Project 1 - Multi Linear Regression

University of Massachusetts - Dartmouth

MTH 522- Project 1

ABSTRACT

This report aims to analyze the numerical and graphical variations among variables including displacement, horsepower, weight, acceleration, and mpg using an automobile dataset. Additionally, we will present comparison plots to explore the interrelationship between these variables.

THE ISSUES

This report utilizes multiple linear regression on the Auto data set. It involves creating a scatterplot matrix of all variables, computing the matrix of correlations, and performing a regression analysis using lm() function. The summary() function is used to display the results, and diagnostic plots are produced using the plot() function. The model is also tested with interaction effects and different variable transformations. The usefulness of the predictors in explaining the response and the accuracy of the model's predictions are evaluated.

DISCUSSION

The summary of the multiple linear regression model on the Auto data set indicates that only the variables of horsepower and weight are statistically significant predictors of mpg. Displacement and acceleration were found to be not significant. The adjusted R-squared value of the initial model was 0.7197, indicating a moderate fit. However, the adjusted R-squared value increased to 0.7207 when the model was modified to include only horsepower and weight as predictors.

This new model can be used to predict the fuel efficiency (mpg) given predictor values for horsepower and weight. It is important to note that other factors, such as the model year, could also be significant predictors but were not included in this analysis. Further exploration and refinement of the model could be done to improve its predictive accuracy.

APPENDIX A: METHOD

multiple linear regression analysis on automobile data using R.

The code read in data from an Excel file, omit any missing data points, and assign relevant columns to variables. The variables used in the analysis are displacement, horsepower, weight, acceleration, and miles per gallon.

The summary(), pairs(), and cor() functions are used to summarize and explore the data.

The lm() function is then used to fit a multiple linear regression model of mpg as a function of the other variables. The model is then plotted using plot().

The resid() function is used to calculate the residuals of the model, which are then plotted using plot().

The summary() function is used to test whether at least one of the predictors is useful, and confint() function is used to calculate the confidence interval of the model.

The predict() function is used to predict the response (mpg) for a new set of predictor values, and print() is used to display the predicted value. The predict() function is also used to calculate the prediction interval for the new set of predictor values.

The plot() function is used to plot hat values and Cook's distance, which are measures of influential observations in the data.

Finally, a three-predictor model is created using lm(), and the significance of the third predictor is tested using summary(). The models with and without the third predictor are compared using anova(), and the summary() function is used to calculate and compare the R-squared values of the two models

##APPENDIX B: CODE AND RESULTS

```
library("readxl")
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Read data and remove missing values
Auto <- read_excel("C:/Users/budap/Downloads/auto_data_budapaneti_usha.xls")
Auto <- na.omit(Auto)

# Define variables
disp <- Auto$`displacement`
hp <- Auto$`horsepower`
weg <- Auto$`weight`
acc <- Auto$`acceleration`
mpg <- Auto$`mpg`

# Summary of data
## Heading: Summary of Auto Data
```

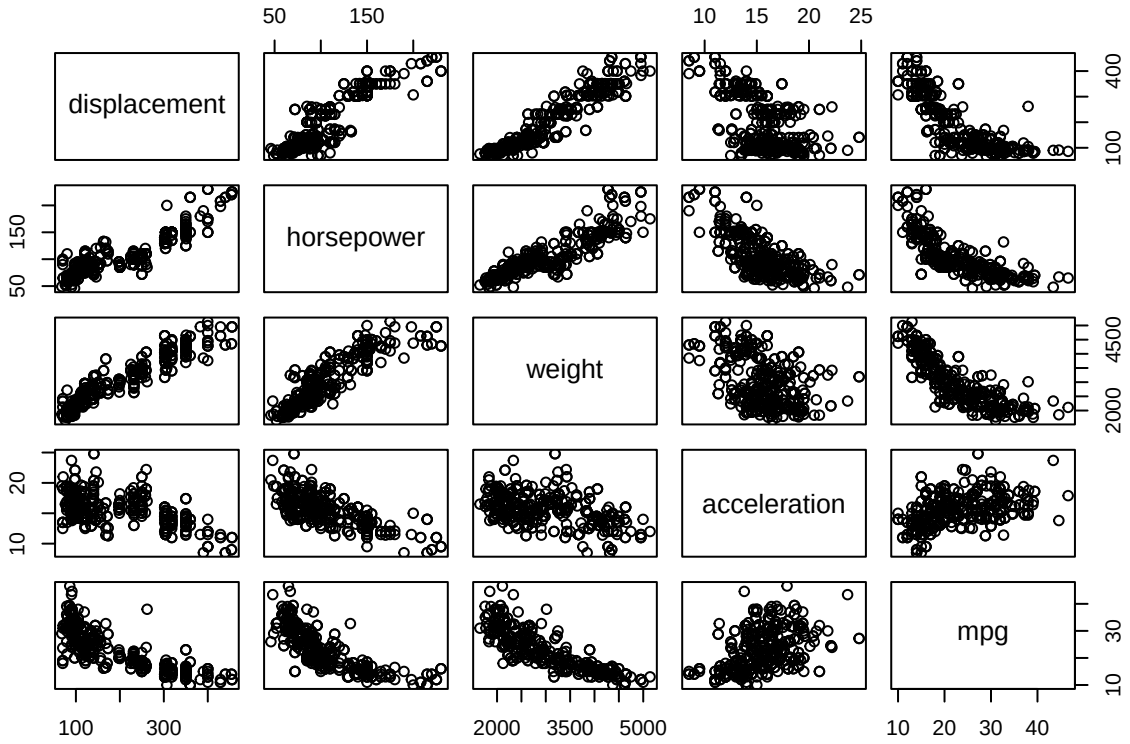
```
summary(Auto)
```

```
##   displacement     horsepower        weight      acceleration        mpg
##  Min.   : 68.0   Min.   : 46.0   Min.   :1649   Min.   : 8.50   Min.   :10.00
##  1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2265   1st Qu.:14.00   1st Qu.:16.05
##  Median :151.0   Median : 95.0   Median :2890   Median :15.50   Median :21.55
##  Mean   :198.7   Mean   :106.2   Mean   :3033   Mean   :15.54   Mean   :22.74
##  3rd Qu.:303.5   3rd Qu.:132.8   3rd Qu.:3830   3rd Qu.:17.00   3rd Qu.:28.10
##  Max.   :455.0   Max.   :230.0   Max.   :5140   Max.   :24.80   Max.   :46.60
```

