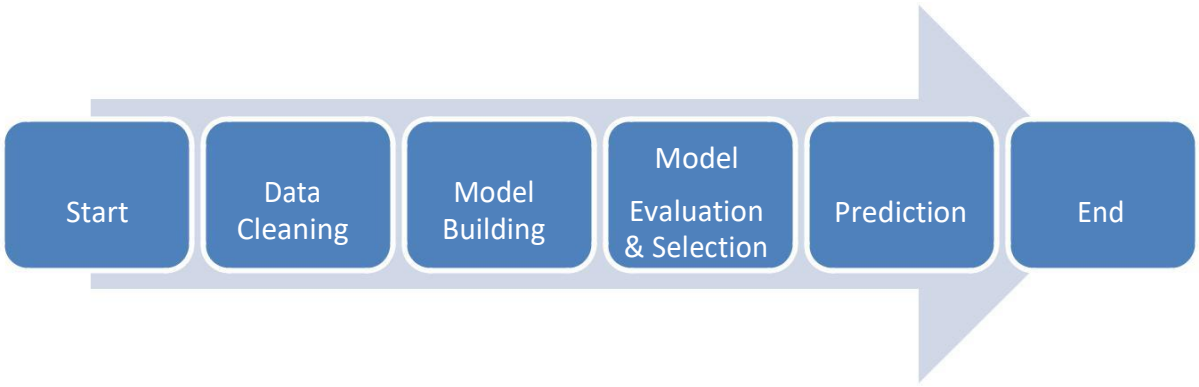
Model for Predicting Auto MPG

Single Dependent Variable Model

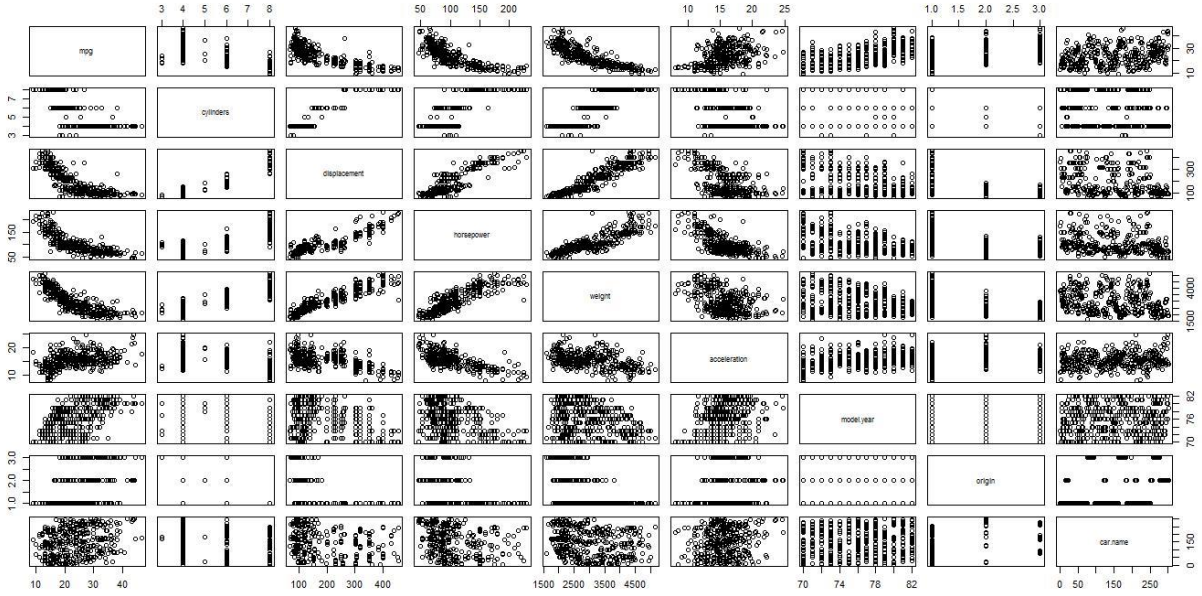


Data Cleaning :

1. Update the data with commas and converted into csv file.
2. Removed the horsepower with null values and imported the data into R

Model Building :

1. A matrix of scatter plot is produced through pairs(car_data_copy)



Only displacement, acceleration, horsepower and weight seem to have a linear relationship with respect to mpg. It is also evident from the fact that all other variables are discrete.

