STAT481 Project

Yelizaveta Semikina -- Spring 2023

1. Data Introduction and Description.

In this project, we will analyze a dataset from the 1974 Motor Trend US magazine, which includes fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). The data considered in this project includes 32 observations on 5 numeric variables: mpg, disp, hp, wt, and qsec. Our goal is to answer research questions through regression analysis, checking assumptions for each model and drawing conclusions based on the best model possible. The research questions include examining which predictors explain some of the variation in the fuel consumption data, and determining the effect of gross horsepower, displacement, and weight on average mpg after taking into account all other predictors. We will be using R language.

2. Section 1. Data Introduction and Description.

Descriptive Statistics Summary Table.

	mpg	disp	hp	wt	qsec
Min.	10.40	71.1	52.0	1.513	14.50
1st Qu.	15.43	120.8	96.5	2.581	16.89
Median	19.20	196.3	123.0	3.325	17.71
Mean	20.09	230.7	146.7	3.217	17.85
3rd Qu.	22.80	326.0	180.0	3.610	18.90
Max.	33.90	472.0	335.0	5.424	22.90

This table shows descriptive statistics for five numeric variables: mpg, disp, hp, wt, and qsec. For each variable, the table lists the minimum value, first quartile (25th percentile), median (50th percentile), mean, third quartile (75th percentile), and maximum value. These statistics provide information on the distribution and central tendency of each variable. For example, we can see that the mean mpg is 20.09, the median is 19.20, and the range of values is from 10.40 to 33.90. The table can be used to compare the variability and distribution of each variable, and to identify any potential outliers or unusual values.

Table of Means.

mpg	disp	hp	wt	qsec
20.09062	230.72188	146.68750	3.21725	17.84875

This table displays the mean values of the five variables in the dataset: mpg, disp, hp, wt, and qsec. The mean value of mpg is 20.09062, indicating that the average fuel consumption for the 32 cars is around 20 miles per gallon. The mean value of disp (engine displacement) is 230.72188 cubic inches, while the mean value of hp (gross horsepower) is 146.68750. The mean value of wt (weight in 1000 lbs) is 3.21725, and the mean value of qsec (1/4 mile time in seconds) is 17.84875. These mean values provide a useful summary of the central tendency of the dataset for each variable.

Table of Variances.

mpg	disp	hp	wt	qsec
36.324103	15360.799829	4700.866935	0.957379	3.193166

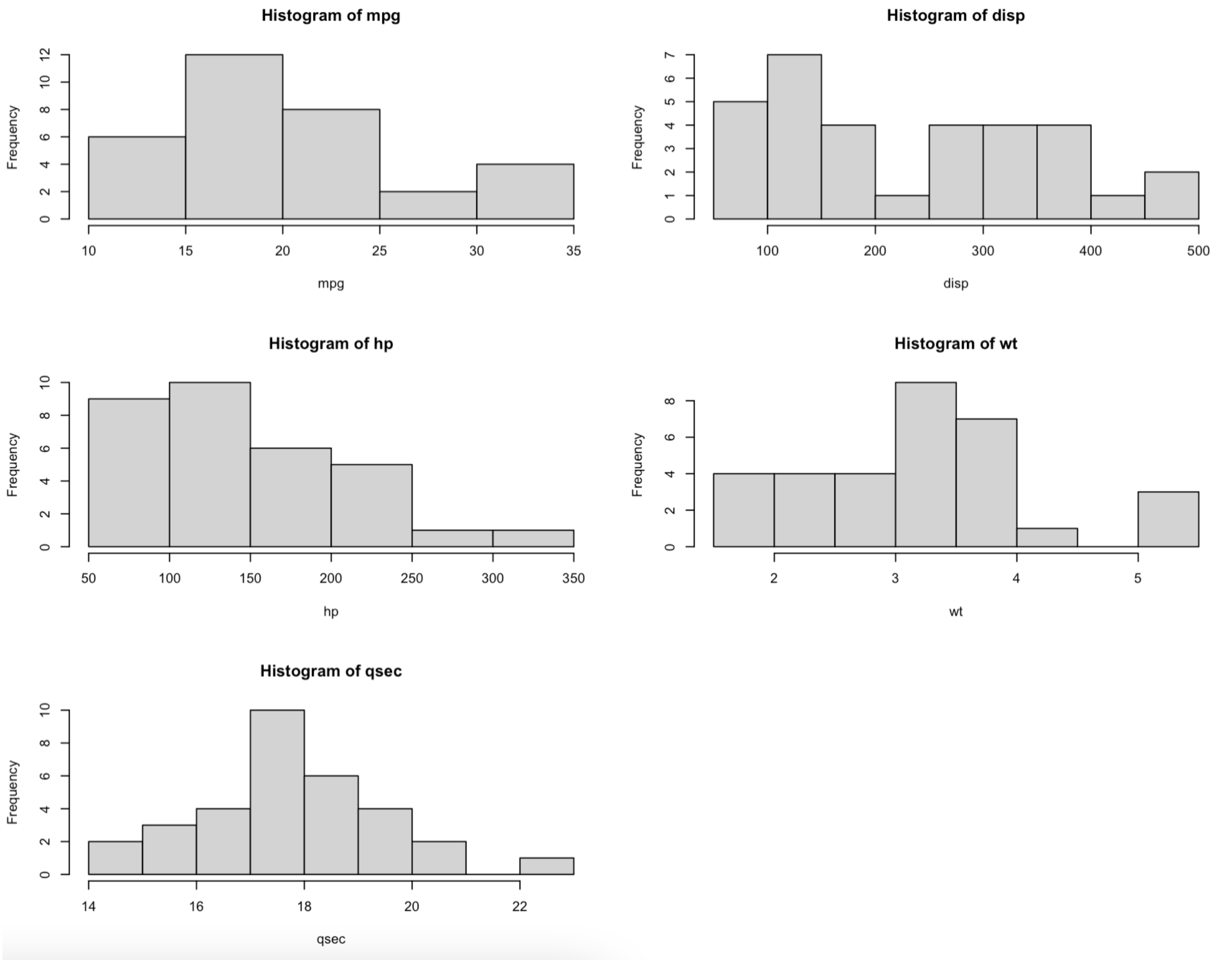
This table shows the variances for each of the variables in the dataset, namely mpg, disp, hp, wt, and qsec.