Relationship between variables

Heading: Correlation and Pairwise Relationships

```
pairs(Auto)
```



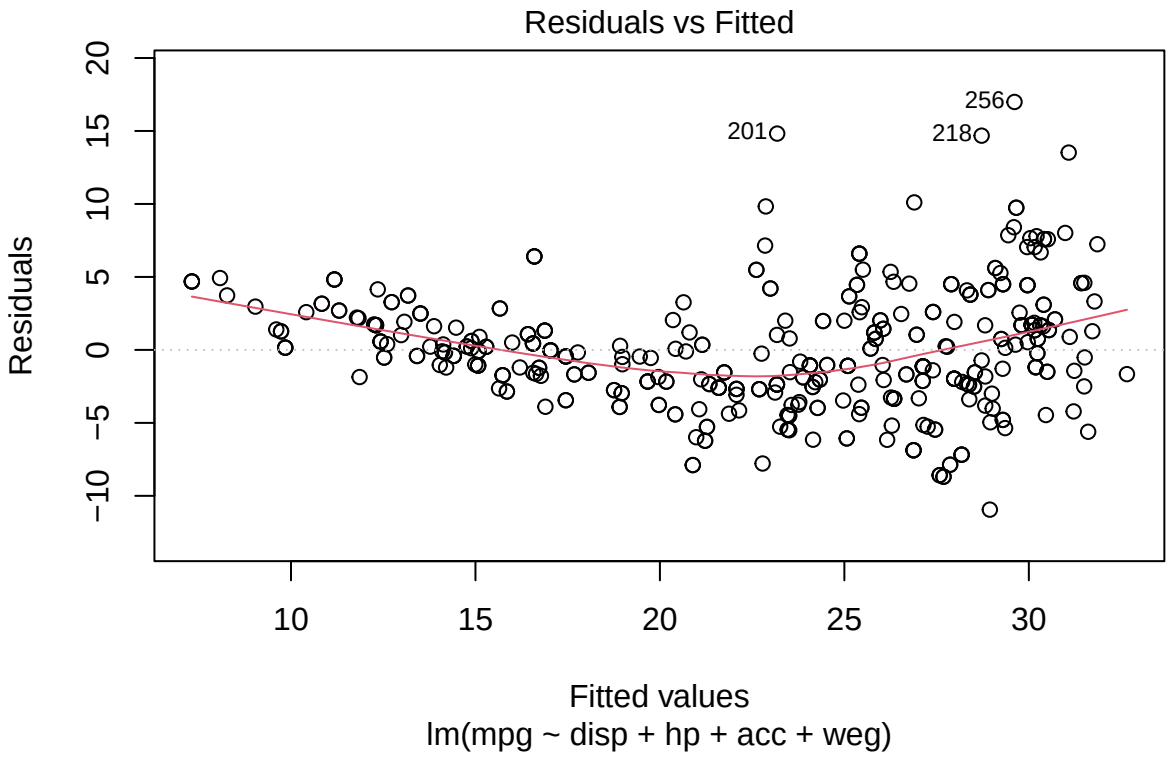

```
cor(Auto)
```

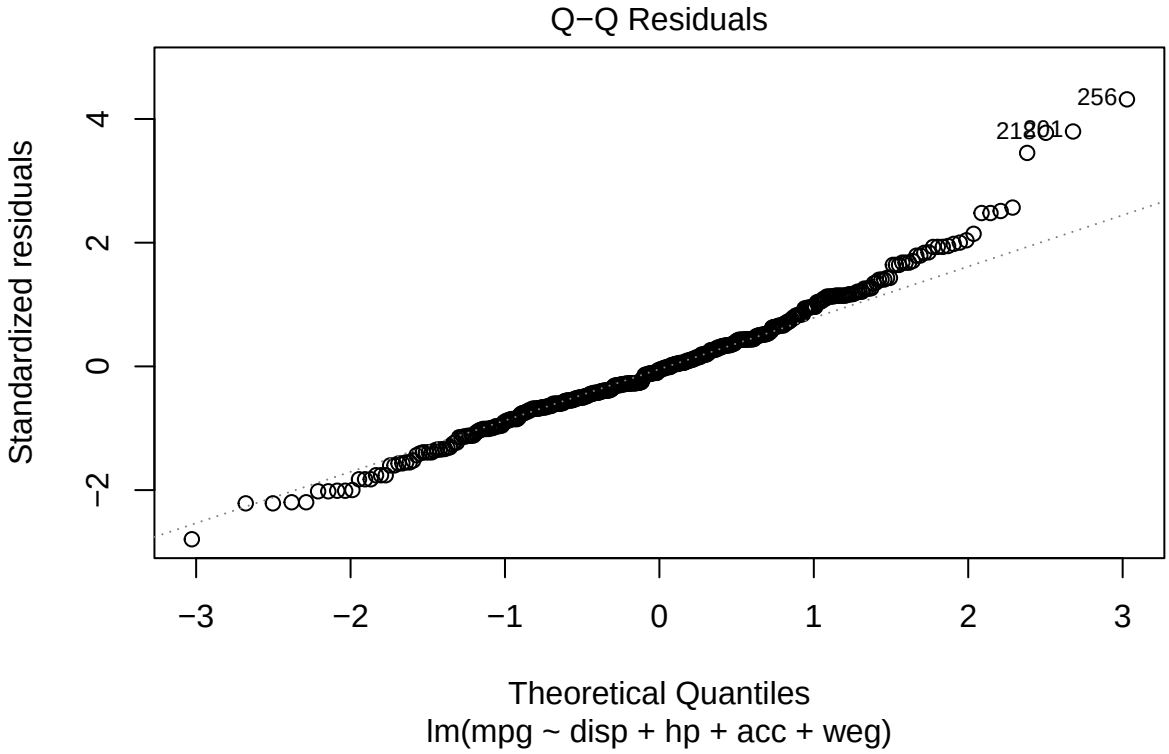
```
##              displacement horsepower     weight acceleration        mpg
## displacement    1.0000000  0.8943033  0.9456484   -0.5481813 -0.8160623
## horsepower      0.8943033  1.0000000  0.8791188   -0.7024110 -0.7821015
## weight          0.9456484  0.8791188  1.0000000   -0.4485435 -0.8459330
## acceleration   -0.5481813 -0.7024110 -0.4485435    1.0000000  0.4303002
## mpg            -0.8160623 -0.7821015 -0.8459330    0.4303002  1.0000000
```