2. Analyzing significance values of all the variables independently

```
         Mpg ~ Weight
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.5619792   0.6461532   62.77   <2e-16 ***
weight       -0.0062905  0.0001984  -31.71   <2e-16 ***

        Mpg ~ Horsepower
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.903508 0.648037 53.86 <2e-16 ***
horsepower  -0.125824   0.005455  -23.07   <2e-16 ***

        Mpg ~
Acceleration Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.0012 1.8352 2.725 0.00681 **
acceleration 1.0379 0.1183 8.770 < 2e-16 ***

        Mpg ~ Displacement

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.352035 0.435875 71.93 <2e-16 ***
displacement -0.048913   0.001809  -27.04   <2e-16 ***
```

All the independent variables have similar significance. Analyzing significance values of all the variables.

summary(lm(mpg ~ ., data=car_data_copy[1:300,]))

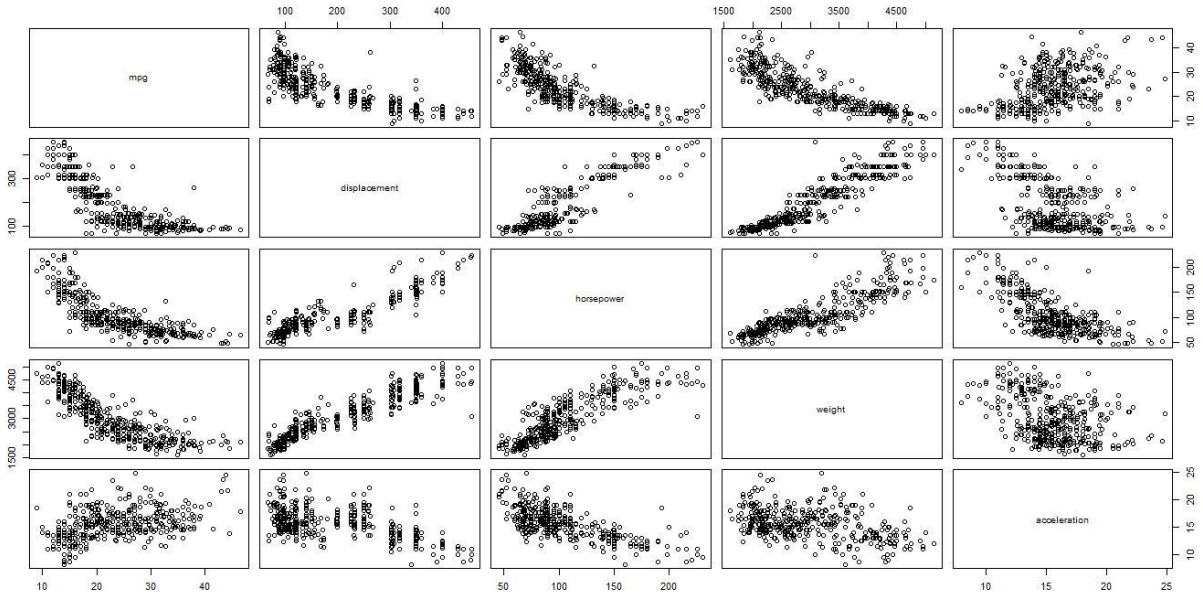
```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   40.5851720  2.0191187  20.100  < 2e-16 ***
displacement  -0.0058876  0.0051269  -1.148   0.2517
horsepower    -0.0270124  0.0124165  -2.176   0.0304 *
weight        -0.0046422  0.0006083  -7.632 3.22e-13 ***
acceleration -0.0593869   0.1032312  -0.575   0.5655
```

Model Building And Selection :

Weight seems to have most significance among other independent variables. And looking at matrix scat -ter plot for all the continuous independent variables weight seems to be more significant than others. We build a model with independent variable as weight and dependent variable as mpg.