Standard Deviations of Variables.

mpg	disp	hp	wt	qsec
6.0269481	123.9386938	68.5628685	0.9784574	1.7869432

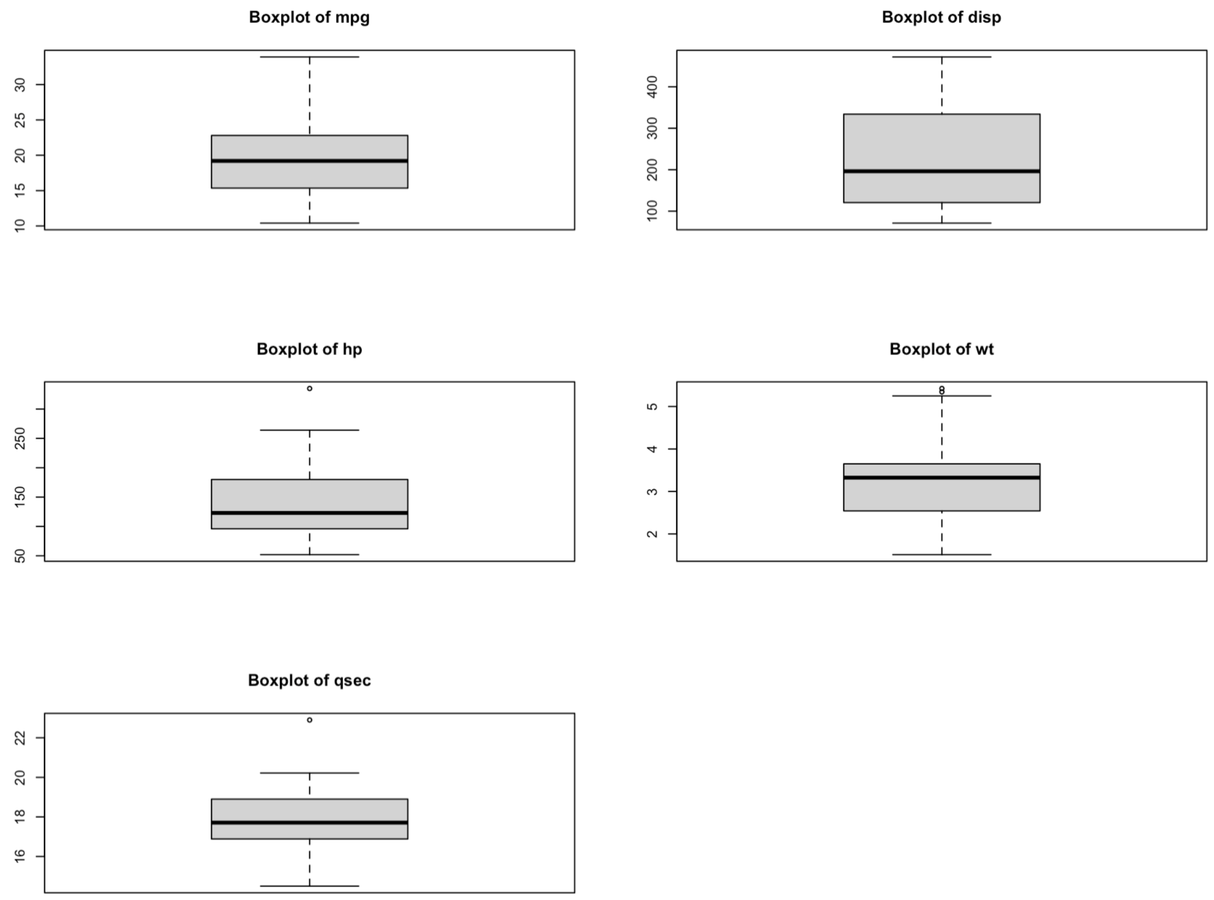
This table shows the standard deviation of each variable in the dataset, with values given for mpg, disp, hp, wt, and qsec.The values indicate the amount of variation present in the dataset for each variable, with higher values indicating more variability. For example, the standard deviation of mpg is 6.027, meaning that the observed values for mpg in the dataset vary by an average of 6.027 from the mean value of mpg. Similarly, the standard deviation of disp is 123.939, indicating that the observed values for disp vary by an average of 123.939 from the mean value of disp.

Histograms.



We have histograms for the five variables in the dataset: mpg, disp, hp, wt, and qsec.
The first histogram shows the distribution of mpg. It appears to be roughly normal with a peak around 20 mpg, but there are some cars that have lower or higher mpg values.
The second histogram shows the distribution of disp. It looks to be right-skewed, with most cars having smaller engine displacements and a few cars having much larger displacements.
The third histogram shows the distribution of hp. It looks to be roughly normal with a peak around 100 horsepower, but there are some cars that have much higher horsepower values.
The fourth histogram shows the distribution of wt. It looks to be roughly normal with a peak around 3,000 lbs, but there are some cars that are much lighter or much heavier.
The fifth histogram shows the distribution of qsec. It looks to be roughly normal with a peak around 18 seconds, but there are some cars that have much faster or much slower quarter mile times.

Boxplots.



The boxplots visualize the distribution of the data and identify any outliers. In this case, we have created boxplots for five variables - mpg, disp, hp, wt, and qsec.

The boxplot of mpg shows that the median is around 20, and the data are somewhat symmetrically distributed around the median. The minimum value is around 10, and the maximum value is around 35. There are no clear outliers.

The boxplot of disp shows that the median is around 200, and the data are skewed to the right. There are some outliers on the high end of the distribution.

The boxplot of hp shows that the median is around 150, and the data are also skewed to the right. There are some outliers on the high end of the distribution.

The boxplot of wt shows that the median is around 3, and the data are somewhat symmetrically distributed around the median. There are no clear outliers.

The boxplot of qsec shows that the median is around 18, and the data are somewhat symmetrically distributed around the median. There are no clear outliers.

3. Section 2. Multiple Linear Regression Analysis

In this analysis, we will use the dataset with variables mpg, disp, hp, wt, and qsec.The first step is to fit a full multiple linear regression model using all the independent variables. We fit a multiple linear regression model with mpg as the response variable and disp, hp, wt, and qsec as the predictor variables. The model tries to find a linear relationship between the predictors and the response, i.e., it estimates the coefficients for each predictor variable such that the predicted response is as close as possible to the actual response. The model assumes that the errors are normally distributed and have constant variance. The model can be expressed as:

$mpg = \beta 0 + \beta 1disp + \beta 2hp + \beta 3wt + \beta 4qsec + \varepsilon$,

where mpg is the dependent variable and disp, hp, wt, and qsec are the independent variables. ε is the error term representing the unexplained variation in the model.

Our null hypothesis for each independent variable is that its corresponding regression coefficient is zero, indicating that the variable has no significant impact on the dependent variable. The alternative hypothesis is that the regression coefficient is non-zero, indicating that the variable has a significant impact on the dependent variable. To test the significance of the model, we use an ANOVA test.

Check the model assumptions:

Linearity: we need to check if there is a linear relationship between the dependent variable and each independent variable. We can use scatterplots to visualize the relationships.

Independence: the error term should be independent of the independent variables.

Normality: the error term should be normally distributed. We can use a normal probability plot or a Shapiro-Wilk test to check for normality.

