STAT306 Course Project

Members
-Junhao Wen
-Weilin Sun
-Edmund Chu

1.Introduction

Nowadays, automobiles are one of the largest contributors to environmental pollution. Automobiles' impact on air quality, greenhouse gas emissions, and resource depletion is largely influenced by the efficiency of fuel they use. The effective utilization rate of fuel for automobiles can be transformed to the value of MPG((miles per gallon). MPG indicates the distance a vehicle can travel per gallon of fuel. A higher MPG value means the vehicle uses less fuel to travel the same distance, which is often interpreted as better fuel efficiency. Our motivation is to identify key factors of fuel efficiency, which can inform manufacturers, consumers, and policymakers about how various vehicle attributes impact fuel economy. Additionally, this analysis could contribute to broader discussions about reducing carbon emissions and promoting the development of more efficient vehicles.

Lower MPG results in higher fuel costs.Our team got curious if there are certain cost variables shown for us to interpret the spending of miles per gallon. Thus, throughout this project, we are going to answer the question "What factors influence city-cycle fuel consumption (miles per gallon) in automobiles?" To answer this question, the dataset “auto-mpg.data” by Quinlan, collected in 1993. Each row corresponds to each automobile and it includes information of automobiles’ such as their MPG, Displacement, Cylinders, Horsepower, Weight, Acceleration and Model_year. In terms of finding the best represented variable we introduce lasso to solve the high related corvariables issue and best fitted model to make sure we find the most significant model.

1.1 Data Description

1.1.1 Dataset Overview

MPG (miles per gallon)(continuous) - response variable, represents the city-cycle fuel consumption for each vehicle. This variable is the target for prediction and indicates fuel efficiency, with higher values suggesting better efficiency, and was measured during city-cycle driving tests in the 1970s and 1980s.
Displacement (Cubic inches)(continuous)- explanatory variable, measures the engine displacement in cubic inches. Displacement is an indicator of engine size and power, potentially influencing fuel consumption rates.
Cylinders(Count)(Integer) - explanatory variable, indicates the number of cylinders in the engine. Common values include 4, 6, and 8 cylinders, with higher values generally implying larger engines and potentially higher fuel consumption.
Horsepower(hp)(continuous) - explanatory variable, specifies the vehicle’s engine power in horsepower. This measure reflects the vehicle’s performance capacity, affecting the energy required and potentially influencing fuel consumption.
Weight(Pounds (lbs)) (continuous) - explanatory variable, represents the vehicle's weight in pounds, a factor influencing fuel consumption as heavier vehicles generally require more energy for movement.
Acceleration(Seconds) (continuous) - explanatory variable, measures the time taken to accelerate from 0 to 60 mph (continuous).
Model_year(Year (e.g., 70 = 1970, 80 = 1980))(Integer) - explanatory variable, indicates the model year of the vehicle (integer)
Car_name(Categorical)-explanatory variables, indicates the different brand of the vehicle(string)
Origin(Categorical)(integer)-explanatory variables, indicates different production distinct

1.1.2. Exploring and Processing Data

This study utilizes a large dataset comprising 398 automobile’s information after data cleaning
The original dataset underwent careful cleaning to:

- Checking the missing value and use the median value to impute the missing value
- Remove the variable car_name

This process resulting in ensuring high data quality by maintaining enough sample size and keeping eight direct quantitative related variables to the target variable mpg

2. Methodology

2.1 Variable Selection

Variable selection is part of the analytical process to determine which predictors are most relevant to the study's goals. We introduced the LASSO and Best subset method in the following research

