A Deep Dive into Fuel Economy: Regression Analysis of the Auto MPG Dataset

Isaac Edem Adoboe

2024-11-27

Contents

Introduction

The purpose of this report is to develop a regression model to predict miles per gallon (MPG) of cars based on various features such as weight, model year, and origin. The dataset used for this analysis, obtained from the UCI Machine Learning Repository, contains detailed information on car attributes, including their weight, model year, origin, and fuel efficiency, among others.

Due to the potential non-linear relationship between the features and MPG, a log-transformation of the dependent variable (MPG) is applied to improve the model's fit. This transformation is particularly useful for addressing skewness and stabilizing the variance of the residuals, which is a common issue in such datasets. The report will detail the steps involved in fitting the regression model, the evaluation of the model's assumptions, and the assessment of its performance in predicting car fuel efficiency.

Key aspects of the analysis include exploring the significance of each predictor, checking for the underlying assumptions of the regression model, and discussing the implications of the results. By applying this approach, the goal is to build a model that can provide reliable predictions of car MPG and yield insights into the relationship between car attributes and fuel efficiency.

Data Overview

The Auto MPG dataset is a dataset from the UCI Machine Learning Repository, commonly used for fuel efficiency analysis. It contains 392 records of automobiles manufactured between 1970 and 1982, along with 9 features that include both numerical and categorical predictors. The primary target variable, mpg (miles per gallon), measures fuel efficiency, while the predictors include attributes such as the number of cylinders, engine displacement, horsepower, weight, acceleration, model year, and country of origin. The dataset also includes the car name as an additional descriptive feature (I dropped this feature for this project analysis).

- Number of Instances: 392 observations
- Number of Features: 7

Preview of data

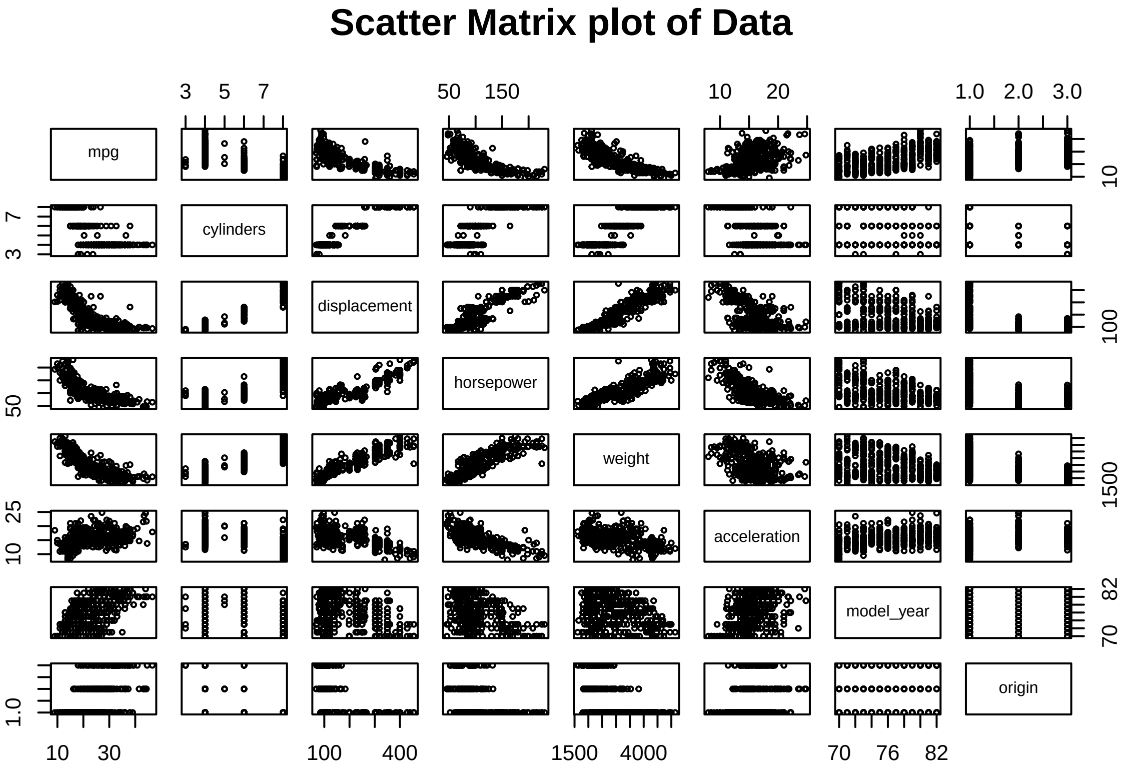
```
##   mpg cylinders displacement horsepower weight acceleration model_year origin
## 1  18         8          307        130   3504         12.0         70      1
## 2  15         8          350        165   3693         11.5         70      1
## 3  18         8          318        150   3436         11.0         70      1
## 4  16         8          304        150   3433         12.0         70      1
## 5  17         8          302        140   3449         10.5         70      1
## 6  15         8          429        198   4341         10.0         70      1
```

1. Y - mpg (Miles per Gallon): *Continuous*, the target variable that measures fuel efficiency.
2. X1 - cylinders: *Categorical*, the number of cylinders in the engine (In data - 3, 4, 5, 6, 8).
3. X2 - displacement: *Continuous*, the engine displacement in cubic inches.
4. X3 - horsepower: *Continuous*, the engine horsepower.
5. X4 - weight: *Continuous*, the weight of the car in pounds.
6. X5 - acceleration: *Continuous*, time taken to accelerate from 0 to 60 mph in seconds.
7. X6 - model year: *Categorical*, the model year of the car (e.g., 70 for 1970; in data - 1970 to 1980).
8. X7 - origin: *Categorical*, the region of manufacture (1 = USA, 2 = Europe, 3 = Japan).

Summary of data

```
##       mpg          cylinders      displacement     horsepower        weight
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
##   acceleration     model_year       origin
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000
##  Median :15.50   Median :76.00   Median :1.000
##  Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :24.80   Max.   :82.00   Max.   :3.000
```

Matrix plot of data



Methodology

Linear regression

In this section, we apply linear regression to model the relationship between the dependent variable, miles per gallon (mpg), and independent variables. The aim is to identify significant predictors of mpg and assess the model's performance using various diagnostic checks.

Regression function

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \epsilon$$

where:

- β_0 are the intercept
- $\beta_1 - - - \beta_7$ is the coefficients
- ϵ is the error term that accounts for the variation in Y not accounted for by X

Dummy variables

We generate dummy variables for the categorical variables in the data. To avoid multicollinearity problem, we drop one of the dummy variables.