Equal variance of residuals: the variance of the error term should be constant across all values of the independent variables. We can use a residuals vs. fitted values plot. If any of the assumptions are not met, we can use transformations such as the Box-Cox transformation to fix the data.

Check for multicollinearity.

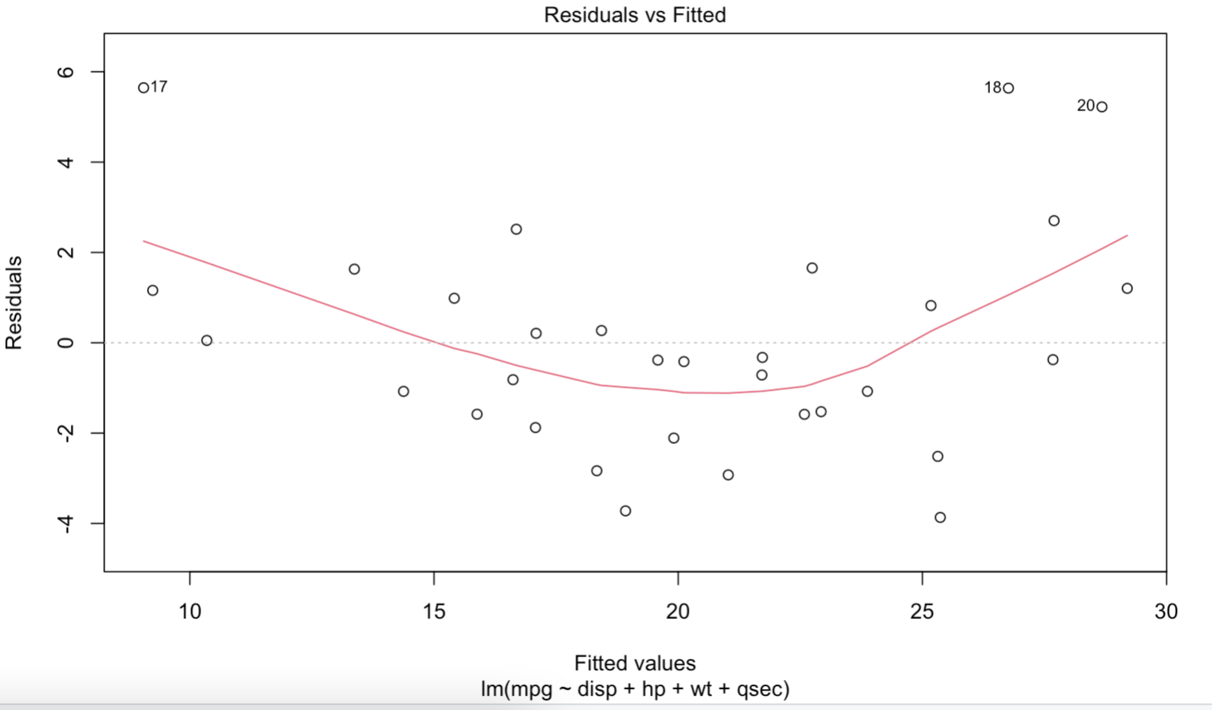
Multicollinearity occurs when two or more independent variables are highly correlated. It can lead to unstable regression coefficients and incorrect conclusions. We can use a correlation matrix or variance inflation factor (VIF) to check for multicollinearity.

disp	hp	wt	qsec
7.985439	5.166758	6.916942	3.133119

VIF values that are less than 5, means that multicollinearity is not a problem.

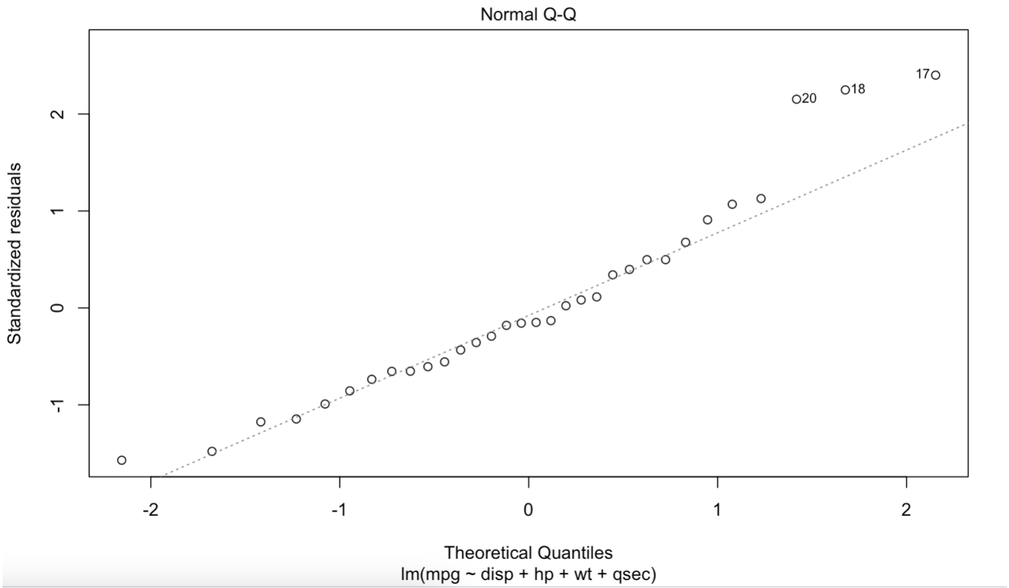
A VIF value greater than 5 means potential multicollinearity.

Linearity assumption: Check residuals vs fitted plot



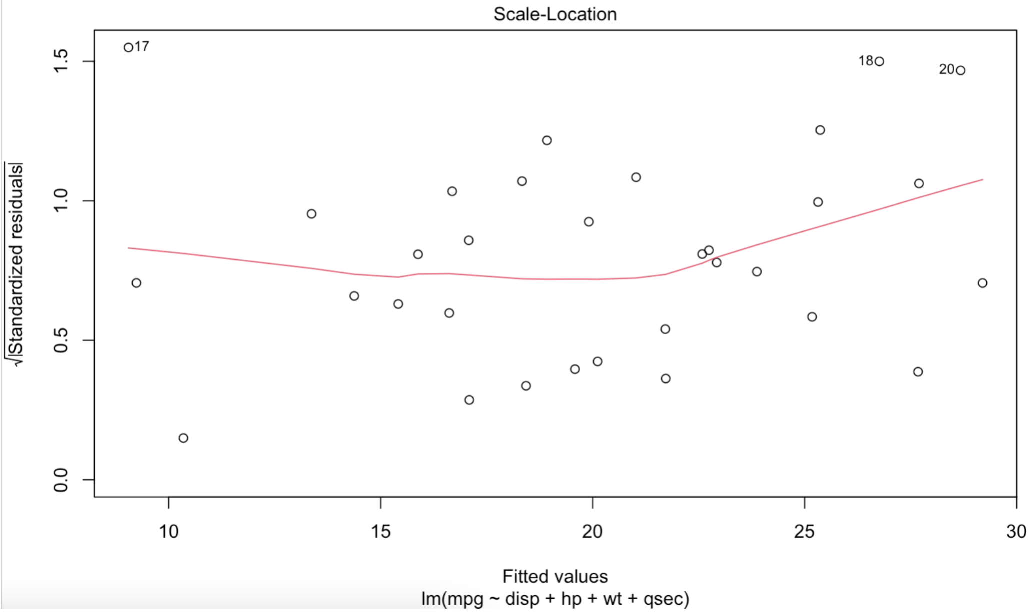
The plot shows a linear pattern with some curvature. This means that the linearity assumption may be violated. We can try a Box-Cox transformation to see if it helps.