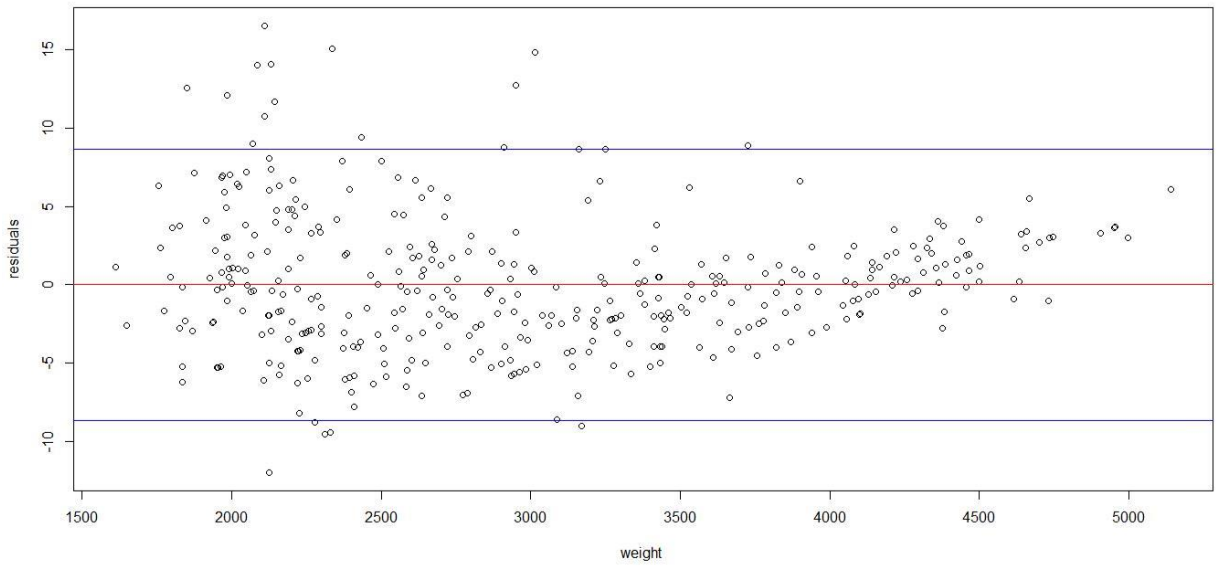
```
lm(formula = mpg ~ weight, data = data_set[1:300, ])
Residuals:     Min       1Q    Median       3Q       Max
            -11.9736  -2.7556  -0.3358   2.1379  16.5194
Coefficients:Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.216524   0.798673   57.87   <2e-16 ***
independent -0.007647   0.000258  -29.64   <2e-16  ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 4.333 on 390 degrees of freedom
Multiple R-squared: 0.6926, Adjusted R-squared: 0.6918 F-
statistic: 878.8 on 1 and 390 DF, p-value: < 2.2e-16
```