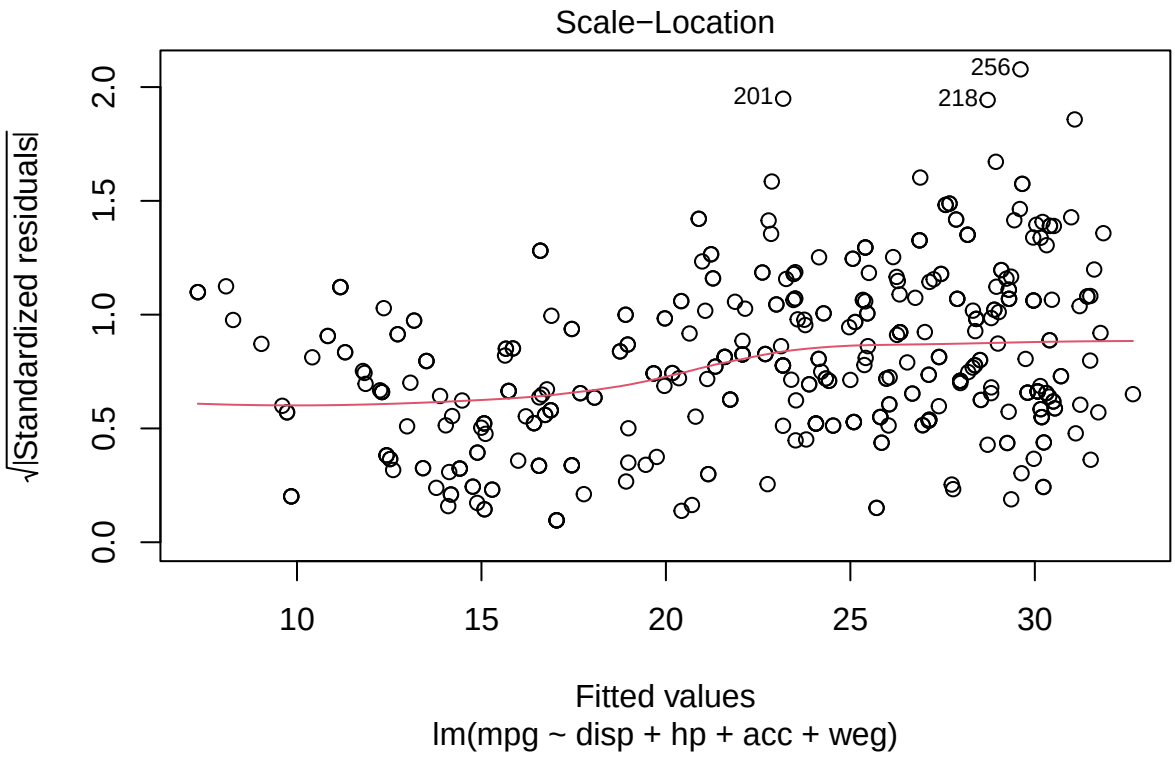
Linear regression

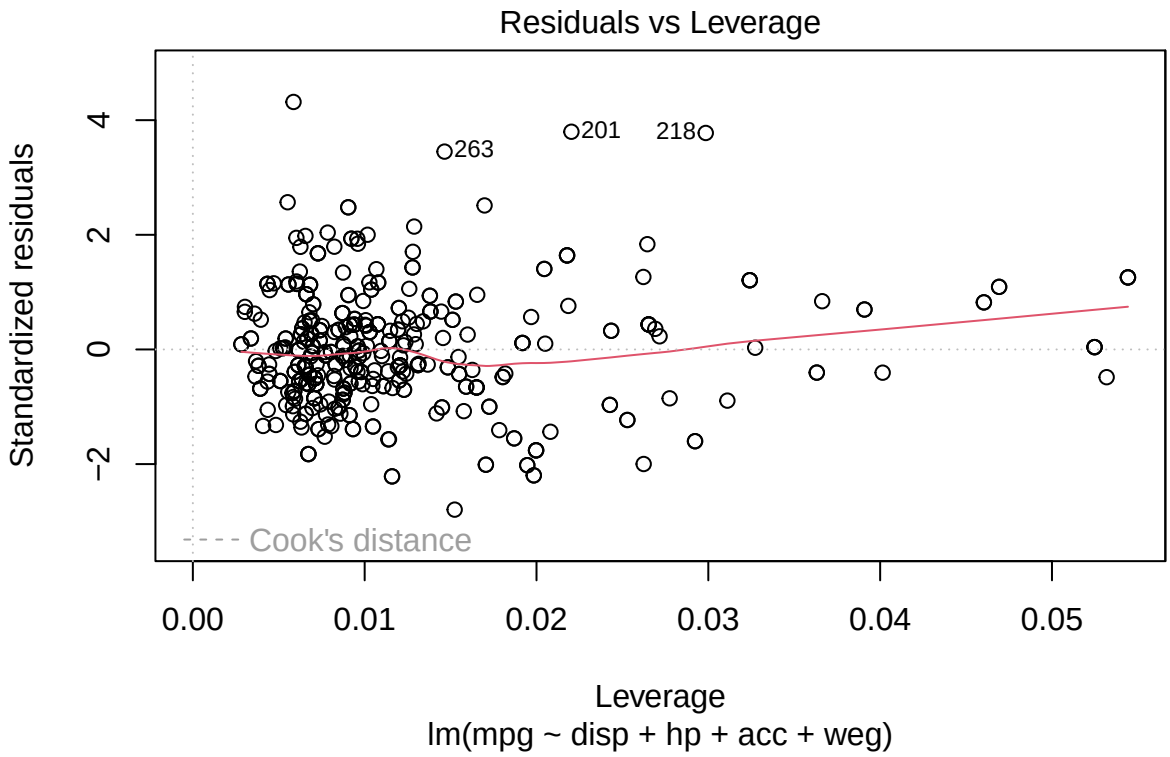
Heading: Linear Regression

```
# Fit a multiple linear regression model
model <- lm(mpg ~ disp + hp + acc + weg, data = Auto)
plot(model)
```

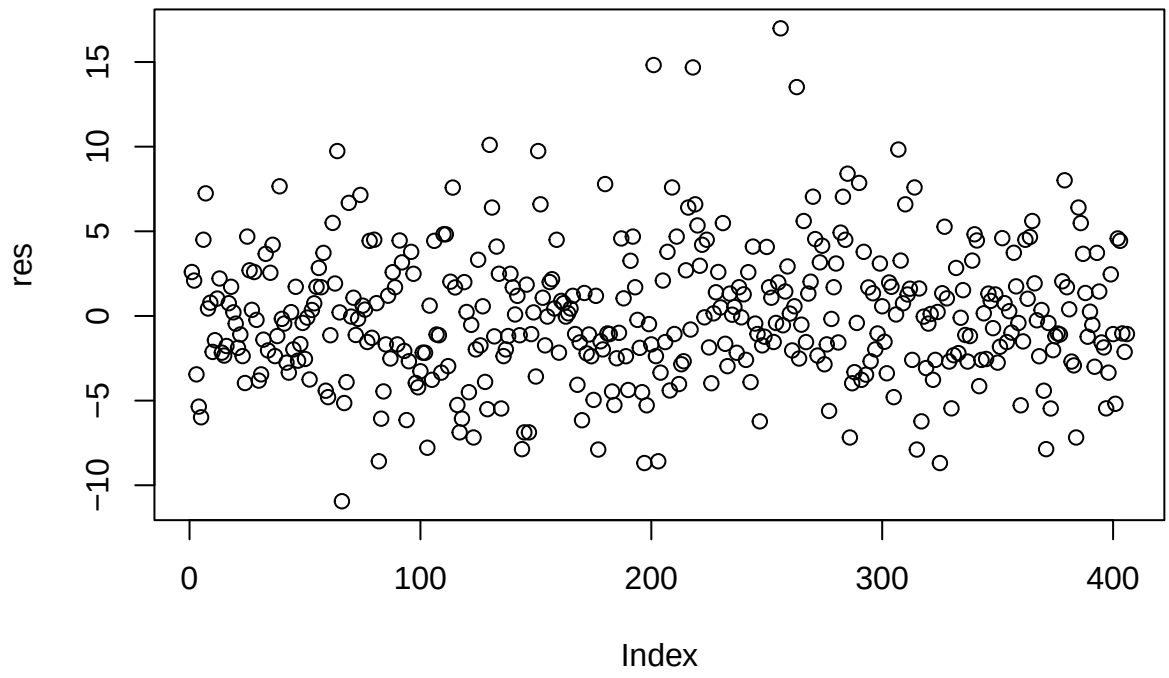

Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ disp + hp + acc + weg)
201
218
256
−10
−5
0
5
10
15
20
10
15
20
25
30