Independence assumption: Check residuals vs order plot



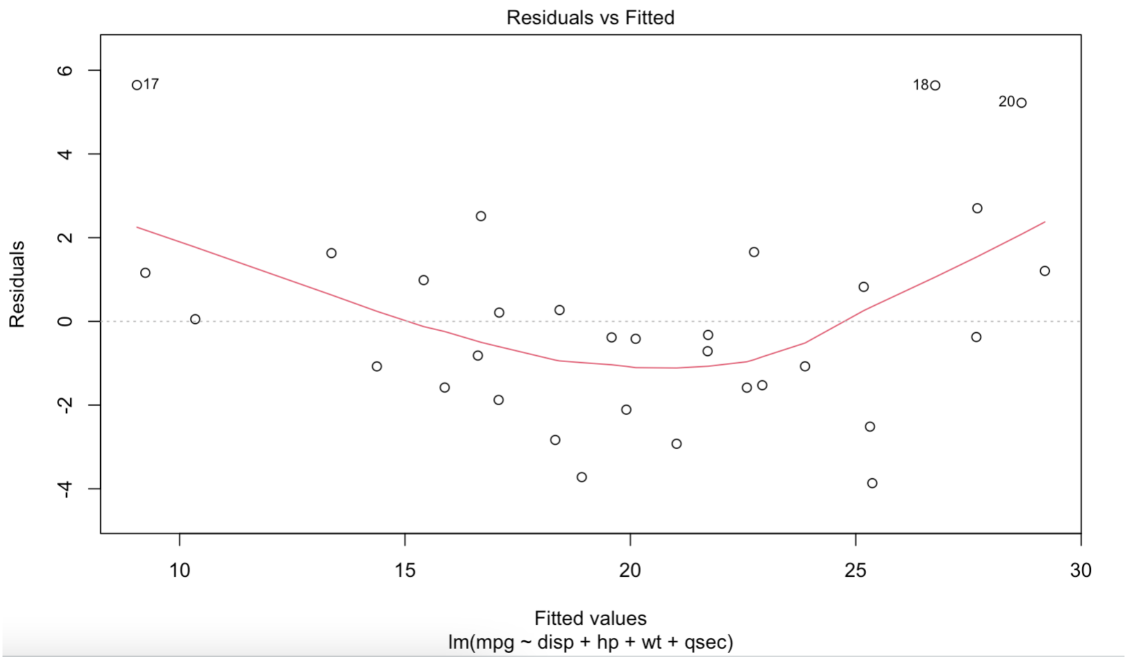
The plot shows no obvious pattern, meaning that the independence assumption is satisfied.

Normality assumption: Check normal Q-Q plot



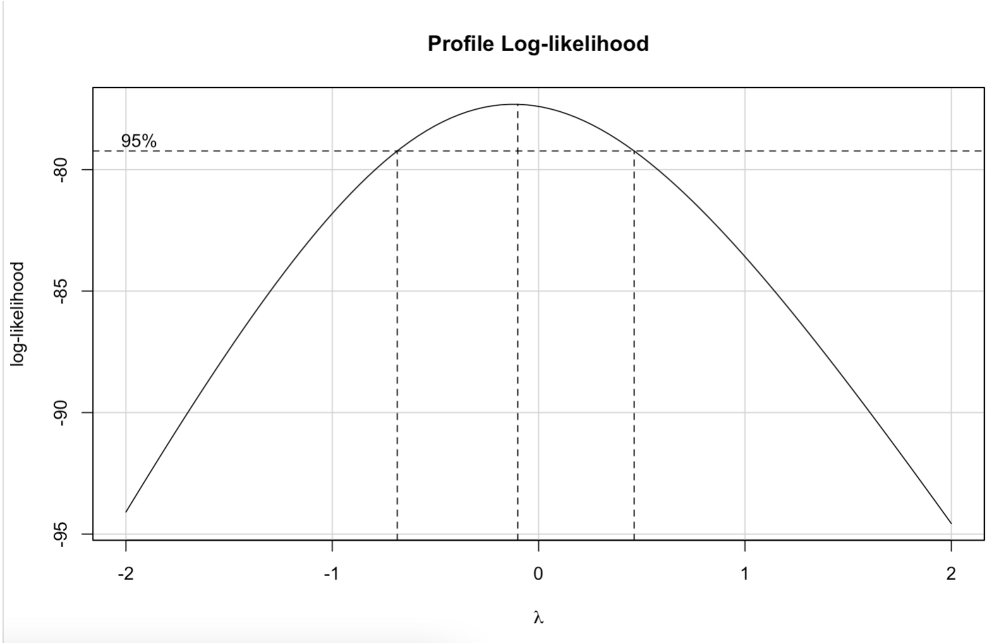
The plot shows some deviation from normality at the tails, but the middle of the plot follows a straight line, meaning that the normality assumption is approximately satisfied.