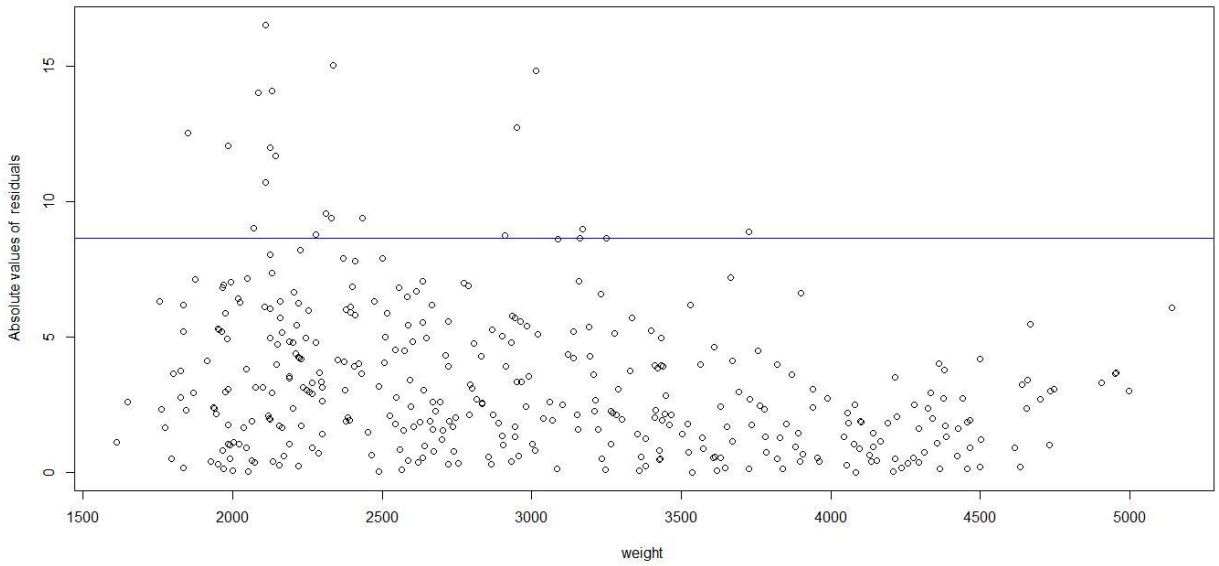
Residuals vs. the predictor variable

Most of the values are under 2 * standard deviation

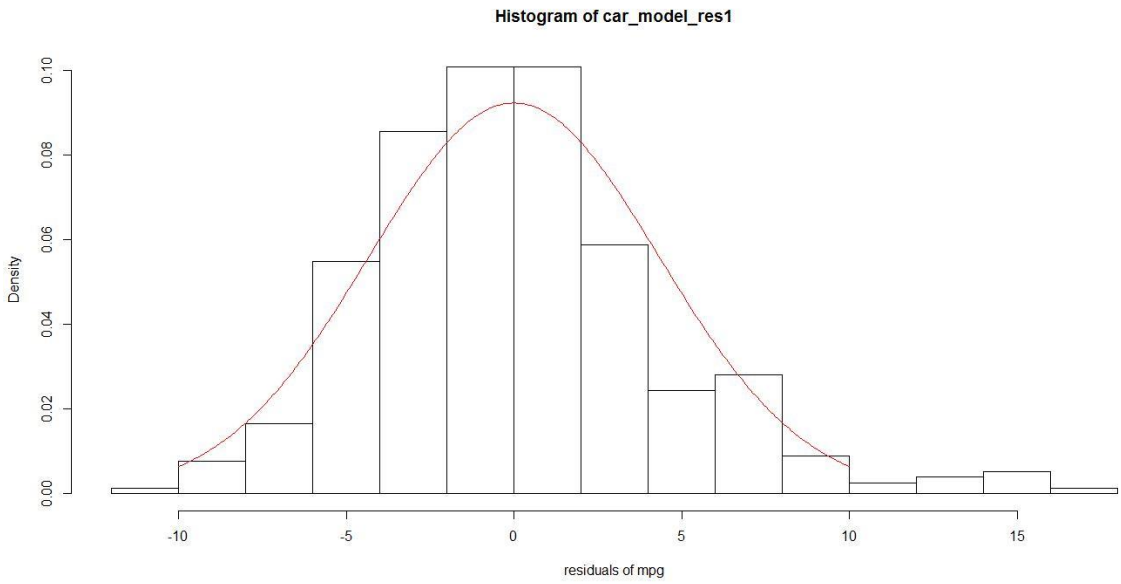


Absolute value of the residuals vs. the predictor variable

