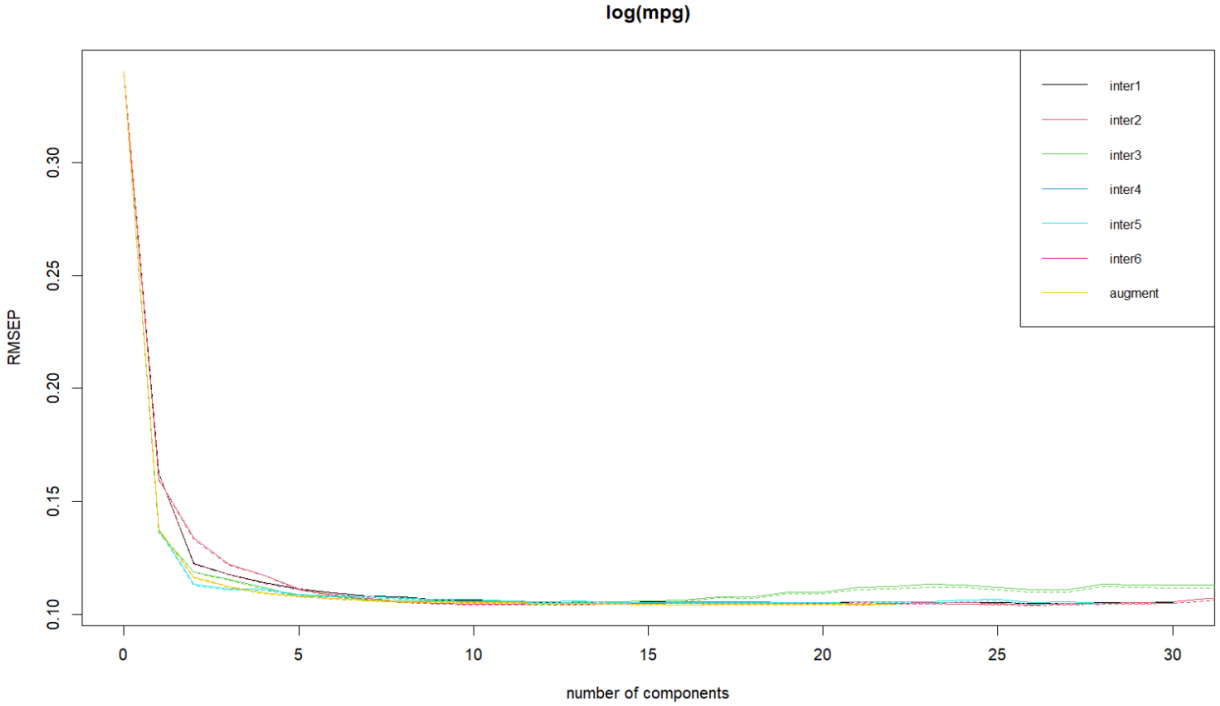

Fig. 1 PLS Models Validation Plot

For modeling simplicity, taking the basic augmented model benefits more in interpretability. The first latent score t_1 explained 83.9% of $Var[\log(mpg)]$ and $\{t_1, \dots, t_5\}$ explained 91%. The latent scores summary can be found in Table 1.

Table 1. Latent Score Summary

Name	Latent Score Expression
t_1	$-0.417\,Z[log(disp)] - 0.412\,Z[log(hp)] - 0.424\,Z[log(wt)] + 0.226\,Z[log(accel)] + \{factors_{t_1}\}$
t_2	$0.103\,Z[log(disp)] - 0.006\,Z[log(hp)] - 0.082\,Z[log(wt)] - 0.425\,Z[log(accel)] + \{factors_{t_2}\}$
t_3	$0.069\,Z[log(disp)] - 0.331\,Z[log(hp)] - 0.283\,Z[log(wt)] - 0.323\,Z[log(accel)] + \{factors_{t_3}\}$
t_4	$0.036\,Z[log(disp)] - 0.237\,Z[log(hp)] - 0.315\,Z[log(wt)] - 0.314\,Z[log(accel)] + \{factors_{t_4}\}$
t_5	$0.193\,Z[log(disp)] - 0.329\,Z[log(hp)] - 0.018\,Z[log(wt)] + 0.446\,Z[log(accel)] + \{factors_{t_5}\}$

The PLS regression model appears to be reasonable and aligns with the real-world philosophy of fuel efficiency. First, t_1 is dominated by negative weights for logged continuous predictors, being consistent with the physics and mechanical perspective and agreed with previous models. Since t_1

captures the largest proportion of the variance relevant for predicting the mpg, it effectively represents the major trade-off between engine size / power and efficiency. On the other hand, t_2 to t_5 explains smaller orthogonal variations in some uncommon models. Besides, cylinder effects reveal that 4-cylindered vehicles have positive scores while 8-cylindered vehicles have negative scores in t_1. t_1 also produces scores 0.126 and 0.213 for origin effects. Lastly t_1 and t_2 both displayed an increasing effect of model year on mpg, and the scores turn strictly positive in year 1977.