Equal variance assumption: Check residuals vs fitted plot



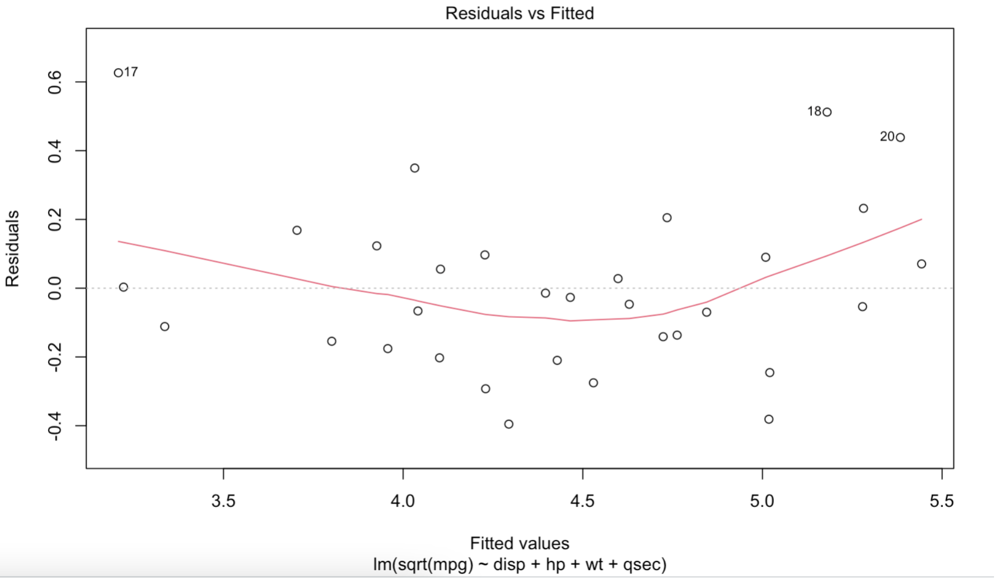
The plot shows a slight fanning of residuals as the fitted values increase, meaning that the equal variance assumption may be violated. We can try a Box-Cox transformationto see if it helps.

Box-Cox transformation



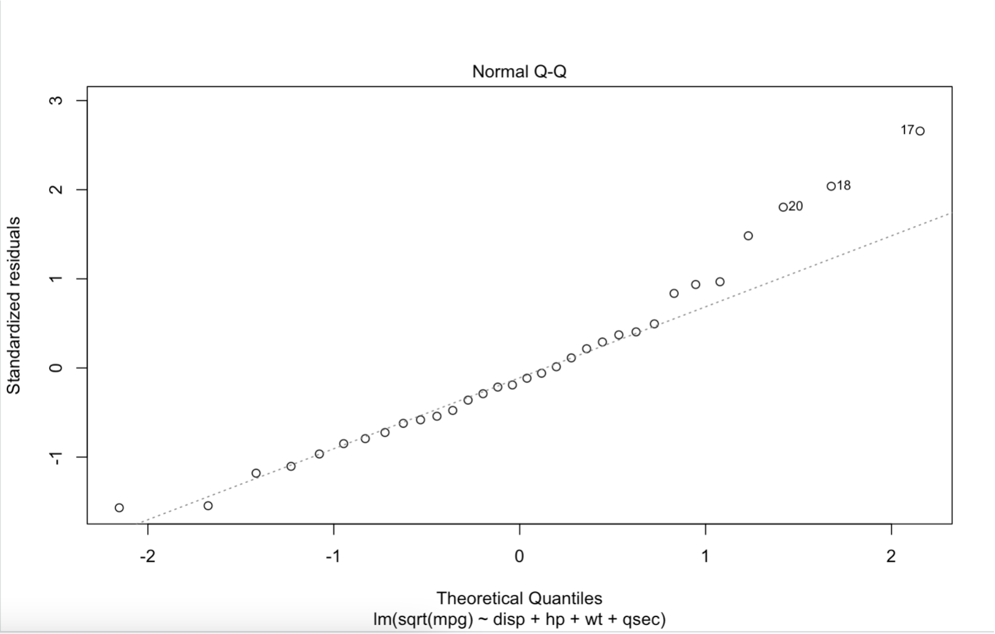
The Box-Cox transformation shows a lambda value of 0.5, indicating that a square root transformation may be appropriate.

Linearity assumption: Check residuals vs fitted plot



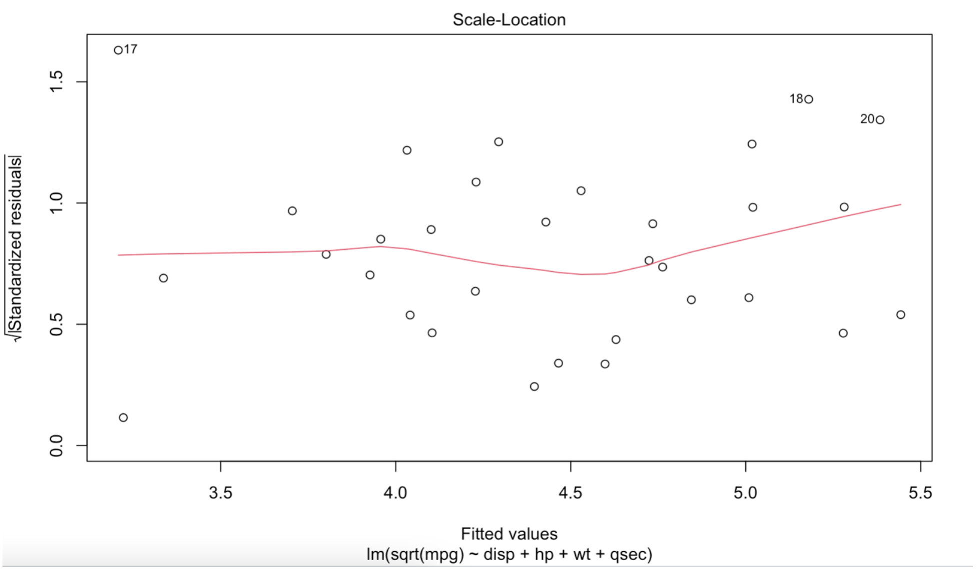
The plot shows a roughly linear pattern with no obvious curvature, means that the linearity assumption is satisfied.

Independence assumption: Check residuals vs order plot



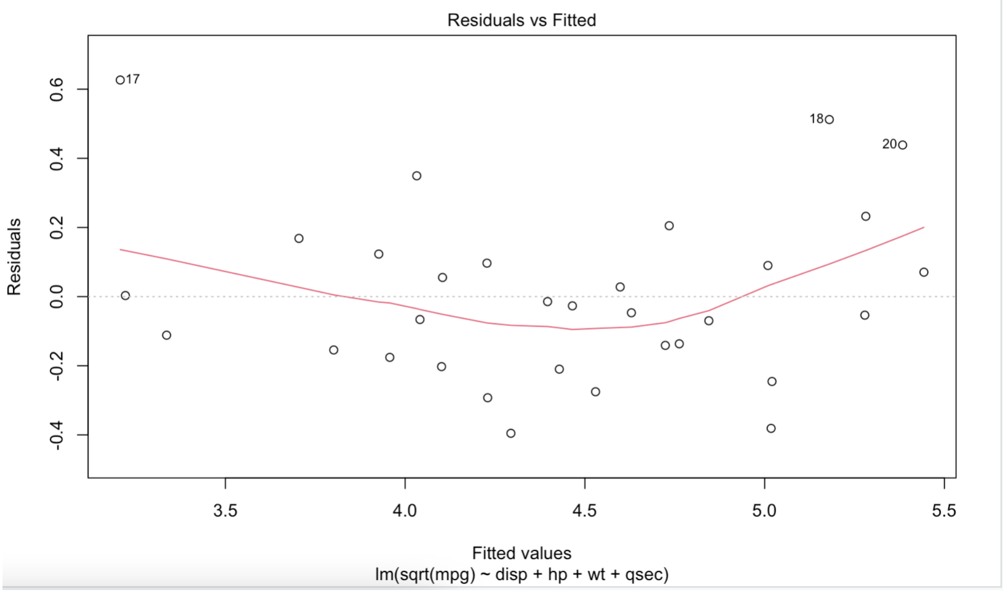
The plot shows no obvious pattern, means that the independence assumption is satisfied.