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(mpg ~ disp + hp + acc + weg)
256
−3
−2
−1
0
1
2
3
−2
0
2
4


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
201
218
256
10
15
20
25
30
Fitted values
lm(mpg ~ disp + hp + acc + weg)




```
res <- resid(model)
plot(res)
```



Predictors

Heading: Predictors

```
# Test whether at least one of the predictors is useful
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ disp + hp + acc + weg, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9438  -2.3767  -0.2317   2.0183  16.9881
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 44.2523721  2.3979335  18.454  < 2e-16 ***
## disp        -0.0046594  0.0063780  -0.731   0.4655
## hp          -0.0316353  0.0155734  -2.031   0.0429 *
## acc         -0.0285644  0.1245217  -0.229   0.8187
## weg         -0.0055318  0.0008161  -6.778 4.36e-11 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.946 on 401 degrees of freedom
## Multiple R-squared:  0.7225, Adjusted R-squared:  0.7197
## F-statistic:   261 on 4 and 401 DF,  p-value: < 2.2e-16
```

Confidence interval

Heading: Confidence Interval

```
confint(model)
```

```
##                    2.5 %       97.5 %
## (Intercept) 39.538280728 48.966463385
## disp        -0.017197894  0.007879077
## hp          -0.062250916 -0.001019689
## acc         -0.273361358  0.216232504
## weg         -0.007136223 -0.003927437
```

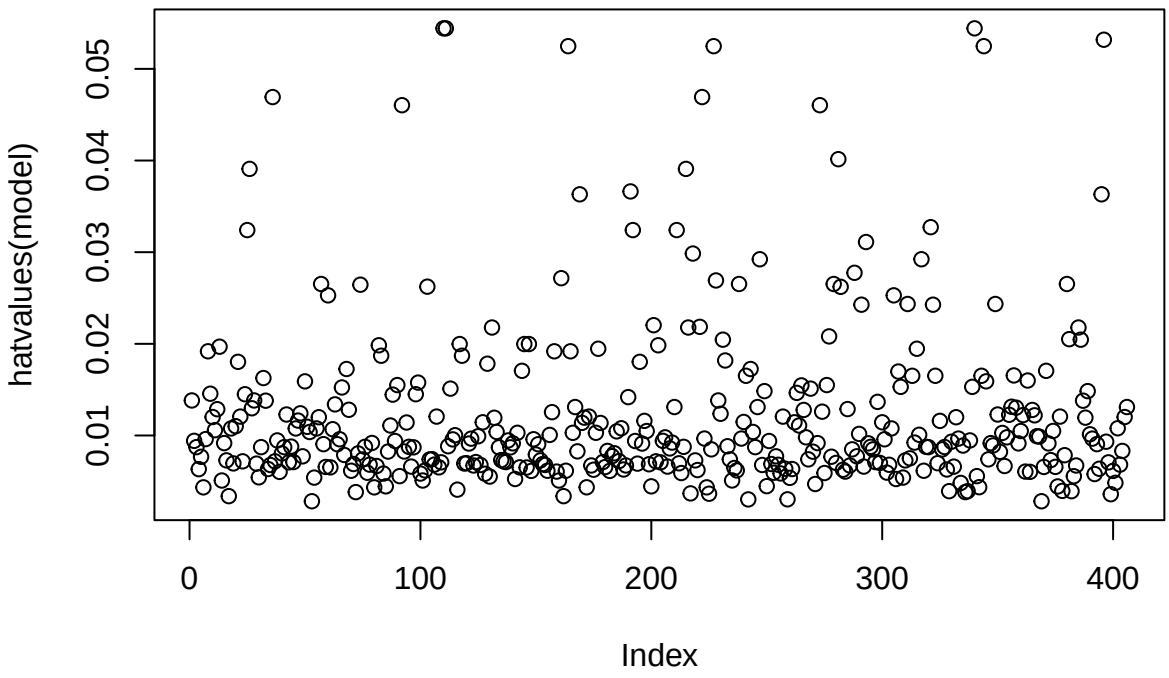
Prediction interval bounds

Heading: Prediction Interval Bounds

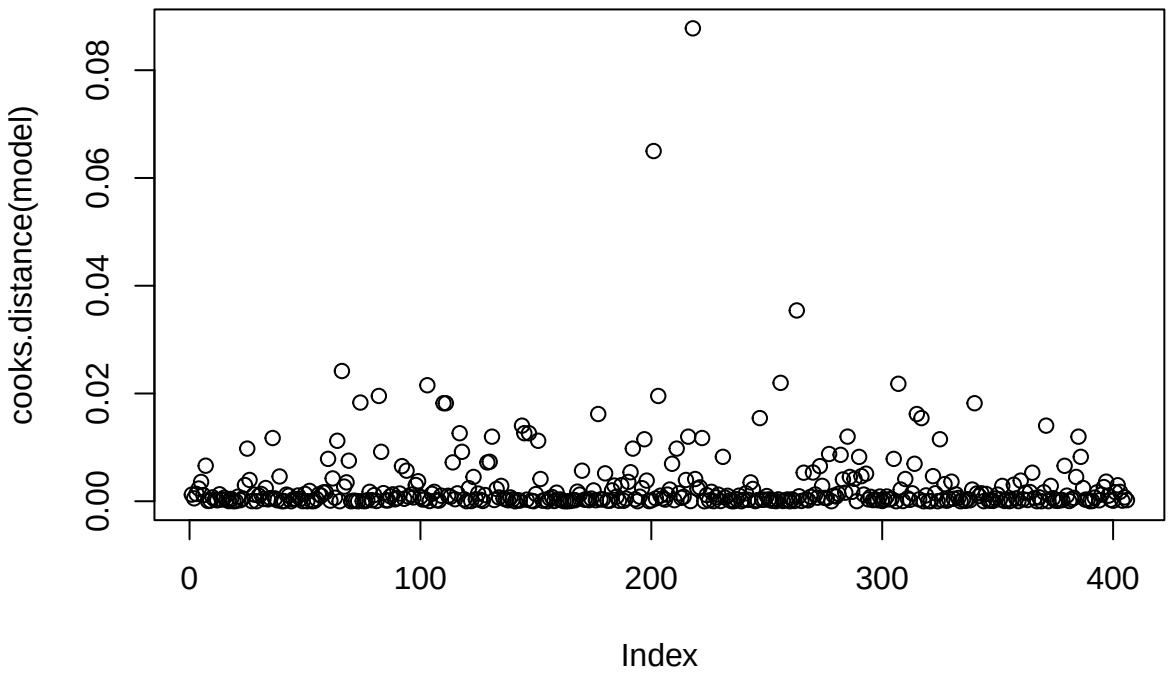
```
# Print the lower and upper bounds of the prediction interval
new_data <- data.frame(disp = 5,hp= 10, acc= 13 ,weg= 15)
prediction_interval <- predict(model, newdata = new_data, interval ="prediction", level = 0.95)
print(prediction_interval[2:4])
```

```
## [1] 35.34803 51.56878       NA
```

```
plot(hatvalues(model))
```




```
plot(cooks.distance(model))
```