Regression function after dummy encoding becomes,

$$Y = \beta_0 + \beta_{14}D_{14} + \beta_{15}D_{15} + \beta_{16}D_{16} + \beta_{18}D_{18} + \beta_2X_2 + \beta_3X_3 + ... + \beta_6X_6 + \beta_{72}D_{72} + \beta_{73}D_{73} + \epsilon$$

```
## 
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight + 
##     acceleration + model_year + origin, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -7.9267 -1.6678 -0.0506  1.4493 11.6002 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  30.9168415  2.3608985  13.095  < 2e-16 ***
## cylinders4    6.9399216  1.5365961   4.516 8.48e-06 ***
## cylinders5    6.6377310  2.3372687   2.840 0.004762 ** 
## cylinders6    4.2973139  1.7057848   2.519 0.012182 *  
## cylinders8    6.3668129  1.9687277   3.234 0.001331 ** 
## displacement  0.0118246  0.0067755   1.745 0.081785 .  
## horsepower   -0.0392323  0.0130356  -3.010 0.002795 ** 
## weight       -0.0051802  0.0006241  -8.300 1.99e-15 ***
## acceleration  0.0036080  0.0868925   0.042 0.966902    
## model_year71  0.9104285  0.8155744   1.116 0.265019    
## model_year72 -0.4903062  0.8038193  -0.610 0.542257    
## model_year73 -0.5528934  0.7214463  -0.766 0.443947    
## model_year74  1.2419976  0.8547434   1.453 0.147056    
## model_year75  0.8704016  0.8374036   1.039 0.299297    
## model_year76  1.4966598  0.8019080   1.866 0.062782 .  
## model_year77  2.9986967  0.8198949   3.657 0.000292 ***
## model_year78  2.9737783  0.7792185   3.816 0.000159 ***
## model_year79  4.8961763  0.8248124   5.936 6.74e-09 ***
## model_year80  9.0589316  0.8751948  10.351  < 2e-16 ***
## model_year81  6.4581580  0.8637018   7.477 5.58e-13 ***
## model_year82  7.8375850  0.8493560   9.228  < 2e-16 ***
## origin2       1.6932853  0.5162117   3.280 0.001136 ** 
## origin3       2.2929268  0.4967645   4.616 5.41e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.848 on 369 degrees of freedom
## Multiple R-squared:  0.8744, Adjusted R-squared:  0.8669 
## F-statistic: 116.8 on 22 and 369 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
## 
## Response: mpg
##               Df  Sum Sq Mean Sq  F value    Pr(>F)    
## cylinders      4 15274.5  3818.6 470.9373 < 2.2e-16 ***
## displacement   1  1098.0  1098.0 135.4069 < 2.2e-16 ***
## horsepower     1   588.0   588.0  72.5161  4.25e-16 ***
## weight         1   715.1   715.1  88.1961 < 2.2e-16 ***
## acceleration   1     7.7     7.7   0.9457    0.3315
## model_year    12  2960.4   246.7  30.4251 < 2.2e-16 ***
## origin         2   183.2    91.6  11.2972  1.73e-05 ***
## Residuals    369  2992.1     8.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Adequacy

After fitting the model of Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7, I conduct all model adequacy tests. I will be doing following checks.

1. Linearity of regression function
2. The assumption of constant error variance
3. Independence of error terms
4. The error terms are normally distributed
5. Lack of fit

Lack of fit

$$\textit{Full model}: \text{mpg}_i = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + \beta_6 \cdot X_6 + \beta_7 \cdot X_7 + \epsilon_i$$

$$\textit{Reduced model}: Y = \mu_{ij} + \epsilon_i$$

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

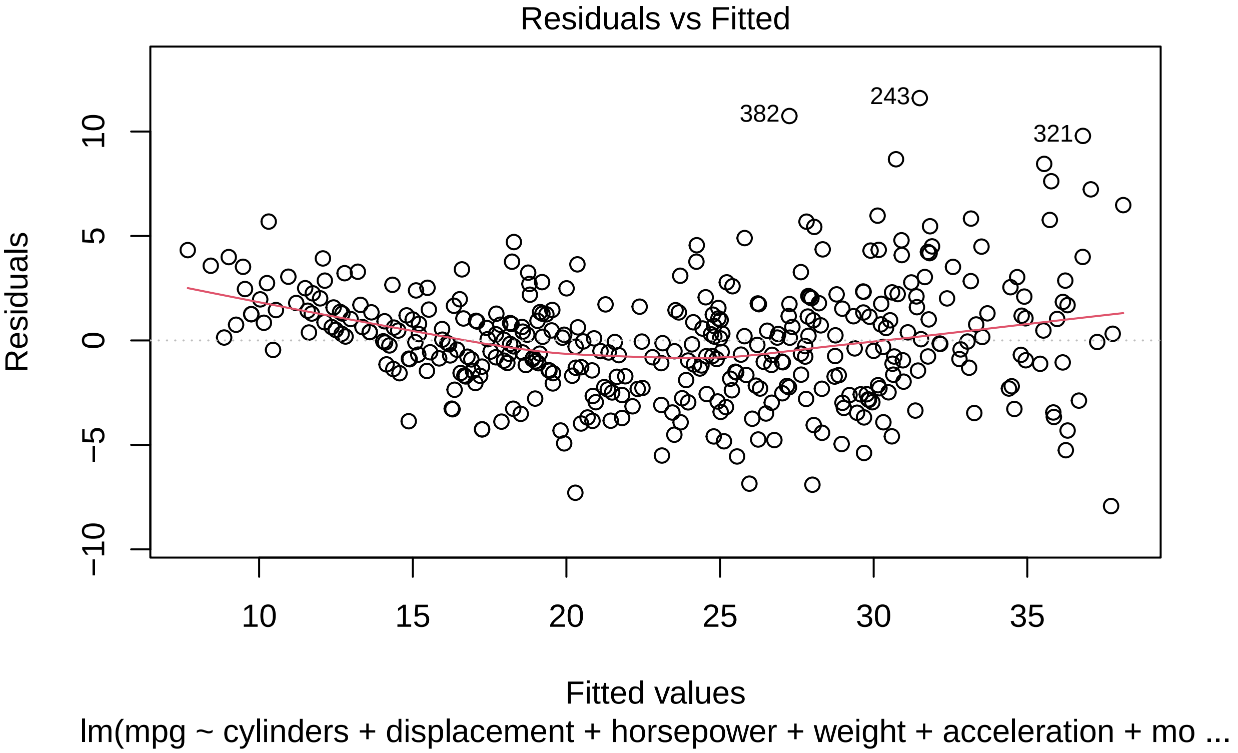
$$H_1: \text{At least one of } \beta_1, \beta_2, \dots, \beta_k \neq 0$$

Significance level $\alpha = 0.05$

```
## Analysis of Variance Table
##
## Model 1: mpg ~ 1
## Model 2: mpg ~ cylinders + displacement + horsepower + weight + acceleration +
##     model_year + origin
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    391 23819.0
## 2    369  2992.1 22     20827 116.75 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: Since the p-value is much smaller than the significance level of 0.05, we reject the null hypothesis. This suggests that the full model, including the predictors significantly improves the fit compared to the reduced model. We can say that the predictors explain a significant amount of the variation in fuel efficiency (mpg).

Linearity of regression function.



The plot of Residuals vs fitted shows a slight curvature. This suggests that there might be some potential non-linearity between the predictors and the response variable. [Check failed.]

The assumption of constant error variance

From the residuals vs fitted plot, we can observe that the residuals follow some kind of imaginary funnel shape. Ideally, the error terms should be randomly dispersed in a horizontal band. We can perform a Breusch-Pagan test to aid our decision making.

The Breush-Pagan test states that

for $e = \gamma_0 + \gamma_1 X_1 + - - - - + \gamma_7 X_7$,

$H_0 : \gamma_1, ..., \gamma_7 = 0$ - Constant variance

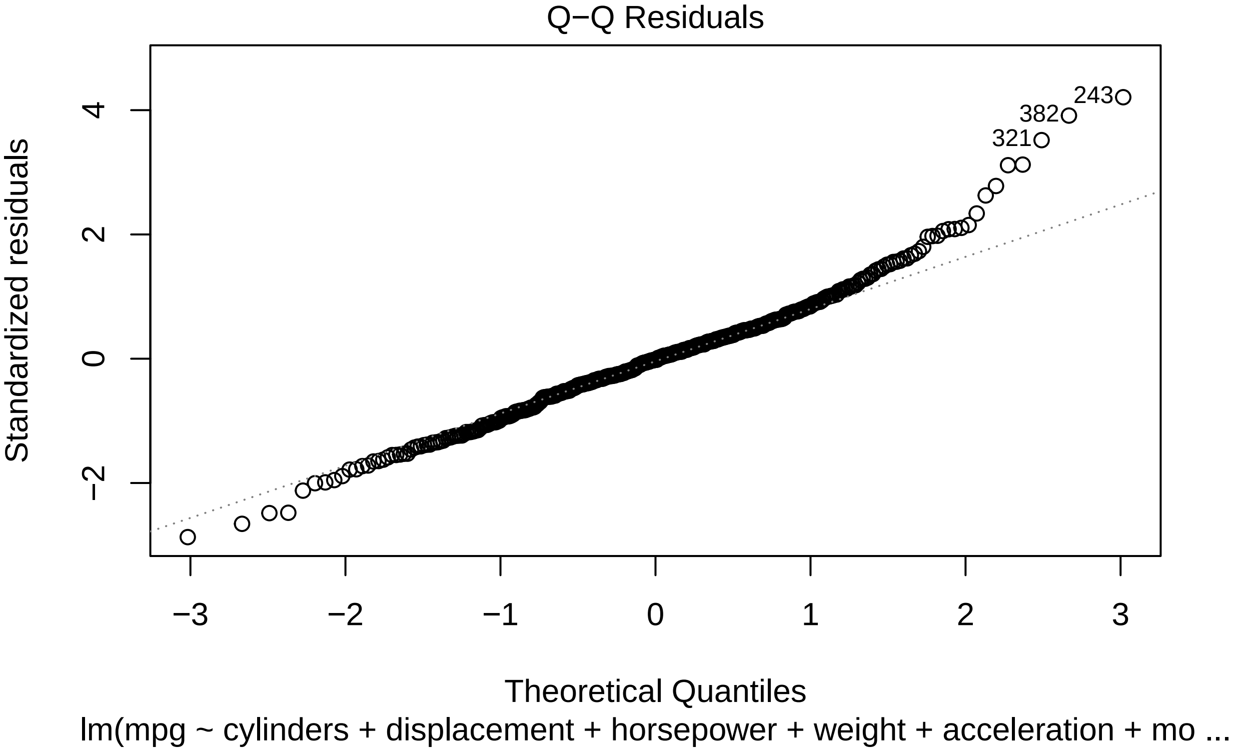
$H_1 : \gamma_1, ..., \gamma_7 = 0 \neq 0$ - Non-constant variance