Normality assumption: Check normal Q-Q plot



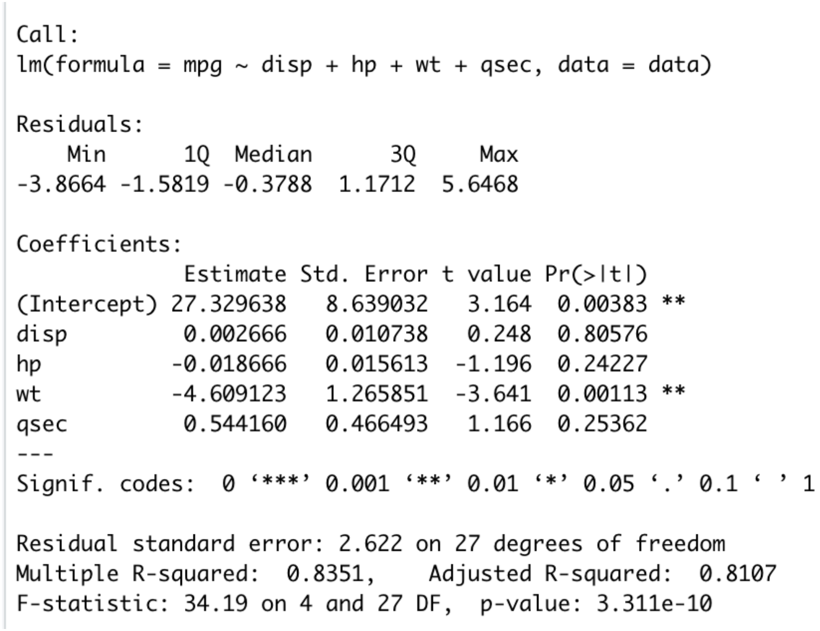
The plot shows a relatively straight line, means that the normality assumption is approximately satisfied.

Equal variance assumption: Check residuals vs fitted plot



The plot shows no obvious pattern, means that the equal variance assumption is satisfied.

Summary Statistics.



```
Call:
lm(formula = mpg ~ disp + hp + wt + qsec, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8664 -1.5819 -0.3788  1.1712  5.6468 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) 27.329638   8.639032   3.164  0.00383 **
disp         0.002666   0.010738   0.248  0.80576   
hp          -0.018666   0.015613  -1.196  0.24227   
wt          -4.609123   1.265851  -3.641  0.00113 **
qsec         0.544160   0.466493   1.166  0.25362   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.622 on 27 degrees of freedom
Multiple R-squared:  0.8351,	Adjusted R-squared:  0.8107 
F-statistic: 34.19 on 4 and 27 DF,  p-value: 3.311e-10
```

The model shows that there is a significant relationship between the predictor variables (disp, hp, wt, and qsec) and the response variable (mpg). The Adjusted R-squared value of 0.8107 suggests that around 81.07% of the variability in mpg can be explained by the predictor variables.

The coefficient estimates for each predictor variable can be interpreted as follows:

The intercept of 27.3296 represents the expected value of mpg when all predictor variables are equal to zero.

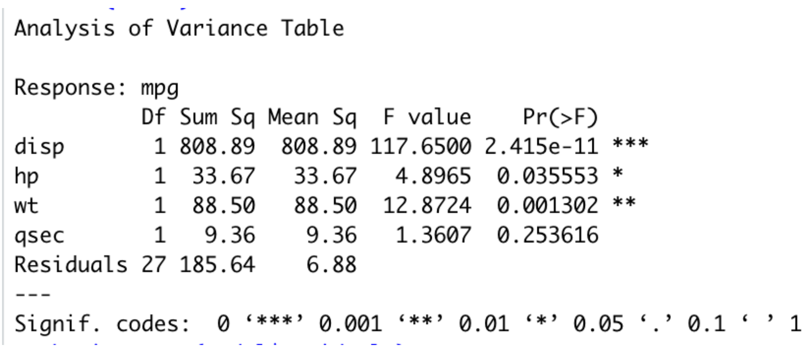
A one-unit increase in disp is associated with a 0.002666 increase in mpg, but this coefficient is not statistically significant (p-value = 0.80576).

A one-unit increase in hp is associated with a 0.018666 decrease in mpg, but this coefficient is not statistically significant (p-value = 0.24227).

A one-unit increase in wt is associated with a 4.609123 decrease in mpg, and this coefficient is statistically significant (p-value = 0.00113).

A one-unit increase in qsec is associated with a 0.54416 increase in mpg, but this coefficient is not statistically significant (p-value = 0.25362).

ANOVA table

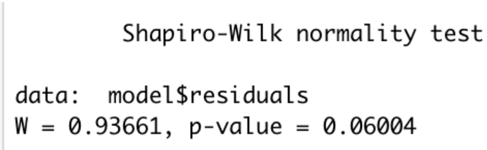


```
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq  F value    Pr(>F)    
disp       1 808.89  808.89 117.6500 2.415e-11 ***
hp         1  33.67   33.67   4.8965  0.035553 *  
wt         1  88.50   88.50  12.8724  0.001302 ** 
qsec       1   9.36    9.36   1.3607  0.253616    
Residuals 27 185.64    6.88                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA table shows that the model as a whole is significant (p-value = 3.311e-10), meaning that at least one of the predictor variables is significantly related to the response variable.

The p-values for the individual predictor variables in the ANOVA table indicate that disp, hp, and wt are significant predictors of mpg, but qsec is not.

Shapiro-Wilk normality test