Anova

Heading: Anova

```
# Perform an F-test for the overall significance of the regression model
anova(model)
```

```
## Analysis of Variance Table
##
## Response: mpg
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## disp        1 14985.0 14985.0 962.257 < 2.2e-16 ***
## hp          1   307.3   307.3  19.735 1.152e-05 ***
## acc         1   248.9   248.9  15.985 7.600e-05 ***
## weg         1   715.5   715.5  45.945 4.362e-11 ***
## Residuals 401  6244.7    15.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R-squared

Heading: R-Squared

```
summary(model)$r.squared
```

```
## [1] 0.7224765
```

Confidence interval

Heading: Confidence Interval

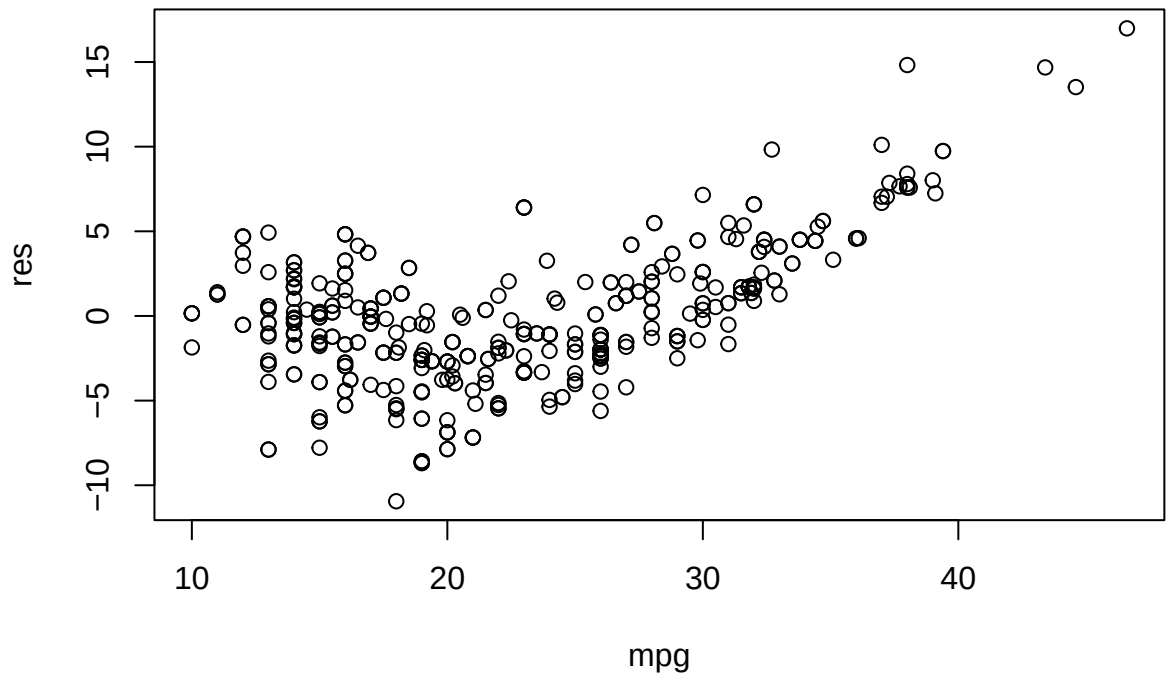
```
confint(model)
```