At significance level, $\alpha = 0.05$,

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##              Data
##  -------------------------------
##  Response : mpg
##  Variables: fitted values of mpg
##
##         Test Summary
## ----------------------------
## DF            =    1
## Chi2          =    44.88062
## Prob > Chi2   =    2.0942e-11
```

Conclusion: p-value < 0.05: Since the p-value is approximately zero, we reject H_0 at significance level of 0.05. There is strong evidence to suggest that the variance of the errors is not constant, and heteroskedasticity is present in the model. [Check failed]

Normality of error terms

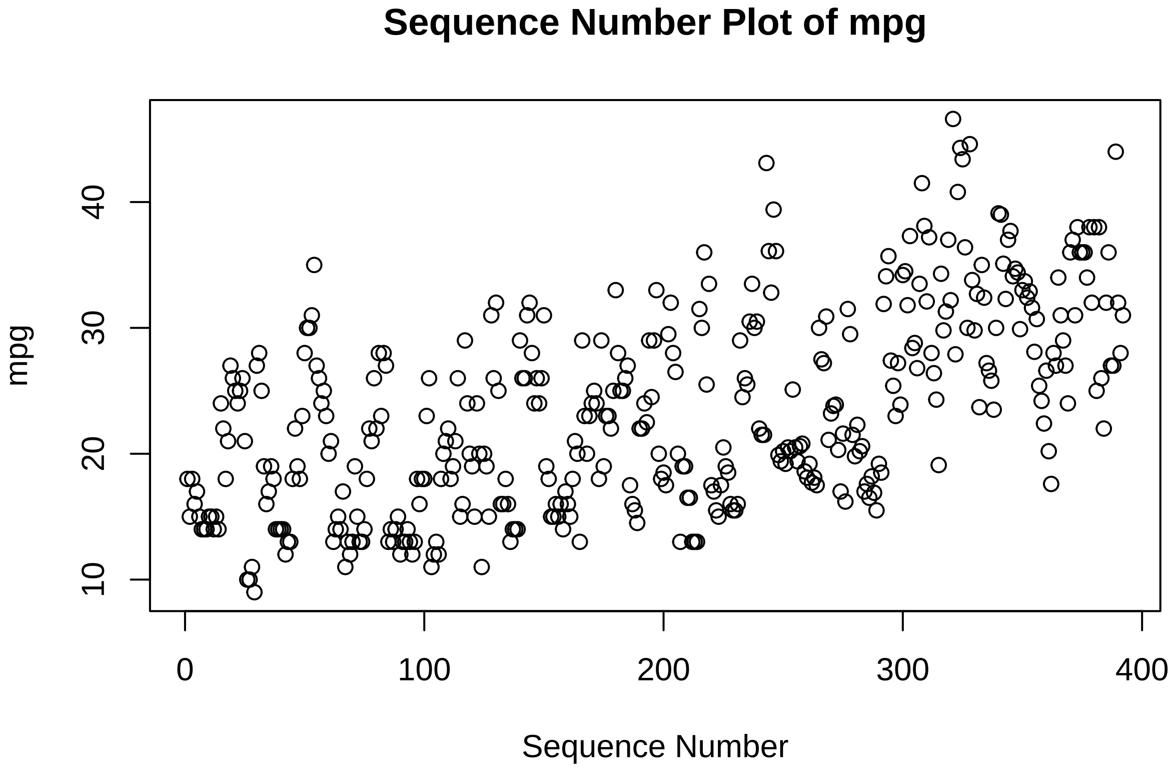


```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9783         0.0000
## Kolmogorov-Smirnov        0.0559         0.1725
## Cramer-von Mises         25.6061         0.0000
## Anderson-Darling          1.4226         0.0011
## -----------------------------------------------
```

Observation: We can see that most of the residuals are on the normal line and suggesting a central normality. We can see that at the tails of the data, there may be some potential outliers and leverage points.

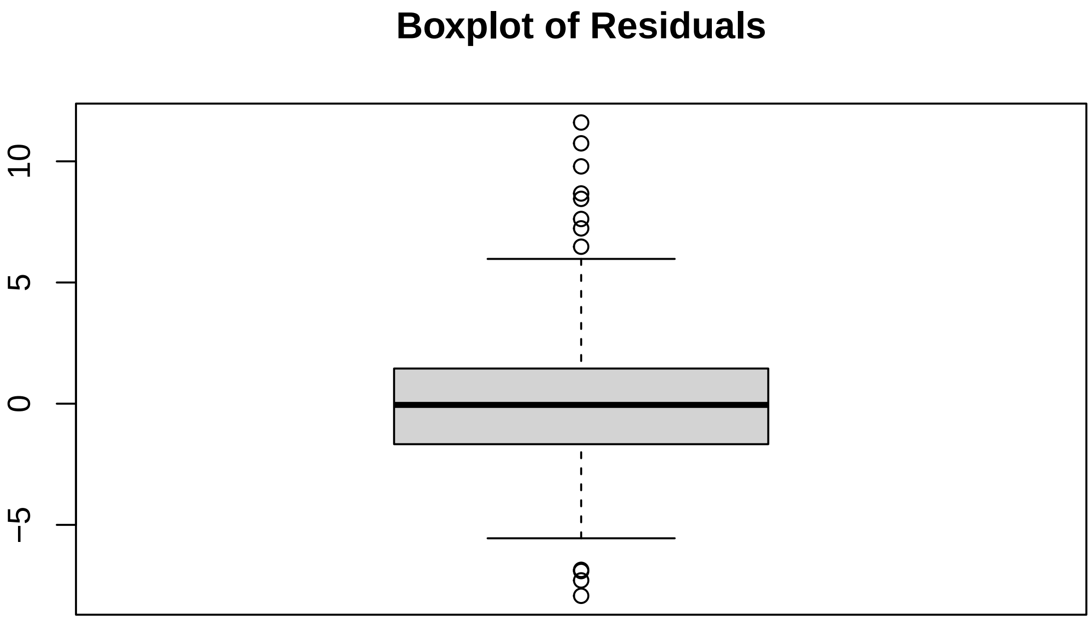
Conclusion: Assumption of normality of residuals was met. [Check passed]

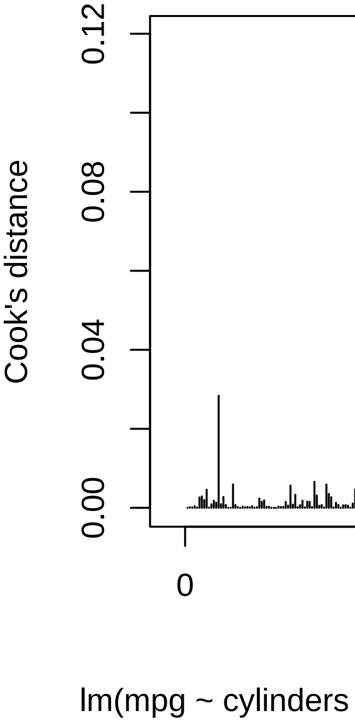
Independence of error terms



Because the sequence in which the observations were measured is known, I will construct a plot of residuals against sequence numbers. Because the cars were produced year after year, we assume that the measurements were taken sequentially which may cause adjacent measurements to be dependent which may in turn result in serial dependence and correlation between the observations, in particular, if the measurements were taken sequentially in time (like in this case). We do not see any particular pattern or alternating pattern between negative and positive values of residuals. [Check passed]

Outliers and leverage points





The plots of boxplot and cook's distance show that there are a number of outliers and leverage points in the data. There are 25 number of influential leverage points in the data.