2.1.1 model linear regression objective

a. Use adjusted R-squared to evaluate how well the model explains the variation in the dataset.
b. Avoid assumption violations and avoid overfitting.
c. Check the p-values of each variable to ensure statistical significance, keeping the model simple and interpretable.

```
Call:
lm(formula = mpg ~ ., data = mpg_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5958 -2.1547 -0.1183  1.9118 12.9942

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.791e+01  4.604e+00  -3.890 0.000118 ***
cylinders    -4.205e-01  3.220e-01  -1.306 0.192393
displacement  1.905e-02  7.515e-03   2.535 0.011620 *
horsepower   -1.239e-02  1.347e-02  -0.919 0.358415
weight       -6.697e-03  6.439e-04 -10.401  < 2e-16 ***
acceleration  9.811e-02  9.705e-02   1.011 0.312686
model_year    7.558e-01  5.051e-02  14.964  < 2e-16 ***
origin        1.424e+00  2.759e-01   5.163 3.88e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.337 on 390 degrees of freedom
Multiple R-squared:  0.8209,    Adjusted R-squared:  0.8177
F-statistic: 255.4 on 7 and 390 DF,  p-value: < 2.2e-16
```

Figure1. Full model linear regression summary

Observations: In the full model, MPG is the dependent variable.. Adjusted R^2=0.817 suggests a strong fit, but the presence of insignificant predictors such as cylinders,horsepower,acceleration might indicate that multicollinearity is inflating the overall R^2.

2.1.1 correlation and VIF objective

a. A high correlation (r>0.8 or r<−0.8) between any two variables indicates potential multicollinearity
b. VIF > 5 indicates problematic multicollinearity

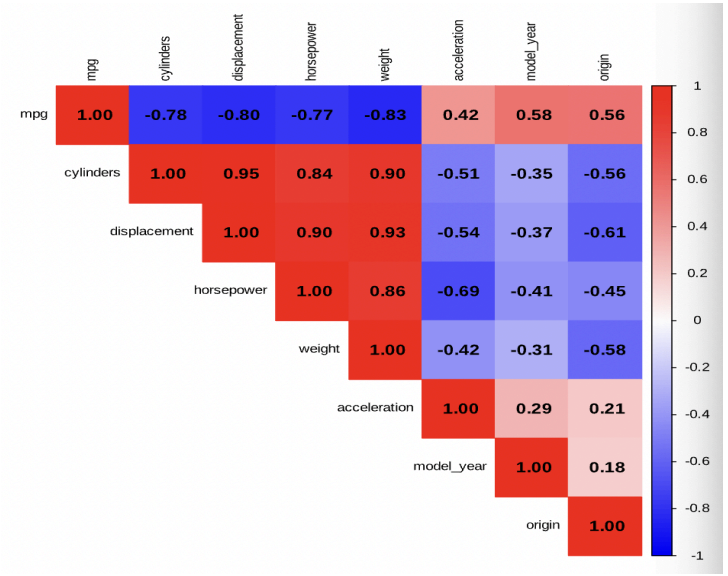


Figure 2. correlation matrix

```
cylinders    -0.7753963  1.0000000    0.9507214  0.8412845  0.8960168
displacement -0.8042028  0.9507214    1.0000000  0.8957782  0.9328241
```

Figure 3. VIF results

Observations: In the VIF plot, we can see some variables have high VIF values (displacement(21.89) and weight(10.60)), indicating potential multicollinearity issues in the data. From the correlation matrix, we have variables displacement and horsepower that have high correlation with nearly all other variables, which also indicate multicollinearity. Given the presence of multicollinearity, directly proceeding with model selection is not advisable. Multicollinearity can lead to unstable coefficient estimates, making it difficult to assess the true relationship between predictors and the response variable, so we cannot just start model selection directly (ie. forward selection)

2.1.3 Lasso Objective

A. LASSO regularization can shrink some coefficients to zero, effectively removing redundant variables
B. Address multicollinearity issues by selecting one predictor from a group of correlated variables

Observations: The variables selected by Lasso are **horsepower, weight, acceleration, model_year, and origin**. This selection helps reduce the likelihood of multicollinearity. To ensure a highly efficient model, we aim to simplify it by reducing the number of parameters while maintaining strong explanatory power for **mpg** in the model.