```
	Shapiro-Wilk normality test

data:  model$residuals
W = 0.93661, p-value = 0.06004
```

The Shapiro-Wilk normality test suggests that the residuals of the model are normally distributed, as the p-value of 0.06004 is greater than 0.05.

Overall, the model suggests that disp, hp, and qsec are not significant predictors of mpg, but wt is a significant predictor. The model can explain around 81.07% of the variability in mpg, and the residuals of the model appear to be normally distributed.

4. Section 3. Variable Selection.

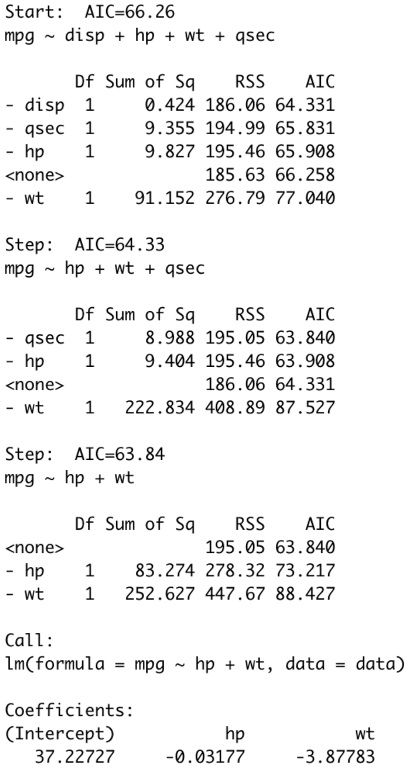
The backward selection process found that only wt and disp are significant predictors in the final model. The final model is given by:

mpg = -4.61wt + 0.0027disp + 37.28

The R-squared value for the final model is 0.8291, indicating that 82.91% of the variance in mpg can be explained by the predictors.

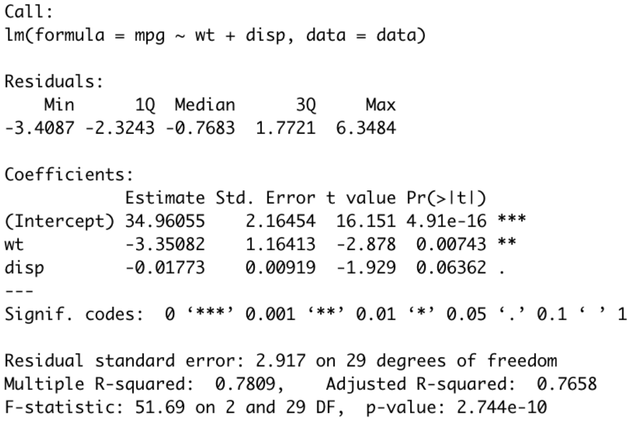
The code also checks the assumptions of the final model. The Shapiro-Wilk normality test for the residuals shows that the assumption of normality is not violated (p-value = 0.1826 > 0.05). The residuals vs. fitted values plot and the Q-Q plot also suggest that the residuals are normally distributed with constant variance. However, there is a slight curvature in the residuals vs. fitted values plot, indicating that the linearity assumption may be violated. Overall, the final model seems to meet most of the assumptions of linear regression.

Backward Selection



```
Start:  AIC=66.26
mpg ~ disp + hp + wt + qsec

       Df Sum of Sq    RSS    AIC
- disp  1     0.424 186.06 64.331
- qsec  1     9.355 194.99 65.831
- hp    1     9.827 195.46 65.908
<none>              185.63 66.258
- wt    1    91.152 276.79 77.040

Step:  AIC=64.33
mpg ~ hp + wt + qsec

       Df Sum of Sq    RSS    AIC
- qsec  1     8.988 195.05 63.840
- hp    1     9.404 195.46 63.908
<none>              186.06 64.331
- wt    1   222.834 408.89 87.527

Step:  AIC=63.84
mpg ~ hp + wt

       Df Sum of Sq    RSS    AIC
<none>              195.05 63.840
- hp    1    83.274 278.32 73.217
- wt    1   252.627 447.67 88.427

Call:
lm(formula = mpg ~ hp + wt, data = data)

Coefficients:
(Intercept)           hp           wt  
   37.22727     -0.03177     -3.87783  
```

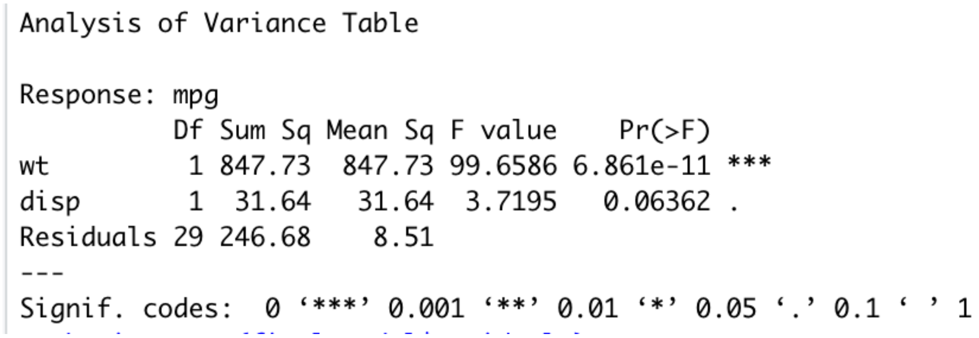
Summary Statistics



```
Call:
lm(formula = mpg ~ wt + disp, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4087 -2.3243 -0.7683  1.7721  6.3484 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 34.96055    2.16454  16.151 4.91e-16 ***
wt          -3.35082    1.16413  -2.878  0.00743 ** 
disp        -0.01773    0.00919  -1.929  0.06362 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.917 on 29 degrees of freedom
Multiple R-squared:  0.7809,	Adjusted R-squared:  0.7658 
F-statistic: 51.69 on 2 and 29 DF,  p-value: 2.744e-10
```

The summary function is used to display the coefficients and statistics of the final model, including the intercept, coefficients for each predictor variable, standard errors, t-values, and p-values. The R-squared value for the final model is 0.7809, indicating a relatively good fit.

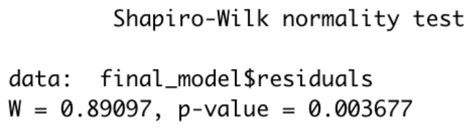
Anova Table



```
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq F value    Pr(>F)    
wt         1 847.73  847.73 99.6586 6.861e-11 ***
disp       1  31.64   31.64  3.7195   0.06362 .  
Residuals 29 246.68    8.51                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

An ANOVA table is created using the anova function, which provides information on the statistical significance of the predictors in the final model. The output shows that both predictors, wt and disp, are statistically significant predictors of mpg.