```
cat("Number of influential leverage points detected:", length(outliers))
```

```
## Number of influential leverage points detected: 25
```

I performed some analysis on the outliers and find that these data points are unusual and do not generalize the trend of the data. They may be caused by cases of one of one sports cars and terrible vehicles with production failure. I decided to drop them.

```
##
## Call:
## lm(formula = lm(mpg ~ cylinders + displacement + horsepower +
##     weight + acceleration + model_year + origin, data = data_without_outliers))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.277 -1.475  0.000  1.512  6.004
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32.1860096  2.1820617  14.750  < 2e-16 ***
## cylinders4    7.4426436  1.6559815   4.494 9.54e-06 ***
## cylinders5    8.3509343  2.9011428   2.878 0.004246 **
## cylinders6    4.3015550  1.7636172   2.439 0.015231 *
## cylinders8    5.9044605  1.9385196   3.046 0.002499 **
## displacement  0.0080104  0.0056922   1.407 0.160250
## horsepower   -0.0373100  0.0110538  -3.375 0.000822 ***
## weight       -0.0046394  0.0005579  -8.316 2.15e-15 ***
## acceleration -0.1385315  0.0745897  -1.857 0.064130 .
## model_year71  0.8031849  0.6412433   1.253 0.211223
## model_year72 -0.5982364  0.6325445  -0.946 0.344934
## model_year73 -0.7654097  0.5728316  -1.336 0.182372
## model_year74  1.2376446  0.6750144   1.834 0.067591 .
## model_year75  1.1013179  0.6721217   1.639 0.102217
## model_year76  1.5278473  0.6333048   2.412 0.016366 *
## model_year77  2.7965159  0.6491184   4.308 2.15e-05 ***
## model_year78  2.4056137  0.6309467   3.813 0.000163 ***
## model_year79  4.6749205  0.6559235   7.127 6.06e-12 ***
## model_year80  7.5192792  0.7917373   9.497  < 2e-16 ***
## model_year81  5.8778241  0.6948481   8.459 7.82e-16 ***
## model_year82  6.8115348  0.6874848   9.908  < 2e-16 ***
## origin2       0.7565744  0.4306130   1.757 0.079812 .
## origin3       2.1256295  0.4028709   5.276 2.34e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.225 on 344 degrees of freedom
## Multiple R-squared:  0.911,  Adjusted R-squared:  0.9054
## F-statistic: 160.1 on 22 and 344 DF,  p-value: < 2.2e-16
```

We can compare the summary of the model without outlier and the model with outlier.

Metric	Model with outliers	Model without outliers
R-squared	0.8744	0.911
Adjusted R-squared	0.8669	0.9054

We can see a noticeable increase in the R-squared metrics for both models. Going forward, we will be using the data without the outliers for our modelling and analysis.

Remeditions

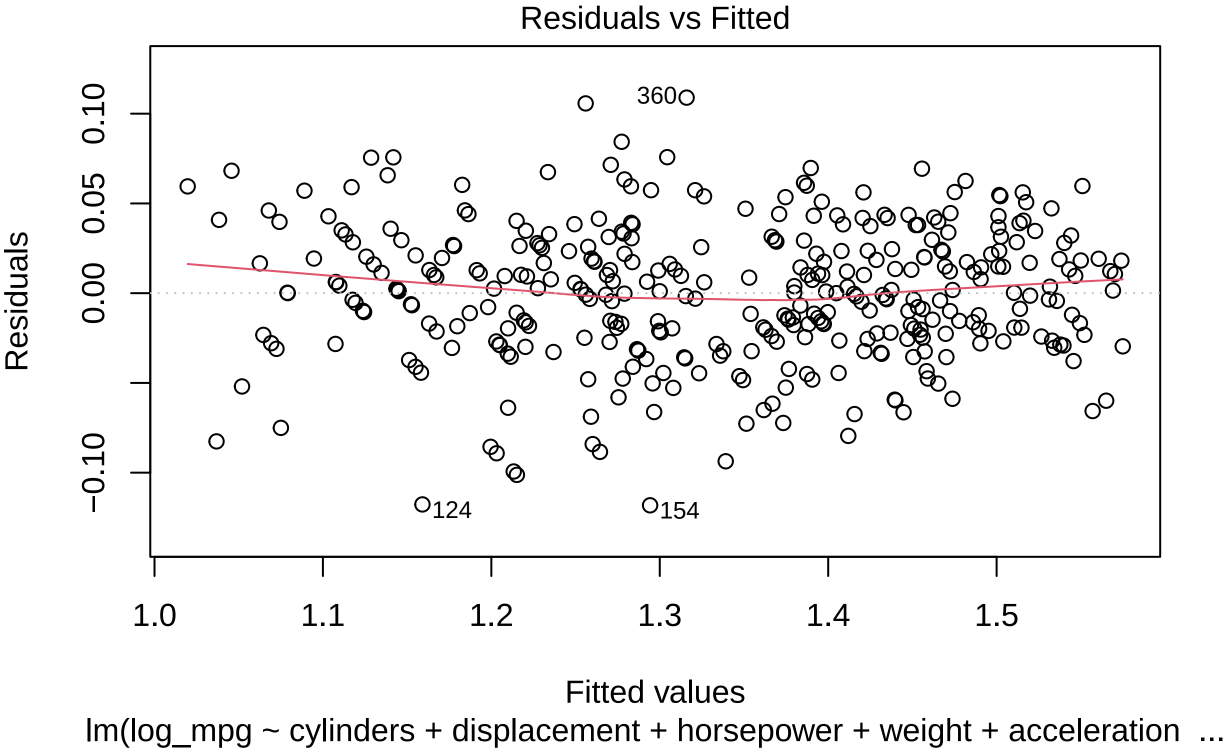
After performing model adequacy checks, we found that the model suffered from 2 main problems.

1. Linearity of the regression function
2. The assumption of constant error variance

Linearity issue fixed

For the linearity of the regression function challenge, I performed first a log transformation.

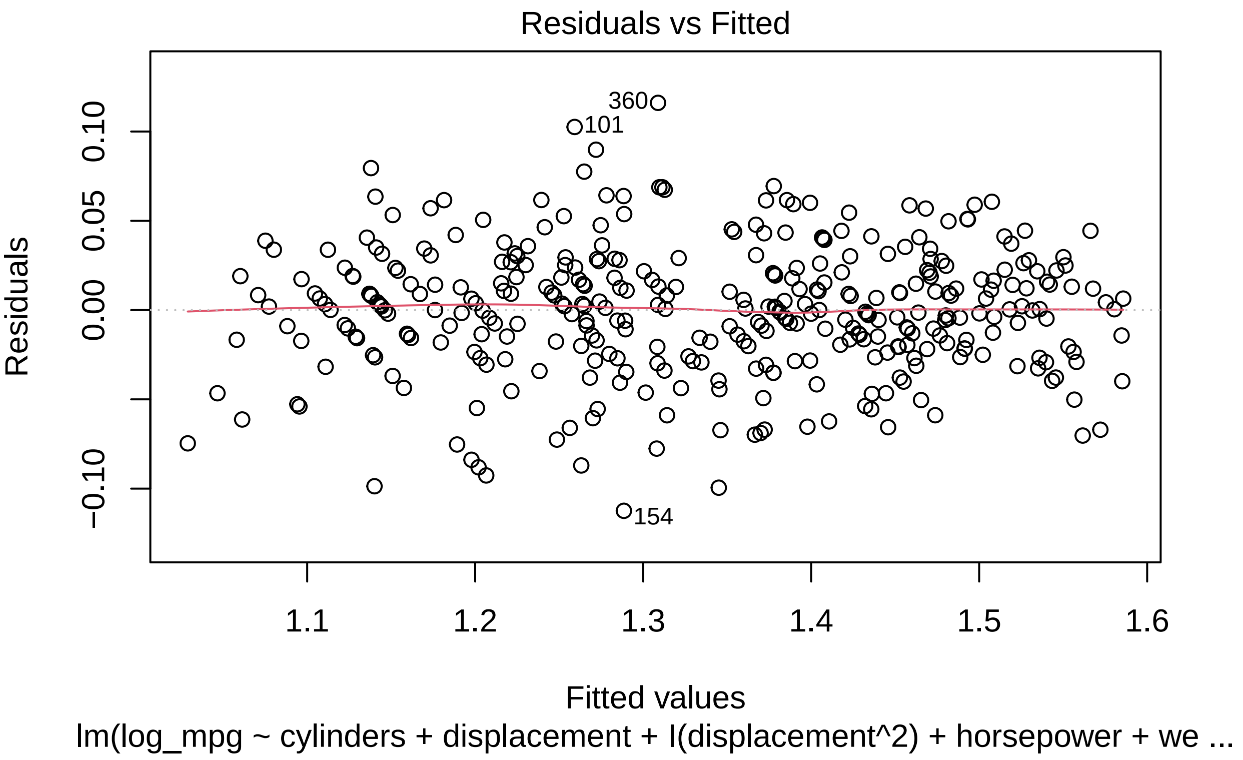
```
##
## Call:
## lm(formula = log_mpg ~ cylinders + displacement + horsepower +
##     weight + acceleration + model_year + origin, data = data_without_outliers)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.118172 -0.024030  0.001056  0.025736  0.108940
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.574e+00  3.801e-02  41.408  < 2e-16 ***
## cylinders4    1.258e-01  2.885e-02   4.362 1.71e-05 ***
## cylinders5    1.519e-01  5.054e-02   3.006 0.002843 **
## cylinders6    7.173e-02  3.072e-02   2.335 0.020132 *
## cylinders8    7.847e-02  3.377e-02   2.324 0.020726 *
## displacement  1.596e-04  9.915e-05   1.610 0.108332
## horsepower   -9.116e-04  1.926e-04  -4.734 3.22e-06 ***
## weight       -8.945e-05  9.719e-06  -9.204  < 2e-16 ***
## acceleration -3.509e-03  1.299e-03  -2.701 0.007263 **
## model_year71  1.303e-02  1.117e-02   1.167 0.244144
## model_year72 -9.005e-03  1.102e-02  -0.817 0.414373
## model_year73 -1.878e-02  9.978e-03  -1.882 0.060699 .
## model_year74  2.381e-02  1.176e-02   2.025 0.043679 *
## model_year75  2.522e-02  1.171e-02   2.154 0.031928 *
## model_year76  3.185e-02  1.103e-02   2.887 0.004132 **
## model_year77  5.733e-02  1.131e-02   5.070 6.52e-07 ***
## model_year78  5.421e-02  1.099e-02   4.933 1.27e-06 ***
## model_year79  9.296e-02  1.143e-02   8.136 7.54e-15 ***
## model_year80  1.254e-01  1.379e-02   9.091  < 2e-16 ***
## model_year81  1.032e-01  1.210e-02   8.530 4.73e-16 ***
## model_year82  1.151e-01  1.198e-02   9.611  < 2e-16 ***
## origin2       1.329e-02  7.501e-03   1.771 0.077423 .
## origin3       2.759e-02  7.018e-03   3.931 0.000102 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03875 on 344 degrees of freedom
## Multiple R-squared:  0.9301, Adjusted R-squared:  0.9256
## F-statistic: 208.1 on 22 and 344 DF,  p-value: < 2.2e-16
```



Notice that the curvature in the residuals vs fitted values plot, improves (i.e. it straightens).

When we introduce a polynomial of displacement based on the relationship displacement has with mpg (from matrix plot), the relationship between the residuals and the fitted values improves drastically.