4. Summary and Analysis

4.1. Model Summary

This Section will compare three models acquired from Section 3. Despite their structural differences and disparate levels of interpretability. The simple logged factor interaction MLR model (17), composite variable logged factor interaction MLR model (20), and PLS logged factor MLR model (27) mapped onto the same physical truth that fuel efficiency is generally negatively correlated with the predictor variables' weight, acceleration, displacement, cylinders, and horsepower provided by (13). Besides, consistent ascending trends of mpg from 1975 to 1981 across three models reflect technological adjustment to internal combustion engine fuel efficiency under the pressure of energy crisis in the 1970s. Model (20) particularly highlights the peak in fuel efficiency in 1981, showing how 4 and 6 cylindered vehicles further diminishes fuel consumption under tightening standards. Besides, model (17) emphasized regional distinctions of how displacement and acceleration penalties on mpg exhibit. However, models (20) and (27) have origin being insignificant in the models because origin's effects on mpg are likely overridden by other predictor variables such as horsepower and cylinders. Based on Section 2.5, mpg distinction between US preference for V8 and European and Japanese reliance on I4 and I6 is largely captured by horsepower and cylinder variables and left little residual explanatory power for the origin factor.

Conversely, unlike how model (17) directly reflects a direct inverse proportional relationship between mpg and its predictor variables, model (20) further addressed multicollinearity issue by introducing composite indices ESI and VMI. These indices consider joint effects of four continuous predictors by assigning the same slopes to closely related variables, thereby improving model stability and reducing noise from overlapping effects. Its robustness ensures compatibility with broader design patterns and details, such as displaying an efficiency peak in 1981, which might be less evident in simpler models. However, model (20)'s interpretation becomes less direct since standardized variables and overlapping effects only capture overall efficiency trends rather than capturing nuance among different effects. Lastly, model (27) further sacrifices transparency as latent variables are abstract but offers stronger generalization and robust performance on hidden data while addressing multicollinearity meanwhile. The latent scores represent compressed efficiency dimensions and stabilize inference under multicollinearity. The PLS framework preserves critical information, such as the dominating negative effects of horsepower and weight on mpg. Meanwhile, it filters noise from the overlapping effects of variables, which provides robustness allowing the model to reveal long-term efficiency patterns.

4.2. Impact of NEVs to ICEVs Market

Looking forward, ICEVs may undergo another major transformation soon as NEVs get more popularized. Some principal challenges with NEVs today are their low charging efficiency and high electricity charges. However, as major breakthroughs in room temperature superconducting materials are achieved and removable better technology becomes more mature, those challenges are likely to be substantially mitigated. Enhance battery module swapping infrastructures complimentary to charging stations, will drastically reduce waiting time for recharge, which further benefits NEV markets, but the issue of standardization of battery modules across various manufacturers remains a challenge [11].

While NEVs markets' growing trend is certain, ICEVs still retain advantages that NEVs have yet to fully replicate. For example, consistent power output and enhanced driving feel make ICEVs typically perform better on long-distance journeys. In addition, NEVs faces an additional reliability challenge that electric drivetrains are generally more prone to malfunction compared with mechanical transmission [12]. Additionally, as shown in AUTO MPG dataset, optimizing mpg involves trade-offs with other variables such as acceleration and horsepower. Engineers from the 1970s and 1980s tried to navigate these compromises to optimize customers' experience. Still, some drivers are reluctant to completely abandon ICEVs for their ultimate driving experience and performance demands in rally racing [13]. Thus, fuel consumption does not appear to be the causation that eliminates the market of ICEVs, but rather a factor among many shaping consumer preferences [14, 15]. Therefore, NEVs are unlikely to fully take over the ICEV markets for various irreplaceable advantages of ICEVs. However, future ICEVs may achieve further breakthroughs in fuel efficiency with continuous process in engineering, material, and chemistry.