2.1.3 Best Subset selection Objective

C. Evaluate all possible combinations of predictors and choose the best fitted predictors
D. By selecting only the most relevant subset, the model reduces noise and variance which address overfitting

```
The results of best subset selection:
         horsepower weight acceleration model_year origin
1  ( 1 ) " "        "*"    " "          " "        " "
2  ( 1 ) " "        "*"    " "          "*"        " "
3  ( 1 ) " "        "*"    " "          "*"        "*"
4  ( 1 ) " "        "*"    "*"          "*"        "*"
5  ( 1 ) "*"        "*"    "*"          "*"        "*"
```

Table1. Best Subset selection table

Observations: Since we have less than 8 variables then using the best subset can find the best optimal combination for better than forward, backward or stepwise selection. Find the Model with the Maximum Adjusted R^2 Using Best Subset Selection, coming up with the model with four variables that has the max adjusted R^2. According to the best subset selection, we successfully chose weight, acceleration, model_year, origin as significant variables. And we will finding the relationships between the four variables and the MPG

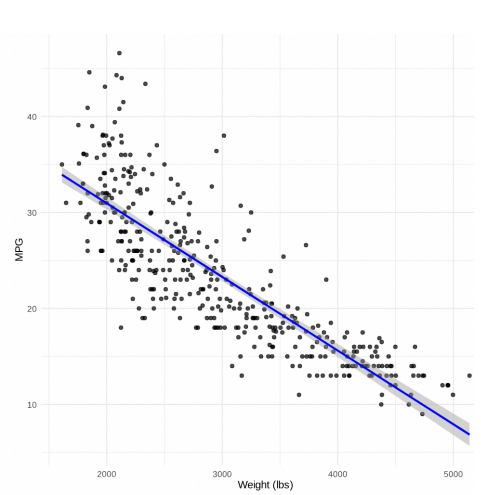
3. Explore the relationship between selected variables and MPG

a) Graph 1. Weight vs mpg:

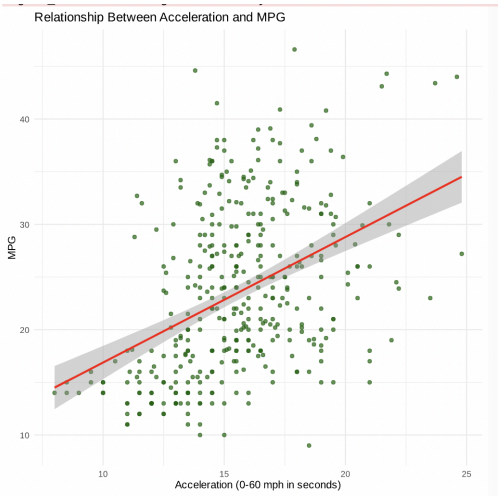
There is a clear negative linear relationship between Weight and MPG. Data points are most concentrated at the lower weights which indicate the Weight variable lighter vehicles tend to have higher fuel efficiency. The shaded area around the regression line represents the confidence interval and wider at the higher and lower end which interpret the uncertainty of the explanatory for extreme value.

b) Graph 2. acceleration vs mpg:

Most data points are centered around 1.5-2.0. The positive relationship between acceleration and MPG are formed. This plot suggests that vehicles with faster acceleration (longer times to reach 60 mph) are more fuel-efficient. However the large CI highlights the uncertainty for observations' variability. The two end-points with large variance reflect less certainty and reliability in predictions for very low or very high acceleration values.



Graph1



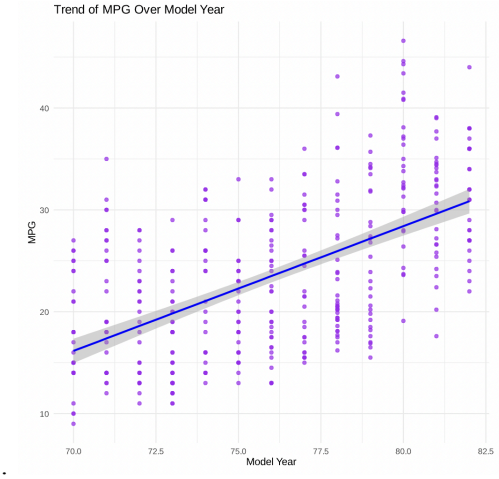
Graph2

c) Graph 3model_year vs mpg:

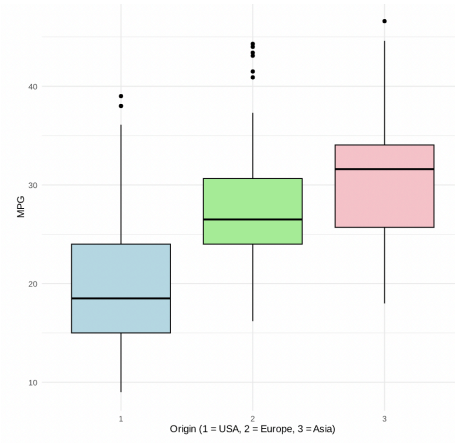
The vehicle with the same model year can have an extremely different amount of MPG. While in general, there is a positive relationship between model year and MPG, the newer model year coming with a higher MPG(lower efficiency).

d) Graph 4 origin vs mpg

The origin was divided into USA, Europe and Asia.The median of MPG for cars from the USA is noticeably lower than the other regions, reflecting lower average fuel efficiency. Asian cars have the highest median MPG, indicating the best average fuel efficiency. Since there is not a big difference between the USA and Europe, we combine them to form a new origin group. To see the influence of MPG between Asia group and USA combine Europe group.



Graph 3



Graph 4

4. Model Comparison and Selection

To better understand the impact of the selected variables, we conducted a multiple linear regression analysis using weight, acceleration, model_year, and origin, which were identified through the Best Subset Selection method. Since these variables were deemed essential, they were retained in all subsequent models. Instead of removing any variables, we focused on evaluating their contributions and interactions within the models.

4.1 full_additive_model and full_interaction_model

The full additive model yielded an adjusted R^2 of 0.816, indicating that 81.6% of the variation in MPG is explained by the model. While this is a strong fit, there is room for improvement. In the figure below , we can see that the VIF values of the full additive model were around 1.5, suggesting minimal multicollinearity. We managed to solve the problem by Lasso.

weight: 1.80881963764691 **acceleration:** 1.25732957979402 **model_year:** 1.1431020745088 **origin:** 1.51360279096128

To explore this, we tried the full interaction model, which includes all interaction terms, resulting in a higher adjusted R-squared of 0.8608 and a lower residual standard error of 2.916 compared to the additive model. However, the full interaction model includes many higher-order terms (e.g., weight*model_year*origin), making it difficult to interpret and lacking clear theoretical justification

```
Call:
lm(formula = mpg ~ weight + acceleration + model_year + origin,
    data = mpg_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8230 -2.1282 -0.0342  1.7693 13.1528

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.858e+01  4.012e+00  -4.630 4.97e-06 ***
weight       -5.930e-03  2.672e-04 -22.192  < 2e-16 ***
acceleration  7.195e-02  6.842e-02   1.052    0.294
model_year    7.464e-01  4.865e-02  15.341  < 2e-16 ***
origin        1.180e+00  2.581e-01   4.573 6.46e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.353 on 393 degrees of freedom
Multiple R-squared:  0.8179,    Adjusted R-squared:  0.816
F-statistic: 441.2 on 4 and 393 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = mpg ~ weight * acceleration * model_year * origin,
    data = mpg_dataset)

Residuals:
   Min     1Q Median     3Q    Max
-8.986 -1.905 -0.045  1.441 11.723

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                           -2.056e+02  2.407e+02  -0.854    0.394
weight                                 4.529e-02  9.006e-02   0.503    0.615
acceleration                          -1.156e-01  1.480e+01  -0.008    0.994
model_year                             3.019e+00  3.150e+00   0.958    0.338
origin                                 9.857e+01  1.839e+02   0.536    0.592
weight:acceleration                    1.496e-03  5.558e-03   0.269    0.788
weight:model_year                     -5.852e-04  1.172e-03  -0.499    0.618
acceleration:model_year                1.345e-02  1.942e-01   0.069    0.945
weight:origin                         -1.769e-02  7.701e-02  -0.230    0.818
acceleration:origin                    4.939e-02  1.095e+01   0.005    0.996
model_year:origin                     -1.195e+00  2.374e+00  -0.503    0.615
weight:acceleration:model_year        -2.474e-05  7.250e-05  -0.341    0.733
weight:acceleration:origin            -1.614e-03  4.625e-03  -0.349    0.727
weight:model_year:origin               1.773e-04  9.894e-04   0.179    0.858
acceleration:model_year:origin        -3.514e-03  1.416e-01  -0.025    0.980
weight:acceleration:model_year:origin  2.348e-05  5.947e-05   0.395    0.693

Residual standard error: 2.916 on 382 degrees of freedom
Multiple R-squared:  0.866,     Adjusted R-squared:  0.8608
F-statistic: 164.6 on 15 and 382 DF,  p-value: < 2.2e-16
```