```
## 
## Call:
## lm(formula = log_mpg ~ cylinders + displacement + I(displacement^2) + 
##     horsepower + weight + acceleration + model_year + origin, 
##     data = data_without_outliers)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.112421 -0.020495  0.001126  0.022127  0.116046 
## 
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        1.668e+00  3.944e-02  42.301  < 2e-16 ***
## cylinders4         1.533e-01  2.784e-02   5.506 7.21e-08 ***
## cylinders5         2.122e-01  4.915e-02   4.318 2.06e-05 ***
## cylinders6         1.424e-01  3.151e-02   4.518 8.62e-06 ***
## cylinders8         1.495e-01  3.425e-02   4.366 1.68e-05 ***
## displacement      -1.164e-03  2.391e-04  -4.871 1.70e-06 ***
## I(displacement^2)  2.155e-06  3.574e-07   6.029 4.26e-09 ***
## horsepower        -1.196e-03  1.893e-04  -6.318 8.24e-10 ***
## weight            -6.674e-05  9.992e-06  -6.680 9.67e-11 ***
## acceleration      -4.729e-03  1.254e-03  -3.772 0.000191 ***
## model_year71       8.845e-03  1.066e-02   0.830 0.407229    
## model_year72      -1.074e-02  1.050e-02  -1.024 0.306766    
## model_year73      -2.239e-02  9.521e-03  -2.352 0.019254 *  
## model_year74       1.696e-02  1.125e-02   1.507 0.132683    
## model_year75       2.328e-02  1.115e-02   2.087 0.037584 *  
## model_year76       2.852e-02  1.052e-02   2.711 0.007045 ** 
## model_year77       5.286e-02  1.079e-02   4.898 1.50e-06 ***
## model_year78       5.505e-02  1.047e-02   5.259 2.55e-07 ***
## model_year79       8.969e-02  1.089e-02   8.233 3.86e-15 ***
## model_year80       1.269e-01  1.314e-02   9.660  < 2e-16 ***
## model_year81       9.858e-02  1.155e-02   8.533 4.66e-16 ***
## model_year82       1.137e-01  1.141e-02   9.964  < 2e-16 ***
## origin2           -1.825e-03  7.570e-03  -0.241 0.809670    
## origin3            1.043e-02  7.263e-03   1.436 0.151860    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.0369 on 343 degrees of freedom
## Multiple R-squared:  0.9368, Adjusted R-squared:  0.9326 
## F-statistic: 221.1 on 23 and 343 DF,  p-value: < 2.2e-16
```



Non constant variance issue fixed

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##               Data
##  -----------------------------------
##  Response : log_mpg
##  Variables: fitted values of log_mpg
##
##         Test Summary
##  -----------------------------
##  DF            =    1
##  Chi2          =    3.963681
##  Prob > Chi2   =    0.04649193
```

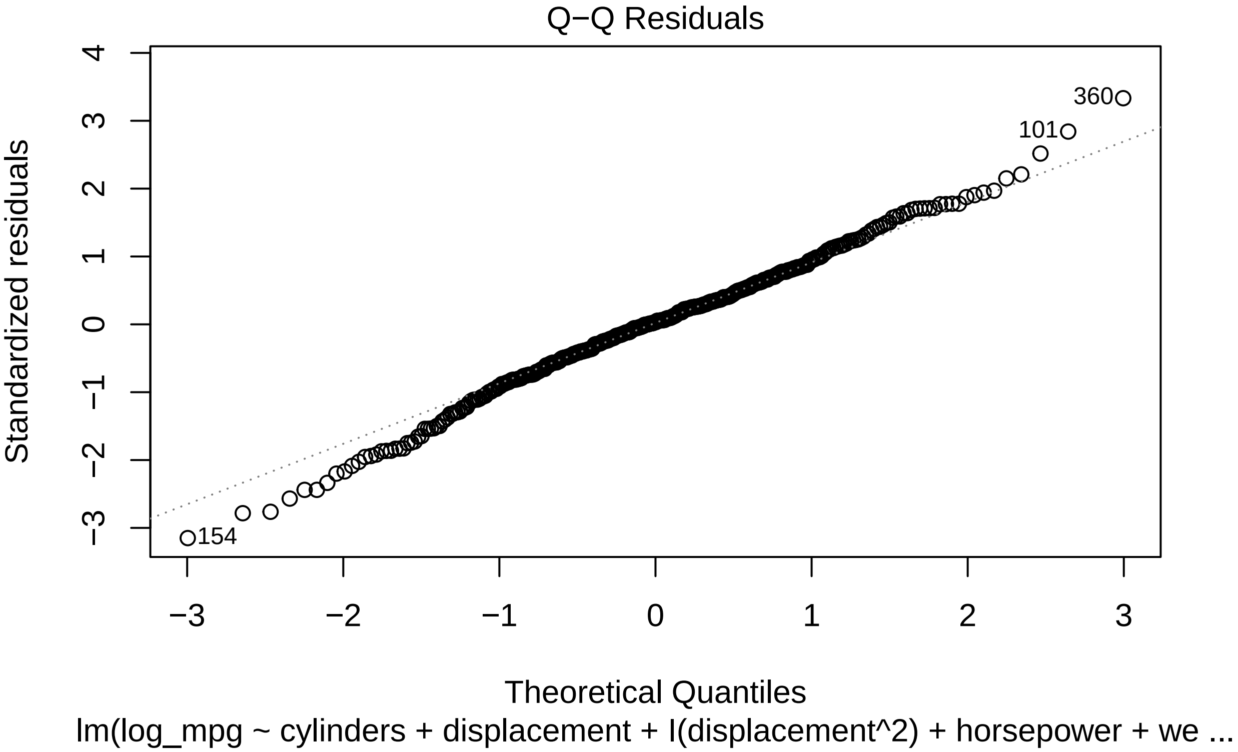
At $\alpha = 0.01$, this model passes the test of constant variance in the error terms.