5. Conclusion

In conclusion, fuel efficiency is not limited by the physical relationship but reflects more of the history. Demonstrates the physical inevitability of efficiency loss with heavier weights and high performance. Yet, luxury and performance models manufacturers still choose to sacrifice fuel efficiency to cater to consumers who prioritize performance and safety rather than fuel efficiency. The regression models illustrate evolution of vehicle designs responding to policies under energy crisis and mitigating environmental degradations. While all three models agreed on the validity of relationship, they possess different levels of interpretability and robustness. The simple logged factor interaction MLR model explicitly illustrates the proportional correlation between mpg and its predictor variables but did not address the multicollinearity issue and had shown signs of underfitting. The composite variable logged factor interaction MLR model illustrates the joint effects of composite variables ESI and VMI on mpg while resolving multicollinearity. Finally, PLS logged factor MLR model compresses correlated predictors into latent scores, further strengthening robustness to weirdly behaved data, but sacrificing transparency. While each model carries distinct advantages, they collectively provide a comprehensive image of how fuel efficiency evolves and correlates with other variables from 1970 to 1982. Across all models, a consistent and significant improvement in fuel efficiency is reflected after 1975, coinciding with tightened policies and engineering adaptation following the energy crisis. However, some additional design variables that are not captured in the dataset might subtly affect mpg, then the regression provides only a rough inferential image of efficiency background and trade-offs. Lastly, the broader market shows that fuel economy is merely one determinant among many. While NEV market has rapidly expanded since the 21st century, ICEVs retain their unique advantages from multiple perspectives. ICEVs and hybrid vehicles' irreparability under various scenarios will still maintain its popularity in the future. Overall, consumer preferences, technological advances, and trade-offs between performance and sustainability will shape the coexistence of all vehicle types

References

[1] Hughes C. The assembly line at Ford and transportation platforms: A historical comparison of labour process reorganisation. New Technology Work and Employment, 2024.

[2] Farghali M, Osman A I, Mohamed I M A, Chen Z, Chen L, Ihara I, et al. Strategies to save energy in the context of the energy crisis: A review. Environmental Chemistry Letters, 2023: 1–37.

[3] Quinlan R. Auto MPG [dataset]. UCI Machine Learning Repository, Irvine (CA), 1993.

[4] Gupta H N. Fundamentals of Internal Combustion Engines. PHI Learning Pvt. Ltd., 2025.

[5] Lee N, Park J, Lee J, Park K, Choi M, Kim W. Estimation of fuel economy improvement in gasoline vehicle using cylinder deactivation. Energies, 2018, 11(11): 3084.

[6] Brusa A, Corti E, Rossi A, Moro D. Enhancement of heavy-duty engines performance and reliability using cylinder pressure information. Energies, 2023, 16(3): 1193.

[7] Sun Z, Premarathna W A A S, Anupam K, Kasbergen C, Erkens S M J G. A state-of-the-art review on rolling resistance of asphalt pavements and its environmental impact. Construction and Building Materials, 2024, 411: 133589.

[8] Saddique M S, Rehman F, Ahmad R, Tareen A, Iqbal M W, Khan M S. Drag reduction technology and devices for road vehicles. Heliyon, 2024, 10(16).

[9] Lakshminarayanan P A, Agarwal A K (eds.). Design and Development of Heavy Duty Diesel Engines: A Handbook. Singapore: Springer, 2019.

[10] Luo J, Lan Y, Jia Z, Chen G, Xu S, Zhang H, Jiang C. Development of flow control for road vehicles based on drag reduction: a review. Journal of the Brazilian Society of Mechanical Sciences and Engineering, 2025, 47: 581

[11] Alhazmi Y A. Electric vehicle battery swap stations: An overview and critical review. Journal of Umm Al-Qura University for Engineering and Architecture, 2024: 1–14.

[12] Kumar A. A comprehensive review of an electric vehicle based on the existing technologies and challenges. Energy Storage, 2024, 6(5): e70000.

[13] Senecal K, Leach F. Racing Toward Zero: The Untold Story of Driving Green. SAE International, 2021.

[14] Mandys F. Electric vehicles and consumer choices. Renewable and Sustainable Energy Reviews, 2021, 142: 110874.

[15] Ghani A A, Abdullah N, Saidur R, Pandey A K. Assessing technological-driven challenges and policies associated with electric vehicle (EV) adoption. Transport Policy, 2025.