Figure4. Full additive model linear regression summary Figure5. Full interaction model linear regression summary

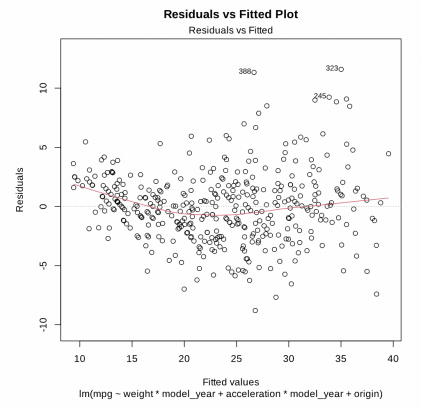
4.2 Selecting the a fitted interaction_model

To balance interpretability and model performance, we selected a simpler interaction model with key terms: **mpg ~ weight * origin + acceleration * origin + model_year.** This model has an adjusted R-squared of 0.8463 and a residual standard error of 3.064. All terms in the model are statistically significant ($p < 0.05$).

This interaction model captures meaningful relationships while avoiding overfitting. For example, interactions like weight * origin and origin * acceleration highlights how production area standards (origin) influence the effects of weight and acceleration on fuel efficiency. Additionally, model_year was included as a main effect to account for technological advancements, but it was excluded from interactions, as there was no evidence that it meaningfully interacted with other variables. Including redundant interactions (e.g., model_year with weight or acceleration) showed no significant improvement to the model and was therefore avoided.

By selecting this model, we achieve a good balance between interpretability and performance, ensuring the results remain both practical and theoretically grounded.

```
Call:
lm(formula = mpg ~ weight * origin + acceleration * origin +
    model_year, data = mpg_dataset)

Residuals:
   Min     1Q Median     3Q    Max
-9.322 -1.984 -0.088  1.669 11.260

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.256e+01  4.901e+00  -2.562  0.01078 *
weight              -4.617e-03  6.462e-04  -7.145 4.43e-12 ***
origin              -4.424e+00  2.250e+00  -1.967  0.04993 *
acceleration        -7.230e-01  1.574e-01  -4.594 5.88e-06 ***
model_year           7.924e-01  4.609e-02  17.191  < 2e-16 ***
weight:origin       -1.294e-03  4.968e-04  -2.605  0.00954 **
origin:acceleration  5.336e-01  9.545e-02   5.590 4.26e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.146 on 391 degrees of freedom
Multiple R-squared:  0.8404,    Adjusted R-squared:  0.838
F-statistic: 343.2 on 6 and 391 DF,  p-value: < 2.2e-16
```



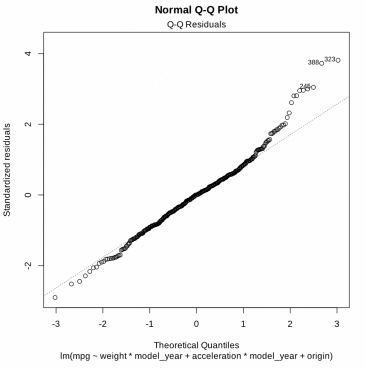


Figure6. fitted model linear regression summary *Graph5: Residual Plot of fitted interaction_model* *Graph6: QQ Plot of fitted interaction_model*