New log-transformed regression function is

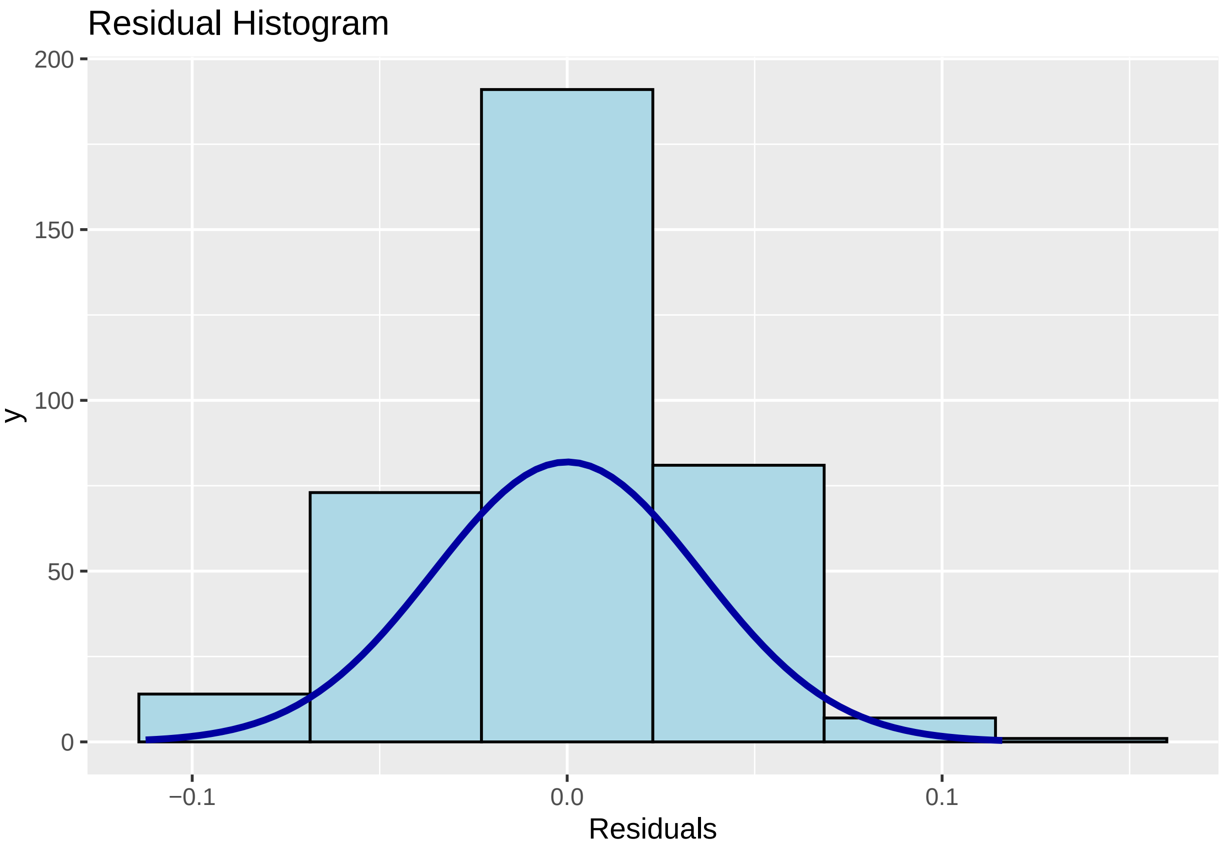
$$\log_{10}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_2^2 + \epsilon$$

where β_8 is the coefficient of the polynomial of the engine displacement.

```
## Warning: not plotting observations with leverage one:
##   287
```



```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9941         0.1708
## Kolmogorov-Smirnov        0.0372         0.6886
## Cramer-von Mises        113.9333         0.0000
## Anderson-Darling          0.7364         0.0545
## -----------------------------------------------
```



Using the Kolmogorov-Smirnov test, the p-value is much greater than 0.05, so we fail to reject the null hypothesis. This implies that there is no significant deviation from a normal distribution, suggesting that the residuals do not violate the normality assumption.

Multicollinearity

```
##            Variables   Tolerance        VIF
## 1         cylinders4 0.019145481  52.231648
## 2         cylinders5 0.565213920   1.769242
## 3         cylinders6 0.022120107  45.207738
## 4         cylinders8 0.015960426  62.654967
## 5       displacement 0.005980896 167.199019
## 6  I(displacement^2) 0.012116507  82.532035
## 7         horsepower 0.072803838  13.735540
## 8             weight 0.050760051  19.700532
## 9       acceleration 0.340170265   2.939704
## 10      model_year71 0.479158151   2.086994
## 11      model_year72 0.477910015   2.092444
## 12      model_year73 0.431021339   2.320071
## 13      model_year74 0.445061913   2.246878
## 14      model_year75 0.409867277   2.439814
## 15      model_year76 0.398901008   2.506888
## 16      model_year77 0.467384842   2.139564
## 17      model_year78 0.437977419   2.283223
## 18      model_year79 0.443685019   2.253851
## 19      model_year80 0.515785015   1.938792
## 20      model_year81 0.422427454   2.367270
## 21      model_year82 0.404705753   2.470931
## 22           origin2 0.479977435   2.083431
## 23           origin3 0.446046705   2.241918
```

Findings

Variables like acceleration, horsepower, and model_year have relatively low VIFs, indicating these predictors are not highly collinear with others.

The predictors displacement, cylinders4, cylinders6, and cylinders8 all have VIFs greater than 10, suggesting that there is severe multicollinearity in your model.

Based on domain knowledge and common sense, I decide to drop some of the variables. I drop "cylinders" and "displacement" (displacement of engine which is the volume in the cylinders) because they are the same thing. These are dropped and the "horsepower" is taken to represent it because the horsepower is determine by the displacement of the engine which in turn determines the power output. When this is done, horsepower correlation reduces.

```
##       Variables  Tolerance       VIF
## 1    horsepower 0.08471256 11.804625
## 2        weight 0.11665330  8.572411
## 3  acceleration 0.38601116  2.590599
## 4  model_year71 0.48782783  2.049904
## 5  model_year72 0.50551246  1.978191
## 6  model_year73 0.43882538  2.278811
## 7  model_year74 0.46670742  2.142670
## 8  model_year75 0.42542335  2.350600
## 9  model_year76 0.41131034  2.431254
## 10 model_year77 0.48525665  2.060765
## 11 model_year78 0.45398989  2.202692
## 12 model_year79 0.45770867  2.184796
## 13 model_year80 0.54939936  1.820170
## 14 model_year81 0.44157565  2.264618
## 15 model_year82 0.44432627  2.250598
## 16      origin2 0.71512318  1.398360
## 17      origin3 0.59670544  1.675869
```

I suspect that this correlation is because of acceleration. This is because, the horsepower of a vehicle is a factor that determines the acceleration. We can decide to drop any of the two. I decide to drop horsepower.

```
##       Variables Tolerance      VIF
## 1        weight 0.4977742 2.008943
## 2  acceleration 0.7441422 1.343829
## 3  model_year71 0.5350064 1.869137
## 4  model_year72 0.5250699 1.904508
## 5  model_year73 0.4594646 2.176446
## 6  model_year74 0.5254390 1.903170
## 7  model_year75 0.5002622 1.998952
## 8  model_year76 0.4705890 2.124996
## 9  model_year77 0.5279563 1.894096
## 10 model_year78 0.4939772 2.024385
## 11 model_year79 0.5122760 1.952073
## 12 model_year80 0.6025053 1.659736
## 13 model_year81 0.5067216 1.973470
## 14 model_year82 0.4851150 2.061367
## 15      origin2 0.7171102 1.394486
## 16      origin3 0.6197449 1.613567
```

Model selection

The goal is to find the simplest model that provides a good fit. We perform step-wise forward propagation to reduce the complexity of the model by selecting only the most relevant predictors.