Most of the values are under 2 * standard deviation

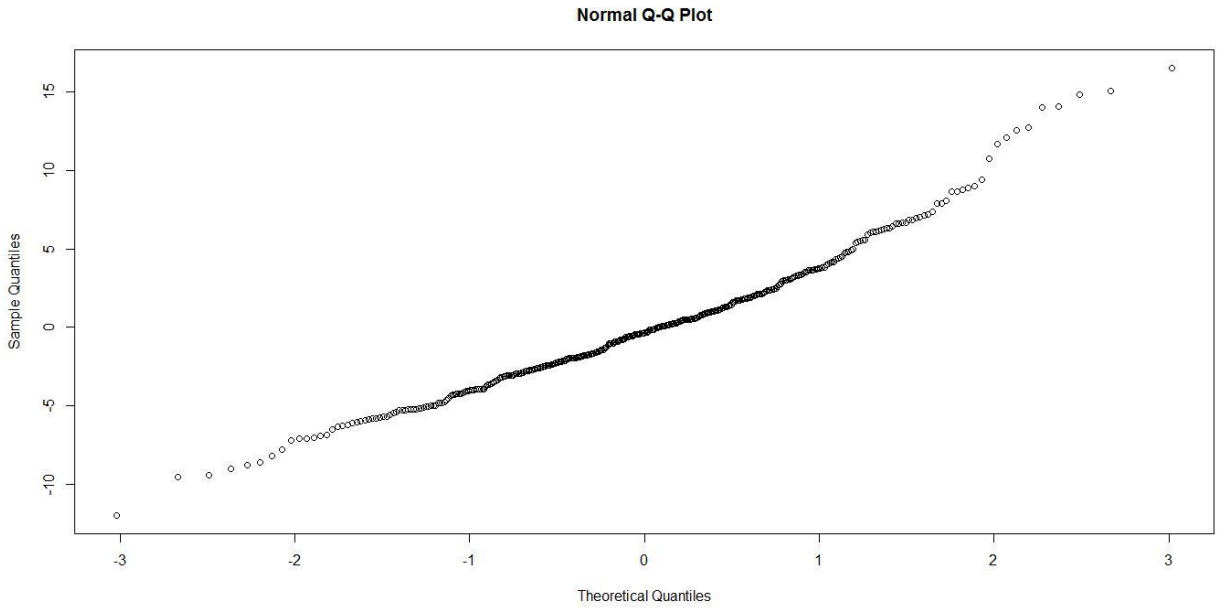


Histogram of the residuals



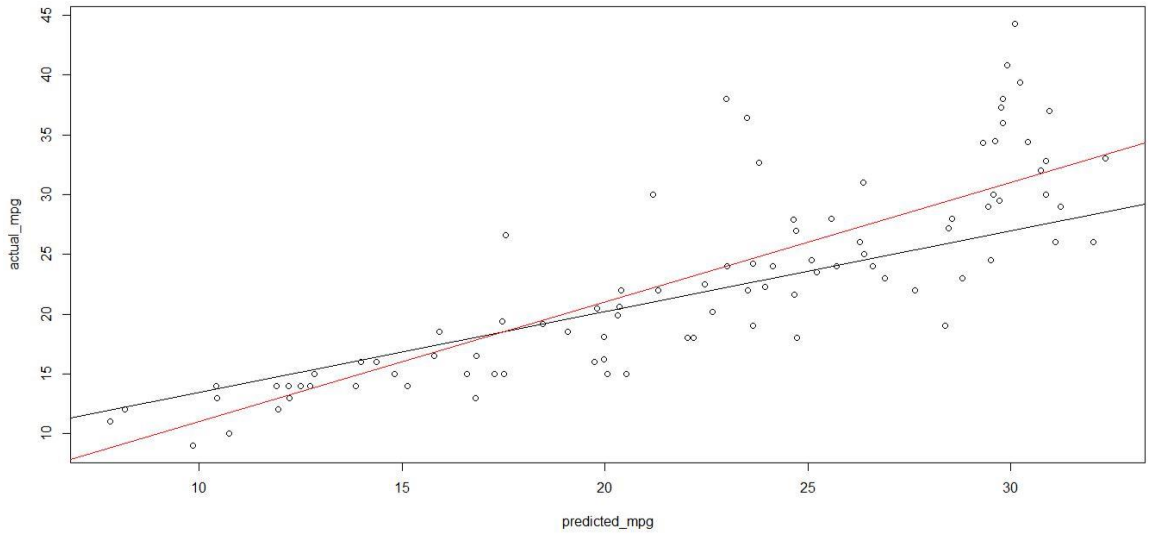
QQ Plot

QQPlot seems slightly linear and becoming non-linear at its far positive end.