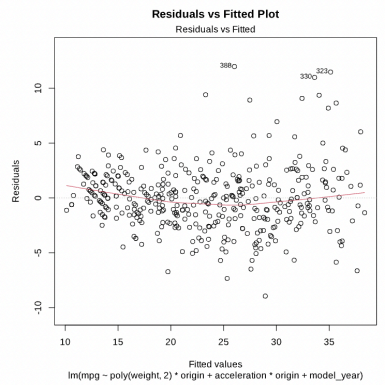
Observation:

a. The residuals are distributed around zero, indicating that the model captures the central trend of the data. However, the presence of a residual pattern suggests a potential non-linear relationship, indicating that the model may not fully capture the underlying relationship between predictors and the response variable. Additionally, heteroscedasticity is observed as the spread of residuals increases with higher fitted values.
b. The Normal Q-Q Plot shows significant deviations from the diagonal line, particularly in the tails, indicating that the residuals are not normally distributed. This violates the normality assumption.

5.Model improvement

a. To address the quadratic relationship observed in the residual plot, we introduce **weight²** into the model to capture the non-linear relationship. To mitigate potential multicollinearity issues, we use the **poly(weight)** function instead. By doing so, we can get a much better residual plot with a little concave pattern. Compared with the previous model it is much better.

```
Call:
lm(formula = mpg ~ poly(weight, 2) * origin + acceleration *
    origin + model_year, data = mpg_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
−8.9443 −1.6591 −0.0738  1.6036 11.9803

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             −30.42005    3.75716  −8.097 7.34e−15 ***
poly(weight, 2)1       −104.23830   23.55448  −4.425 1.25e−05 ***
poly(weight, 2)2         45.74459   12.48698   3.663 0.000283 ***
origin                   −7.64892    1.68193  −4.548 7.26e−06 ***
acceleration             −0.61799    0.14646  −4.220 3.05e−05 ***
model_year                0.83072    0.04275  19.433  < 2e−16 ***
poly(weight, 2)1:origin  −5.67928   22.37521  −0.254 0.799769
poly(weight, 2)2:origin −13.67615   11.36095  −1.204 0.229404
origin:acceleration       0.49852    0.08989   5.546 5.41e−08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.9 on 389 degrees of freedom
Multiple R−squared:  0.8651,    Adjusted R−squared:  0.8624
F−statistic: 311.9 on 8 and 389 DF,  p−value: < 2.2e−16
```



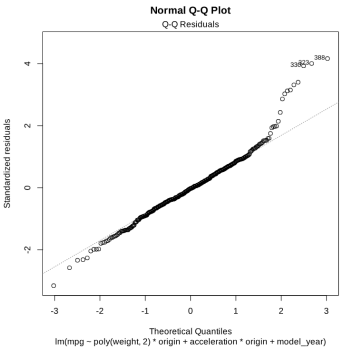
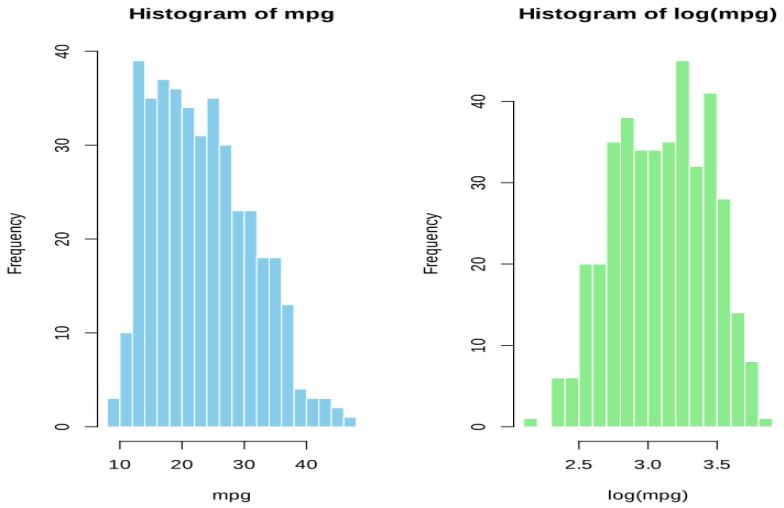
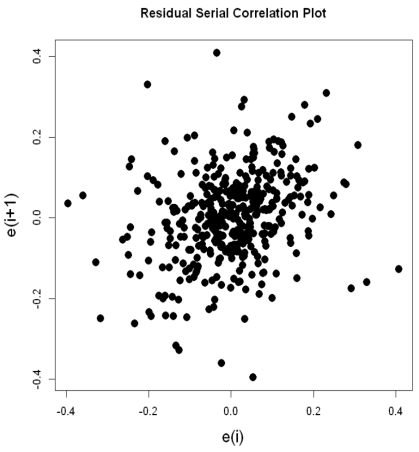