Which criteria to use?

1. R-squared - R-squared always increases as more predictors are added to the model, even if those predictors are irrelevant.
2. MSE - MSE depends on the scale of the dependent variable, so comparing models with different dependent variables can be misleading.
3. AIC - AIC penalizes models that add more parameters to increase the fit.
4. Mallow's Cp - This is a criterion used to select the best model by balancing the goodness of fit (R-squared) and the complexity of the model.

Based on the above, AIC and Mallow's Cp is our preferred criteria to model selection. I will select the model with the lowest AIC.

All possible outcomes

I performed stepwise for all possible models and go found that this was the best.

```
##    Index N                            Predictors  R-Square Adj. R-Square
## 1      1 1                                weight 0.7969716     0.7964154
## 3      2 1                            model_year 0.3793407     0.3583014
## 4      3 1                                origin 0.3360353     0.3323872
## 2      4 1                          acceleration 0.1805948     0.1783499
## 6      5 2                     weight model_year 0.9108499     0.9075667
## 7      6 2                         weight origin 0.8017664     0.8001281
## 5      7 2                   weight acceleration 0.7994361     0.7983342
## 10     8 2                     model_year origin 0.6175224     0.6023103
## 8      9 2               acceleration model_year 0.4579458     0.4379835
## 9     10 2                   acceleration origin 0.4246194     0.4198642
## 13    11 3              weight model_year origin 0.9147399     0.9110963
## 11    12 3        weight acceleration model_year 0.9108507     0.9073049
## 12    13 3            weight acceleration origin 0.8045842     0.8024250
## 14    14 3        acceleration model_year origin 0.6506767     0.6357483
## 15    15 4 weight acceleration model_year origin 0.9147459     0.9108486
##    Mallow's Cp
## 1    470.50747
## 3   2207.03886
## 4   2364.82370
## 2   3000.96531
## 6     26.99459
## 7    454.82311
## 5    462.38976
## 10  1233.21374
## 8   1886.33552
## 9   2003.15289
## 13    15.02478
## 11    28.99146
## 12   445.25490
## 14  1099.10320
## 15    17.00000
```

Forward step propagation

```
##
##
##                                  Stepwise Summary
## --------------------------------------------------------------------------------
## Step    Variable         AIC          SBC         SBIC         R2       Adj. R2
## --------------------------------------------------------------------------------
##  0      Base Model     -387.623     -379.812    -1432.603    0.00000    0.00000
##  1      weight         -970.771     -959.055    -2015.137    0.79697    0.79642
##  2      model_year    -1248.821    -1190.241    -2312.305    0.91085    0.90757
##  3      origin        -1261.195    -1194.803    -2326.312    0.91474    0.91110
## --------------------------------------------------------------------------------
##
## Final Model Output
## ------------------
##
##                            Model Summary
## -----------------------------------------------------------------
## R                        0.956       RMSE                  0.041
## R-Squared                0.915       MSE                   0.002
## Adj. R-Squared           0.911       Coef. Var             3.170
## Pred R-Squared           0.907       AIC               -1261.195
## MAE                      0.033       SBC               -1194.803
## -----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##  AIC: Akaike Information Criteria
##  SBC: Schwarz Bayesian Criteria
##
##                                  ANOVA
## ---------------------------------------------------------------------
##               Sum of
##              Squares         DF    Mean Square       F         Sig.
## ---------------------------------------------------------------------
## Regression      6.762         15          0.451    251.054    0.0000
## Residual        0.630        351          0.002
## Total           7.392        366
## ---------------------------------------------------------------------
##
##                                     Parameter Estimates
## -----------------------------------------------------------------------------------------
##        model      Beta    Std. Error    Std. Beta         t        Sig       lower    upper
## -----------------------------------------------------------------------------------------
##  (Intercept)     1.634         0.015                  109.202    0.000       1.605    1.664
##       weight     0.000         0.000       -0.730     -34.980    0.000       0.000    0.000
## model_year71     0.029         0.011        0.053       2.501    0.013       0.006    0.051
## model_year72     0.003         0.011        0.006       0.277    0.782      -0.019    0.025
## model_year73    -0.012         0.010       -0.026      -1.132    0.259      -0.033    0.009
## model_year74     0.042         0.012        0.075       3.569    0.000       0.019    0.064
## model_year75     0.043         0.011        0.082       3.819    0.000       0.021    0.065
## model_year76     0.047         0.011        0.097       4.374    0.000       0.026    0.069
## model_year77     0.074         0.011        0.136       6.434    0.000       0.051    0.097
## model_year78     0.063         0.011         0.124      5.703    0.000      0.042    0.085
## model_year79     0.111         0.011         0.208      9.785    0.000      0.089    0.134
## model_year80     0.156         0.014         0.224     11.393    0.000      0.129    0.183
## model_year81     0.126         0.012         0.227     10.551    0.000      0.102    0.149
## model_year82     0.142         0.012         0.265     12.030    0.000      0.119    0.165
##      origin2     0.024         0.007         0.061      3.314    0.001      0.010    0.038
##      origin3     0.024         0.007         0.067      3.401    0.001      0.010    0.038
## ----------------------------------------------------------------------------------------
```

Weight significantly reduced the AIC, indicating it had a strong relationship with mpg. Model year, cylinders, horsepower, and other variables were progressively added, improving the model and AIC at each step. After using stepwise forward propagation, based on AIC, it suggests that we drop the "origin" variable from the model. The final model has a high Adj. R-squared of 0.932.

Findings

- The models with weight, model_year, and origin provide the best balance of R-squared, adjusted R-squared, and Mallow's Cp, with very high explanatory power and relatively low complexity.
- Adding more predictors, like acceleration, does not provide significant improvements to the explanatory power, and the complexity (by Mallow's Cp) increases.
- The model is statistically significant (based on the F-statistic and p-values), meaning that the predictors explain the variance in the dependent variable in a meaningful way.

Final model

```
##
## Call:
## lm(formula = log_mpg ~ weight + model_year + origin, data = data_without_outliers)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.137717 -0.025160  0.003443  0.029138  0.115582
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.634e+00  1.497e-02 109.202  < 2e-16 ***
## weight       -1.210e-04  3.459e-06 -34.980  < 2e-16 ***
## model_year71  2.873e-02  1.149e-02   2.501 0.012825 *
## model_year72  3.145e-03  1.134e-02   0.277 0.781768
## model_year73 -1.188e-02  1.050e-02  -1.132 0.258531
## model_year74  4.151e-02  1.163e-02   3.569 0.000409 ***
## model_year75  4.294e-02  1.124e-02   3.819 0.000158 ***
## model_year76  4.745e-02  1.085e-02   4.374 1.61e-05 ***
## model_year77  7.393e-02  1.149e-02   6.434 4.07e-10 ***
## model_year78  6.350e-02  1.113e-02   5.703 2.50e-08 ***
## model_year79  1.114e-01  1.139e-02   9.785  < 2e-16 ***
## model_year80  1.559e-01  1.368e-02  11.393  < 2e-16 ***
## model_year81  1.257e-01  1.191e-02  10.551  < 2e-16 ***
## model_year82  1.419e-01  1.179e-02  12.030  < 2e-16 ***
## origin2       2.357e-02  7.111e-03   3.314 0.001014 **
## origin3       2.403e-02  7.065e-03   3.401 0.000749 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04237 on 351 degrees of freedom
## Multiple R-squared:  0.9147, Adjusted R-squared:  0.9111
## F-statistic: 251.1 on 15 and 351 DF,  p-value: < 2.2e-16
```

