Automobile MPG Analysis

#Load and clean data set

```
# Load the data set
library(readxl)

## Warning: package 'readxl' was built under R version 4.4.2

data = read_excel("C:/Users/filte/Downloads/auto-mpg(1).xlsx")
#use the first 300 rows
data <- data[1:300, ]
#view the first 6 rows of the data
head(data)

## # A tibble: 6 × 9
##     mpg cylinder displacement horsepower weight acceleration `model year`
origin
##   <dbl>    <dbl>        <dbl> <chr>       <dbl>        <dbl>        <dbl>
<dbl>
## 1    18        8          307 130          3504         12             70
1
## 2    15        8          350 165          3693         11.5           70
1
## 3    18        8          318 150          3436         11             70
1
## 4    16        8          304 150          3433         12             70
1
## 5    17        8          302 140          3449         10.5           70
1
## 6    15        8          429 198          4341         10             70
1
## # ℹ 1 more variable: `car name` <chr>

#remove na values:
data <- na.omit(data)
#convert horsepower to numeric data
data$horsepower <- as.numeric(as.character(data$horsepower))

## Warning: NAs introduced by coercion

# Convert 'origin' from numeric to factor
data$origin <- as.factor(data$origin)
```

Evaluate how much wight impacts miles per gallon

```
#preform a simple linear regression on how much wight impacts mpg
linear_regression = lm(mpg ~ weight, data = data)
summary(linear_regression)
```

```
##
## Call:
## lm(formula = mpg ~ weight, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1077 -1.8842 -0.0333  1.7275 15.1232
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.3879027  0.6368804   63.41   <2e-16 ***
## weight      -0.0062524  0.0001957  -31.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.992 on 298 degrees of freedom
## Multiple R-squared:  0.7741, Adjusted R-squared:  0.7733
## F-statistic:  1021 on 1 and 298 DF,  p-value: < 2.2e-16
```

Result of regression

Multiple R-squared: 0.7741 This suggests that about 77.4% of the variation in mpg could be explained by the model, so it is a good fit

Adjusted R-squared: Adjusted R-squared: 0.7733 This is similar to the R-squared value but adjusted for the number of predictors in the model. This is simular to the Multipul R-Squared this suggest that the model is the overfitting

negative correlation indicates that when wight goes down the mpg goes up

equation mpg=40.3879−0.0062524x

#Preform a multiple regression

```
# Multiple Linear Regression with multiple independent variables
multiple_model <- lm(mpg ~ weight + horsepower + displacement + acceleration
+ cylinder, data)
summary(multiple_model)
```

```
##
## Call:
## lm(formula = mpg ~ weight + horsepower + displacement + acceleration +
##     cylinder, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.207 -1.842  0.016  1.604 14.968
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.0054213  2.2389099  18.315  < 2e-16 ***
```

```
## weight       -0.0045575  0.0006155  -7.405 1.41e-12 ***
## horsepower   -0.0259806  0.0125071  -2.077   0.0387 *
## displacement -0.0029379  0.0069194  -0.425   0.6715
## acceleration -0.0615279  0.1052719  -0.584   0.5594
## cylinder     -0.2480496  0.3288403  -0.754   0.4513
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.956 on 292 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.7836, Adjusted R-squared:  0.7799
## F-statistic: 211.5 on 5 and 292 DF,  p-value: < 2.2e-16
```

Multiple R-squared: 0.7836 This suggests that about 78.36% of the variation in mpg could be explained by the model, so it is a good fit

Adjusted R-squared: 0.7799 This is similar to the R-squared value but adjusted for the number of predictors in the model. This is simular to the Multipul R-Squared this suggest that the model is the overfitting

equation mpg=41.0054−0.0045575* weight−0.0259806* horsepower−0.0029379* displacement−0.0615279* acceleration−0.2480496* cylinders

negative correlation indicates that when the other variables goes down the mpg goes up