Figure7. fitted model linear regression summary *Graph7: Residual Plot of fitted interaction_model* *Graph8: QQ Plot of fitted interaction_model*

b. To resolve the violation of the constant variance assumption, we apply a log transformation. This effectively addresses heteroscedasticity, as the residual spread becomes more uniform across the range of fitted values. The residuals are now well-centered around zero, with no apparent patterns, meeting the assumptions of randomness and constant variance.



Graph 9:Histogram Comparison: mpg vs. log(mpg)



Residual Serial Correlation Plot

A. The MPG histogram shows a right-skewed distribution, with more observations at lower values and a long tail at higher values, indicating violations of normality and heteroscedasticity. In contrast, the histogram of log(MPG) is more symmetric and bell-shaped after applying the log transformation. This transformation reduces skewness, stabilizes variance, and addresses non-normality and heteroscedasticity, making the data more suitable for regression modeling and improving model robustness. (Graph 9)
B. The residual serial correlation plot shows a random scatter of points with no discernible pattern, clustering, or systematic structure, suggesting that the residuals are independent and uncorrelated. Most points are centered around the origin consistent with the assumption that residuals have a mean of zero. This supports the independence assumption in linear regression (Residual Serial Correlation Plot)

6. Conclusion and Discussion

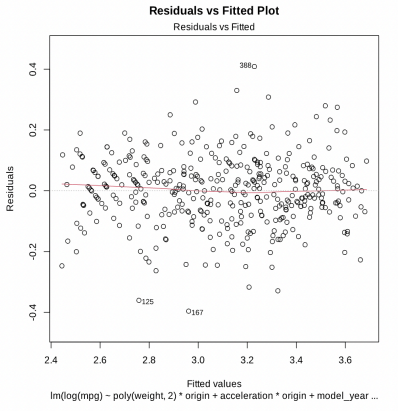
6.1 strengthen of the final model

We chose improved_model2 **lm(log(mpg) ~ poly(weight, 2) * origin + acceleration * origin + model_year**

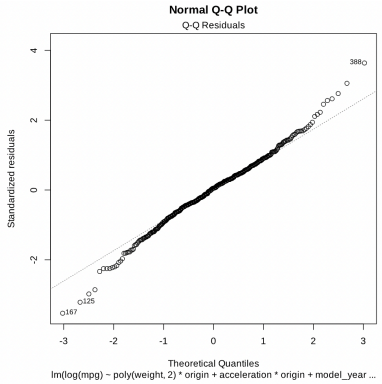
(Figure 10) as our final model, since it has the highest adjusted-R^2 at 0.8892 compared with the previous model. Indicating that approximately 88.9% of the variance in log(MPG) is explained by the predictors. It can be a strongly fitted model. And the RSE shrinked a significant amount to 0.1127 showing that the model's predictions are precise. Most of the terms have p value less than 0.05 which is considered to be significant.

```
Call:
lm(formula = log(mpg) ~ poly(weight, 2) * origin + acceleration *
    origin + model_year, data = mpg_dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.39593 -0.06364  0.00600  0.06636  0.40863

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.868624   0.146512   5.929 6.74e-09 ***
poly(weight, 2)1     -5.391537   0.918515  -5.870 9.35e-09 ***
poly(weight, 2)2      1.251996   0.486934   2.571 0.010506 *
origin               -0.223268   0.065588  -3.404 0.000733 ***
acceleration         -0.017285   0.005711  -3.027 0.002638 **
model_year            0.032583   0.001667  19.546  < 2e-16 ***
poly(weight, 2)1:origin  0.201685   0.872529   0.231 0.817320
poly(weight, 2)2:origin -0.442760   0.443024  -0.999 0.318220
origin:acceleration   0.015439   0.003505   4.405 1.37e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1131 on 389 degrees of freedom
Multiple R-squared:  0.8914,    Adjusted R-squared:  0.8892
F-statistic: 399.2 on 8 and 389 DF,  p-value: < 2.2e-16
```