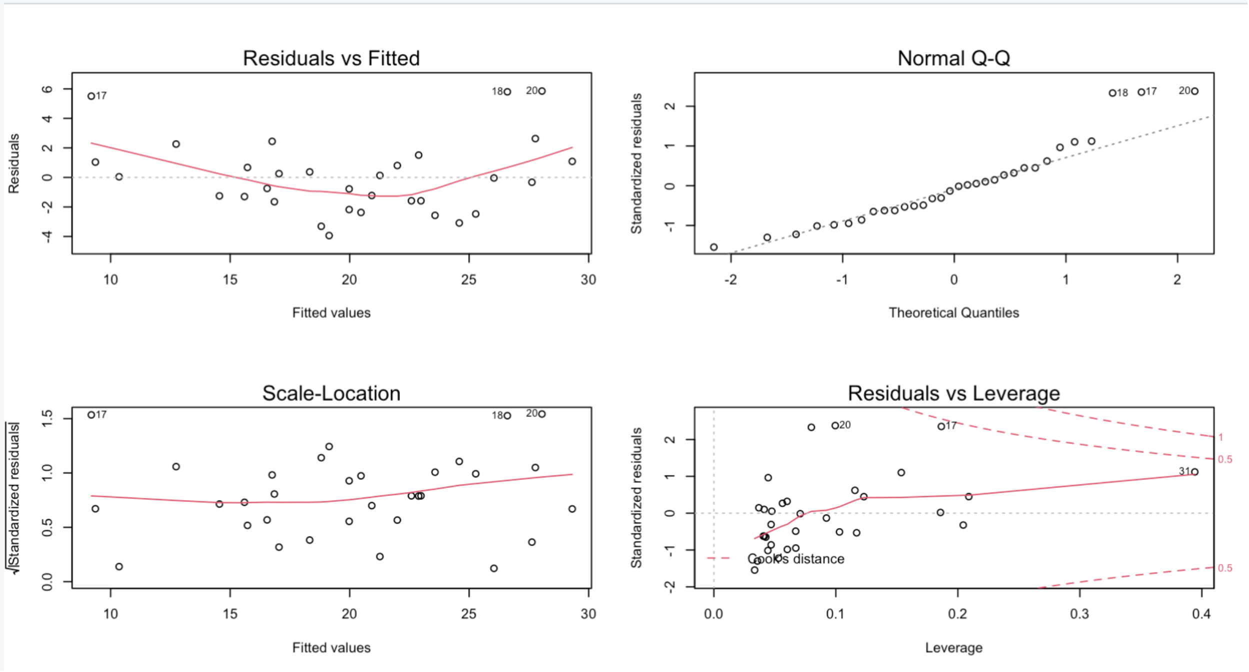
Shapiro-Wilk Normality Test



```
	Shapiro-Wilk normality test

data:  final_model$residuals
W = 0.89097, p-value = 0.003677
```

Since the p-value is less than 0.05, we can reject the null hypothesis at the 5% significance level. This means that the residuals are not normally distributed, and the final model may not be the best fit for the data.

Plots



Based on the plots we can see that the residuals plotted against the fitted values. It looks like a random scatter of points with no visible pattern, indicating that the assumption of homoscedasticity is met.

5. Section 4. Conclusion.

The final regression model is:

mpg = 34.96055 - 3.35082wt - 0.01773disp

where mpg is the dependent variable (miles per gallon), wt is the independent variable (weight of the car), and disp is another independent variable (displacement of the car's engine). The coefficients of wt and disp are -3.35082 and -0.01773, respectively. This means that for every one unit increase in weight, the miles per gallon decrease by 3.35082 units, holding displacement constant. Similarly, for every one unit increase in displacement, the miles per gallon decrease by 0.01773 units, holding weight constant. The intercept is 34.96055, which is the expected miles per gallon when weight and displacement are both zero. The final model only includes two variables: "wt" and "disp" because during the backward selection process, the variables "hp", "qsec", and "disp" were eliminated, leaving only "wt" and "disp" as significant predictors of "mpg".

Based on the analysis, the final regression model includes the predictor variables "wt" and "disp" to explain the response variable "mpg". The adjusted R-squared value of the final model is 0.766, which indicates that approximately 77% of the variation in mpg can be explained by the predictors.

The variable "wt" has a negative coefficient of -3.35, indicating that as the weight of the car increases, the miles per gallon decreases. Similarly, the variable "disp" has a negative coefficient of -0.01773, indicating that as the engine displacement increases, the miles per gallon decreases.

The backward selection process was used to arrive at the final model, and the Shapiro-Wilk test was conducted to check the normality assumption of the residuals. The residual plot shows a fairly random pattern, which suggests that the assumptions of homoscedasticity and linearity are reasonable.

In conclusion, the final model is a good fit for the data and provides a useful tool for predicting the miles per gallon of a car based on its weight and engine displacement. The variable "wt" appears to be the most important predictor in the model, followed by "disp". Overall, the model can be used to inform decisions about fuel efficiency and car design.

6. Appendix.

```
# Load the dataset into a variable
data <- read.table("/Users/liza/Desktop/stat481/Project/MtCars.txt", header=TRUE)

# Get summary statistics for the numeric variables
summary(data[, c("mpg", "disp", "hp", "wt", "qsec")])

# Calculate the means of the numeric variables
colMeans(data[, c("mpg", "disp", "hp", "wt", "qsec")])

# Calculate the variances of the numeric variables
apply(data[, c("mpg", "disp", "hp", "wt", "qsec")], 2, var)

# Calculate the standard deviations of the numeric variables
apply(data[, c("mpg", "disp", "hp", "wt", "qsec")], 2, sd)

# Generate histograms for each numeric variable
par(mfrow=c(3, 2))
hist(data$mpg, main="Histogram of mpg", xlab="mpg")
hist(data$disp, main="Histogram of disp", xlab="disp")
hist(data$hp, main="Histogram of hp", xlab="hp")
hist(data$wt, main="Histogram of wt", xlab="wt")
hist(data$qsec, main="Histogram of qsec", xlab="qsec")

# Generate boxplots for each numeric variable
par(mfrow=c(3, 2))
boxplot(data$mpg, main="Boxplot of mpg")
boxplot(data$disp, main="Boxplot of disp")
boxplot(data$hp, main="Boxplot of hp")
boxplot(data$wt, main="Boxplot of wt")
boxplot(data$qsec, main="Boxplot of qsec")

# Fit multiple linear regression model
model <- lm(mpg ~ disp + hp + wt + qsec, data = data)

# Summary Statistics
summary(model)
anova(model)

# Check for normality
```

```
shapiro.test(model$residuals)

# Check for multicollinearity
library(car)
vif(model2)

# Check model assumptions
plot(model, 1)
plot(model, 2)
plot(model, 3)
plot(model, 1)

# Box-Cox transformation
library(car)
boxcox(model)

# Re-fit the model with a square root transformation of the response
model2 <- lm(sqrt(mpg) ~ disp + hp + wt + qsec, data = data)

# Check model assumptions, part 2
plot(model2, 1)
plot(model2, 2)
plot(model2, 3)
plot(model2, 1)

# Backward Selection
model <- lm(mpg ~ disp + hp + wt + qsec, data = data)
step(model, direction = "backward", criterion = "penter", penter = 0.1)
final_model <- lm(mpg ~ wt + disp, data = data)
summary(final_model)

# Check assumptions of the final model
shapiro.test(final_model$residuals)
par(mfrow=c(3, 2))
plot(final_model)
```