clean new data

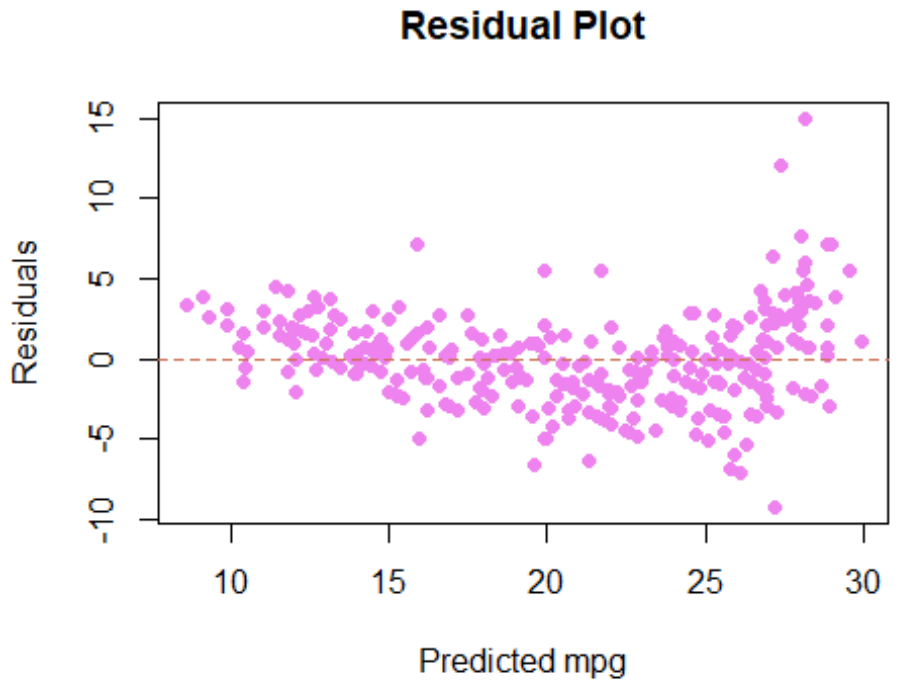
```
#extract the last 98 values
newdata =  data[301:398,]
#remove na values:
newdata <- na.omit(data)
```

Predict mpg using the remaining 98 samples

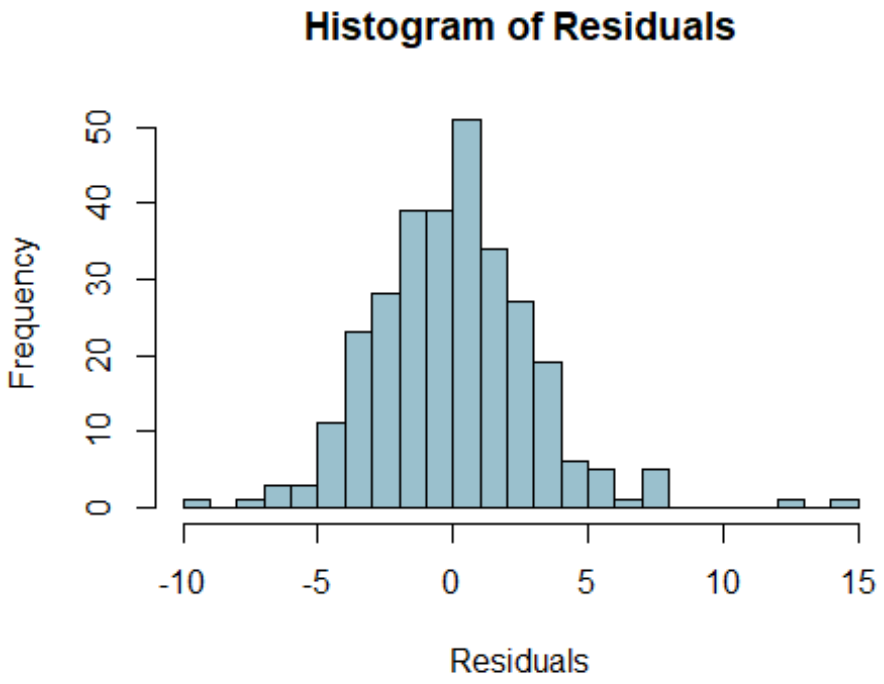
```
#run the predicted model
predicted_mpg <- predict(multiple_model, newdata)

# Actual mpg values for the remaining 98 samples
actual_mpg <- newdata$mpg
#calculate residual
residuals <- actual_mpg - predicted_mpg
# Remove NA or non-finite values from predictions and residuals
valid_indices <- complete.cases(predicted_mpg, residuals) &
is.finite(predicted_mpg) & is.finite(residuals)

# Subset the data to remove invalid entries
predicted_mpg_clean <- predicted_mpg[valid_indices]
residuals_clean <- residuals[valid_indices]
```

Visualize the predicted model

```r
# Residual Plot
plot(predicted_mpg_clean, residuals_clean,
     main = "Residual Plot",
     xlab = "Predicted mpg",
     ylab = "Residuals",
     pch = 19, col = "violet")
abline(h = 0, col = "salmon3", lty = 2)  # Add a horizontal line at 0
```




```r
# Histogram of Residuals
hist(residuals_clean,
     main = "Histogram of Residuals",
     xlab = "Residuals",
     col = "lightblue3",
     breaks = 20)
```



#

The residuals are randomly scattered around 0 with discernible pattern. The model is a good fit The historagram is normaly distrubited indicating a that the data fits the model In conclusion the linear regression model is a good fit to show how the variables affect mpg