We can write the final equation in terms of mpg as $e^{(\beta_0+\beta_1\cdot\text{weight}+\beta_2\cdot\text{model\_year}+\beta_3\cdot\text{origin})}$

Model adequacy of final model

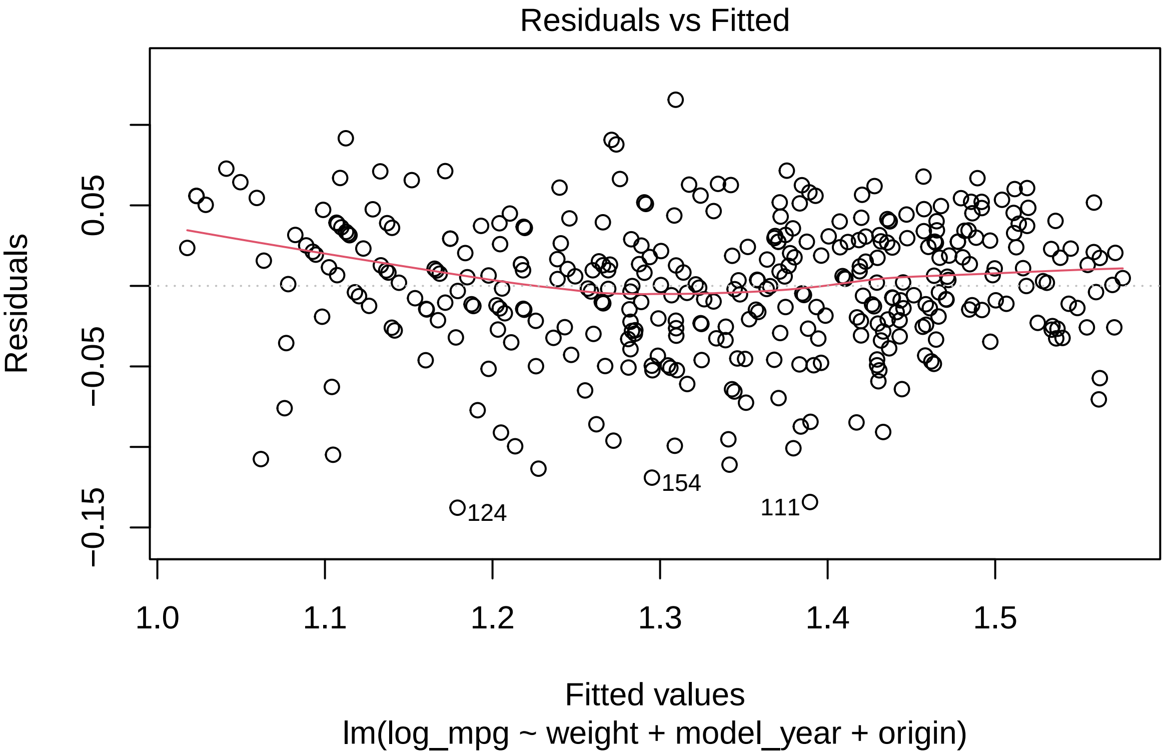
The Breush-Pagan test states that

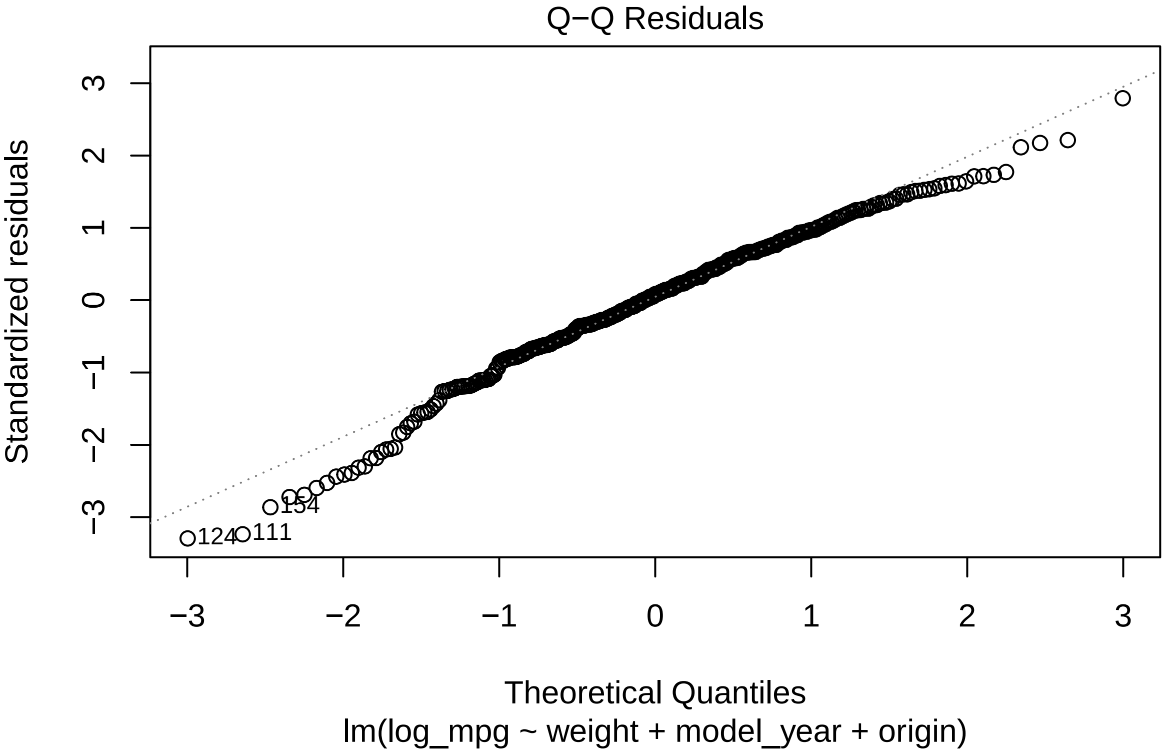
for $e = \gamma_0 + \gamma_1 X_1 + - - - - + \gamma_3 X_3$,

$H_0 : \gamma_1, ..., \gamma_3 = 0$ - Constant variance

$H_1 : \gamma_1, ..., \gamma_3 = 0 \neq 0$ - Non-constant variance

At significance level, $\alpha = 0.05$,





```
##
##  studentized Breusch-Pagan test
##
## data:  model_final
## BP = 18.402, df = 15, p-value = 0.2421
```

This test suggests that there is no significant evidence of heteroscedasticity, meaning the variance of the residuals appears to be constant across the range of fitted values.

From the plots above, we can conclude that, there is a significant linear relationship in the model.

Assumption of normality of error terms and constant variance was met.

Conclusion

In this report, a linear regression model was fitted to predict the log-transformed value of miles per gallon (mpg) based on the predictors: weight, model year, and origin. The model aimed to understand the relationship between these variables and mpg while accounting for the inherent skewness in the mpg data by applying a logarithmic transformation.

Key Findings:

- The analysis revealed significant relationships between the independent variables (weight, model year, and origin) and the log-transformed mpg.
- Weight had a negative association with mpg, indicating that as the weight of the vehicle increases, the mpg tends to decrease.
- Model year showed a positive relationship, suggesting that newer vehicles generally achieve higher mpg.
- Origin also contributed to the model, with differences in mpg depending on the vehicle's country of origin.

Model Interpretation:

- The coefficients of the fitted model can be interpreted on the log scale. However, to convert the model back to its original scale, the exponentiation of the predicted values was performed. This allowed for predictions of mpg in its original unit (miles per gallon).

Limitations:

- While the model provides valuable insights, there are potential limitations, including the exclusion of other potential predictors.

Benefits of results:

- The results can be valuable for stakeholders in the automotive industry, particularly for manufacturers and consumers interested in optimizing vehicle design for better fuel efficiency. Understanding the impact of weight, model year, and origin on mpg can inform vehicle development and purchasing decisions.