Figure10. final model linear regression summary



Graph11: Residual Plot of fitted interaction_model



Graph12: QQ Plot of fitted interaction_model

6.2 Interpretation for the significant variables

Based on the regression output provided in the image:

Weight (poly(weight, 2)):
For Weight, The first-order term of the polynomial transformation of weight has a significantly negative coefficient of -5.391537, suggesting that as vehicle weight increases, the fuel efficiency (log of mpg) decreases significantly. However, the positive coefficient for the second-order term of 1.251996 indicates a non-linear relationship, where the rate of decrease in fuel efficiency diminishes for higher weights. This suggests a nuanced trade-off between weight and fuel efficiency.
Origin:
The negative coefficient of -0.223268 for the variable origin implies that vehicles originating from certain regions (e.g., imported vs. domestic) have a lower fuel efficiency on average, holding other factors constant.
Acceleration:
The negative coefficient of -0.017285 for acceleration shows that vehicles with higher acceleration times of slower vehicles tend to have lower fuel efficiency. This relationship aligns with reality expectations since quicker cars often optimize fuel efficiency less effectively.
Model Year:
The positive coefficient of 0.032583 for model_year indicates that newer vehicles generally have better fuel efficiency, likely due to advancements in technology and stricter environmental regulations.
origin:acceleration:
The positive coefficient of 0.015439 for the interaction term suggests that the relationship between origin and fuel efficiency depends on acceleration, with some origins being better optimized for certain performance levels.
poly(weight, 2):origin:
The coefficients for these interaction terms are not significant, indicating that the interaction between vehicle weight and origin does not significantly influence fuel efficiency.

6.3 Potential Limitation and improvement

The dataset, collected from 1970 to 1982, may not fully reflect current trends due to significant advancements in technology, societal norms, and regulations. Hence, the findings may be outdated and have limited relevance or applicability to present-day scenarios. In further recommendations for research interpreting the results and considering validating the model with more recent data could be beneficial and more practical for the modern day.

From the residual plot, there seems to be some curvature and potential patterns in the residuals plot. This indicates that the model might not fully capture non-linear relationships, which can be improved by Including additional non-linear terms. Heteroscedasticity violation can be improve by doing further transformations (other than log)

6.4 Interpretation and conclusion

To answer the research question: **"**What factors influence city-cycle fuel consumption (miles per gallon) in automobiles?" In this study by Lasso and Best subset we conclude that weight, origin, model_year and acceleration are the most significant factors to influence the MPG. In the model, having the origin of Asia as the baseline, keeping other factors remaining, the log of MPG will decrease when we switch the origin from the Asian group to the USA combined Europe group. Increasing weight and acceleration of automobiles will cause a negative growth to MPG, which can be the most influential factor to the fuel consumption.

In our model, a quadratic term for weight shows a nonlinear relationship with MPG, meaning that as weight increases, its marginal impact on fuel efficiency may vary in quadratic manner. This is critical for manufacturers to consider when designing vehicles, especially in balancing weight and MPG. The acceleration is intersection with origin in the final model. Vehicles that take **higher acceleration time** tend to have lower fuel efficiency (lower MPG). This might reflect old engines that are inefficient in utilizing fuel to achieve acceleration in some origins.

coding part is submitted by Junhao Wen