```
##                    2.5 %       97.5 %
## (Intercept) 39.538280728 48.966463385
## disp        -0.017197894  0.007879077
## hp          -0.062250916 -0.001019689
## acc         -0.273361358  0.216232504
## weg         -0.007136223 -0.003927437
```

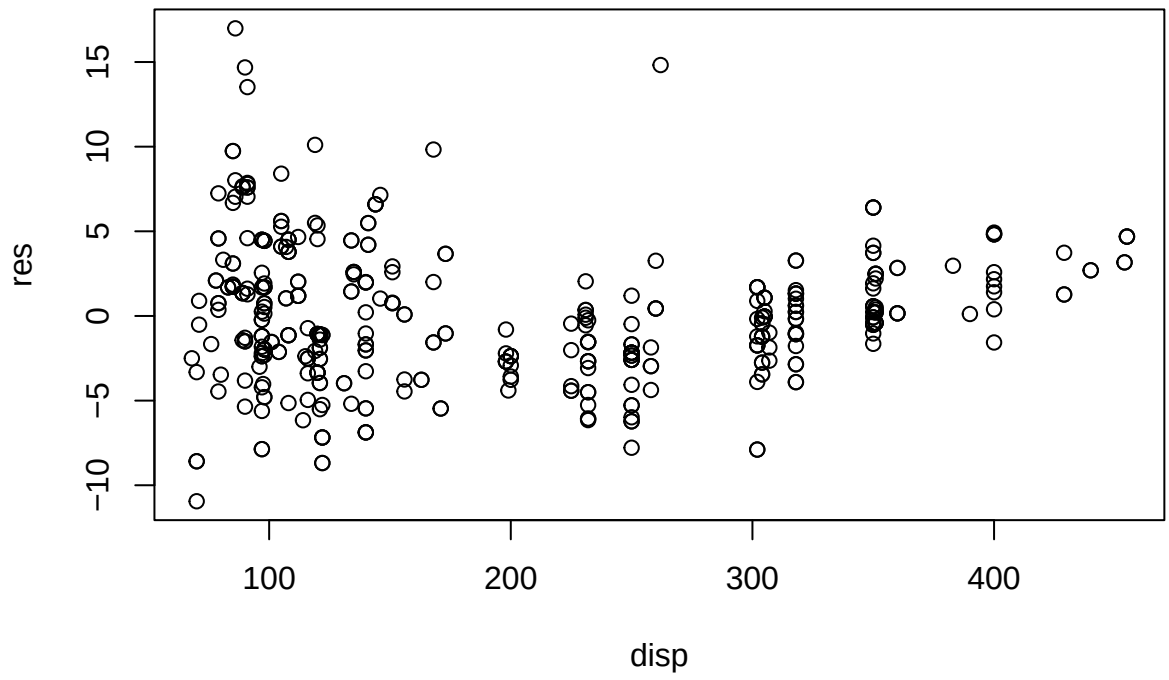
Residual plots

Heading: Residual Plots

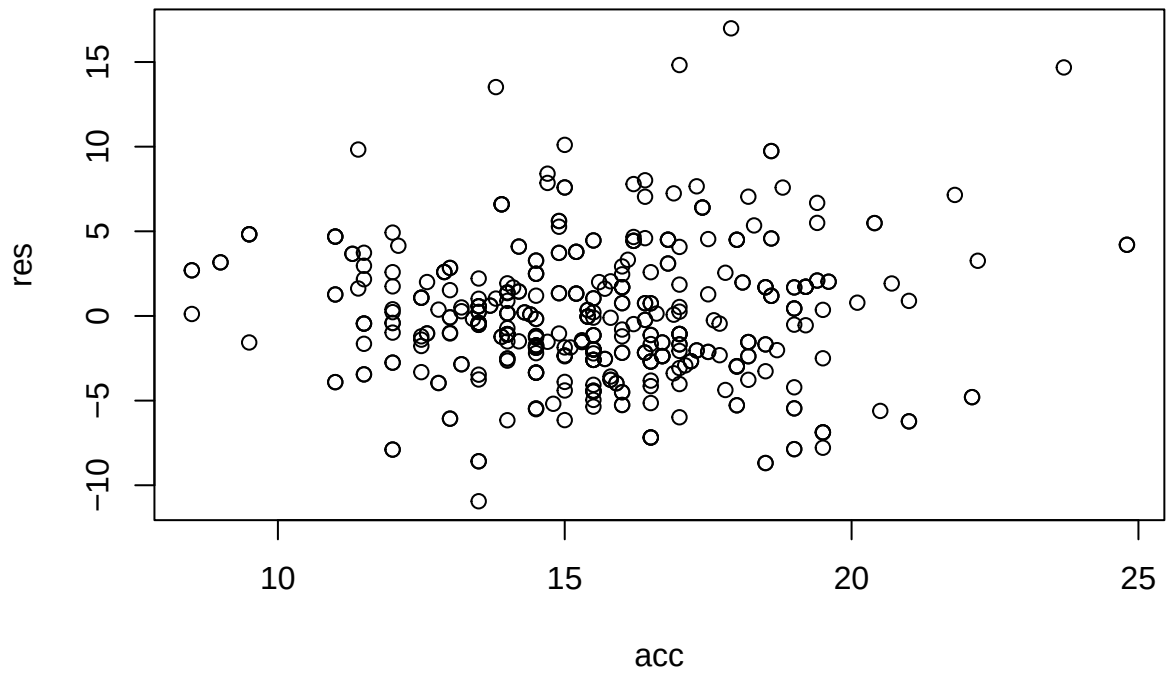
```
# Residual plot against predicted response
plot(mpg, res)
```



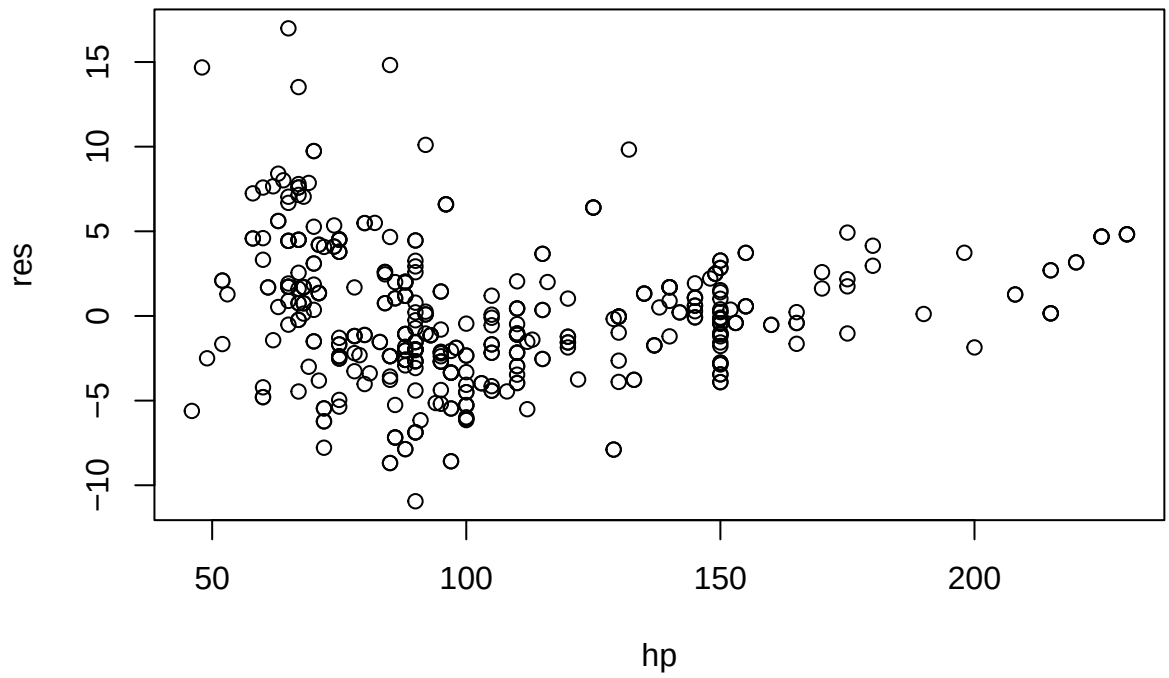

```
# Residual plots against each predictor
plot(disp, res)
```



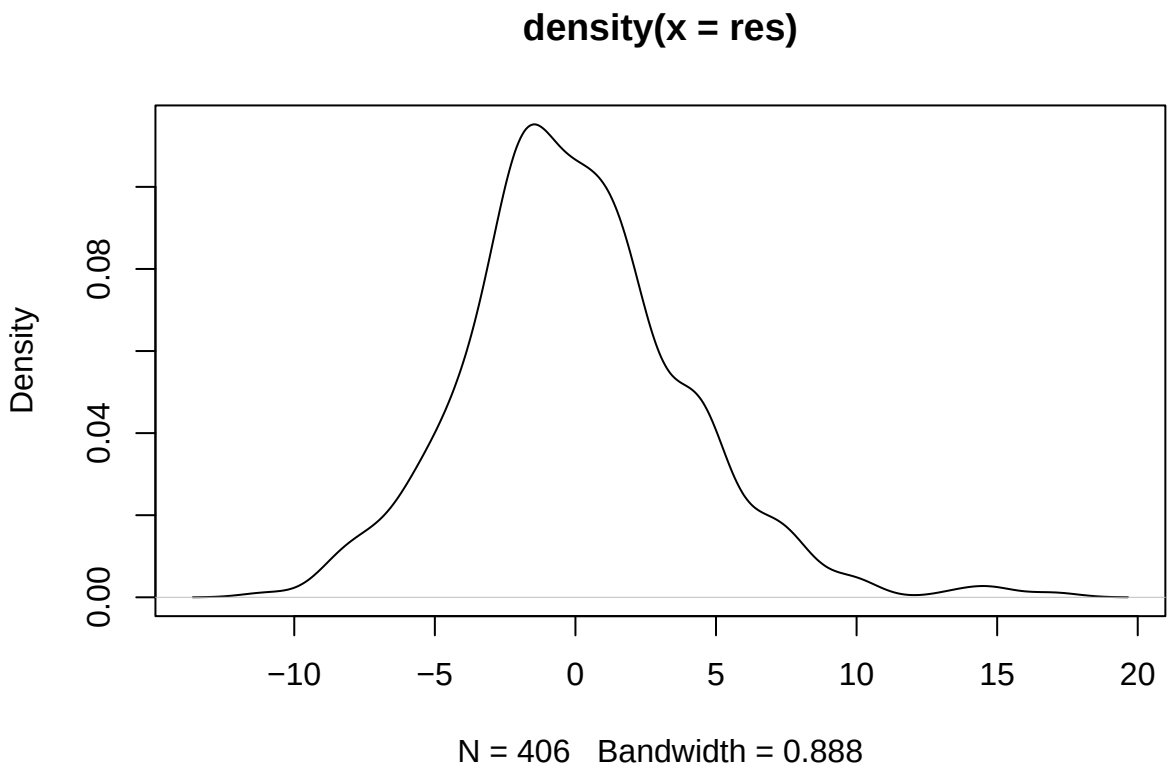

```
plot(acc, res)
```




```
plot(hp, res)
```



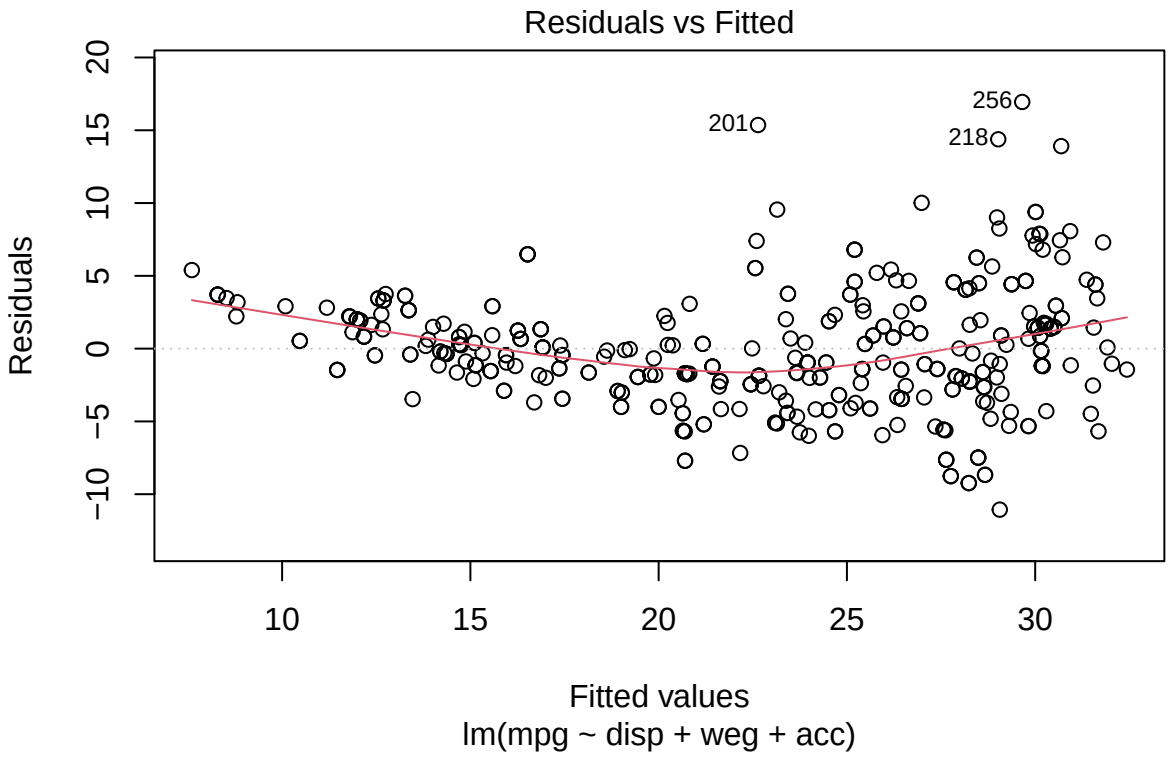

```
plot(density(res))
```

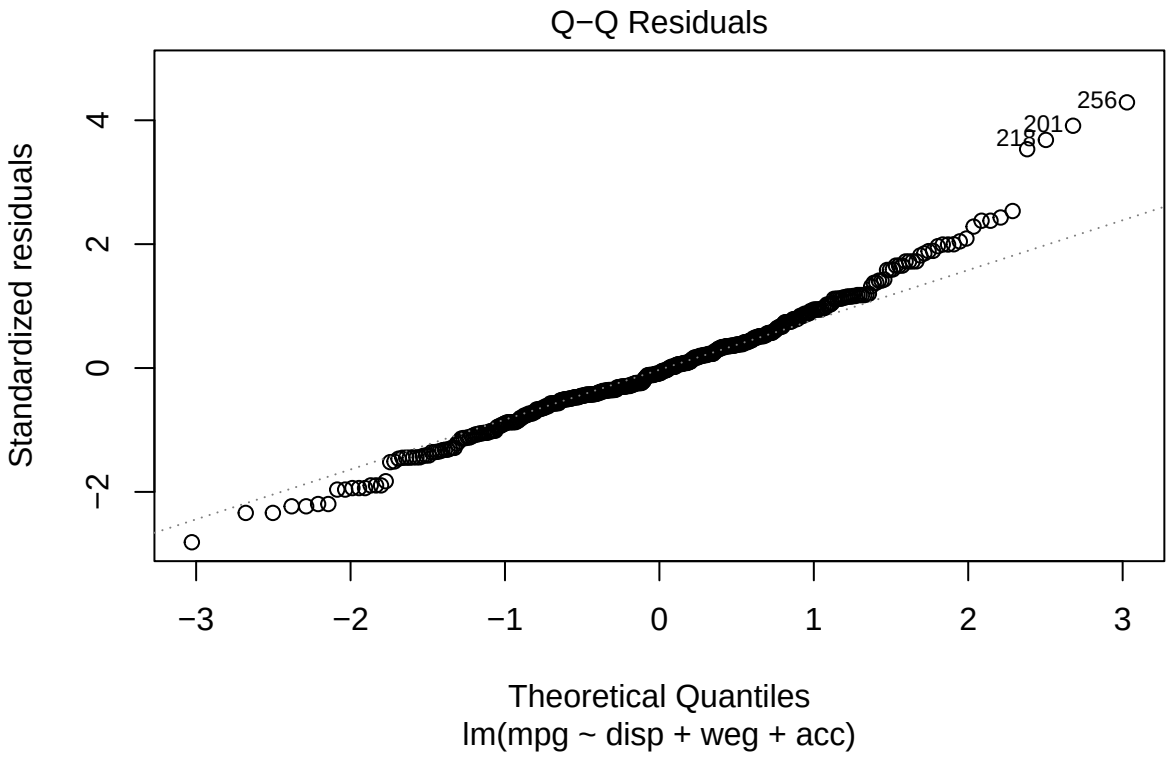


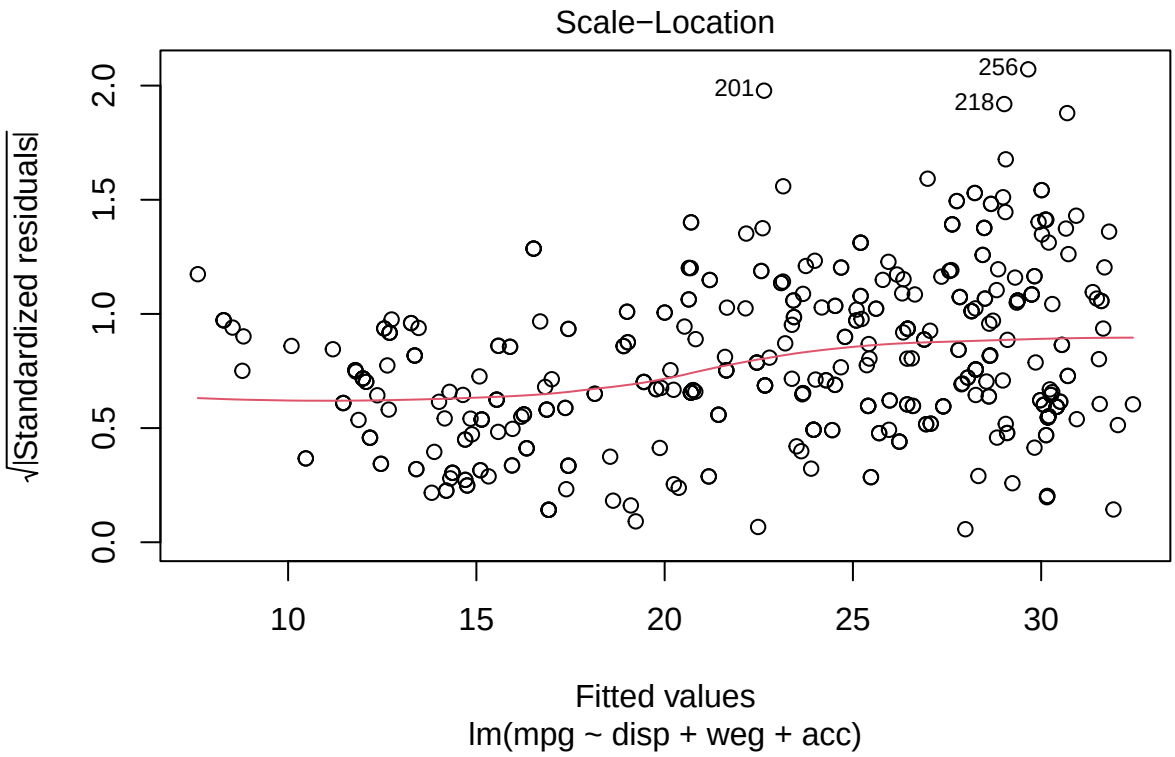
Subset model

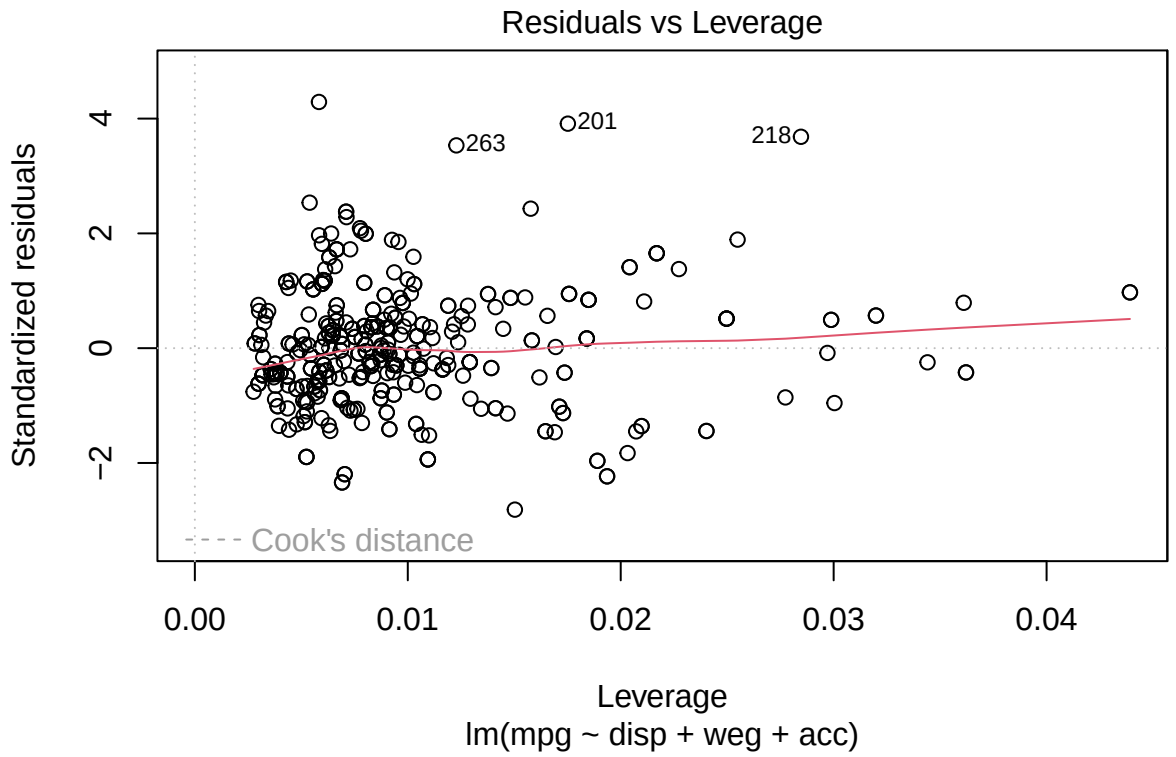
Heading: Subset Model

```
# 3 predictor
model_subset <- lm(mpg ~ disp + weg + acc, data = Auto)
plot(model_subset)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ disp + weg + acc)
201
218
256
−10
−5
0
5
10
15
20
10
15
20
25
30


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(mpg ~ disp + weg + acc)
256
201
218
−3
−2
−1
0
1
2
3
−2
0
2
4


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
201
256
218
10
15
20
25
30
Fitted values
lm(mpg ~ disp + weg + acc)



Summary of subset model

```
# Test the significance of Predictor3
summary(model_subset)
```

```
##
## Call:
## lm(formula = mpg ~ disp + weg + acc, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0594  -2.2555  -0.3287   2.0189  16.9442
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.0501218  1.8139607  22.630   <2e-16 ***
## disp        -0.0066852  0.0063240  -1.057    0.291
## weg         -0.0063133  0.0007225  -8.738   <2e-16 ***
## acc          0.1397572  0.0933108   1.498    0.135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.962 on 402 degrees of freedom
```

```
## Multiple R-squared:  0.7196, Adjusted R-squared:  0.7175
## F-statistic: 343.9 on 3 and 402 DF,  p-value: < 2.2e-16
```

Anova for subset model

```
# Compare the models with and without Predictor3
anova(model_subset, model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ disp + weg + acc
## Model 2: mpg ~ disp + hp + acc + weg
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    402 6308.9
## 2    401 6244.7  1    64.261 4.1265 0.04288 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R-squared for both models

Heading: R-Squared for Both Models

```
summary(model)$r.squared
```

```
## [1] 0.7224765
```

```
summary(model_subset)$r.squared
```

```
## [1] 0.7196207
```

New data prediction

Heading: New Data Prediction

```
new_data <- data.frame(disp = 5, hp = 10, acc = 13, weg = 15)
prediction <- predict(model, newdata = new_data)
print(prediction)
```

```
##        1
## 43.45841
```

Prediction interval for new data

Heading: Prediction Interval for New Data

```
# Calculate the prediction interval for the new set of predictor values
prediction_interval <- predict(model, newdata = new_data, interval = "prediction", level = 0.95)
```