Actual_mpg vs predicted_mpg

Red line shows if actual_mpg = predicted_mpg. Black line fitted values of scatter plot between actual_mpg and predicted_mpg.



We randomly select 300 rows of data from 392 rows and we repeat the process for 10 times. We then c ollect coefficients, intercepts, predicted and actual values for test and training data by calling the functio n predict_significant_values.

We find average of all the means of actual and training data

```
mean(predictedlist_weight$mean_actual_mpg_test_92) [1]
23.11543
mean(predictedlist_weight$mean_predicted_mpg_test_92)
[1] 23.09936
```

Prediction :

We build a model using average values β_0 and β_1 coefficients, intercepts

```
mc <- mean(predictedlist_weight$coefficients)
mi <- mean(predictedlist_weight$intercept)
```

β_0 = 46.35704 and β_1 = 0.003773136

```
best_model_mpg <- mi + mc * car_data_copy$weight
```

Average values of actual and predicted -

```
mean(car_data_copy$mpg)
23.44592
mean(best_model_mpg)
23.44215
```

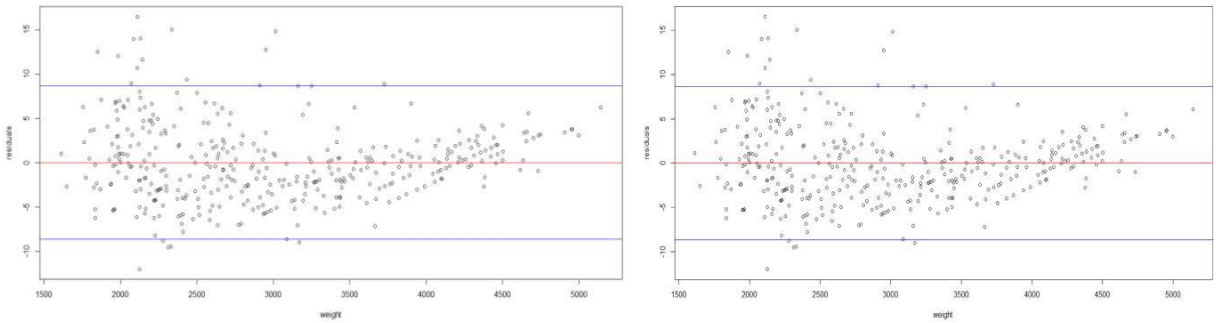
```
best_model_mpg_res <- car_data_copy$mpg - best_model_mpg
```

mean(best_model_mpg_res) = 0.003773136

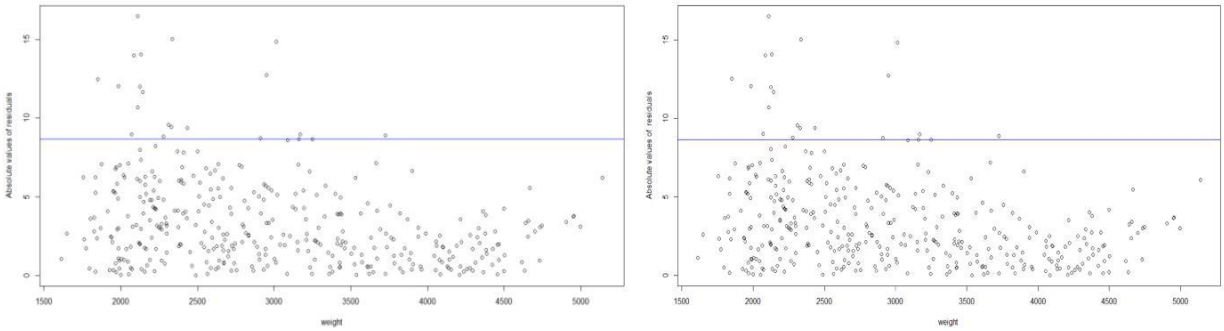
Mean of the best model comes out to be close to zero (0.003773136)

Best Model vs Random Model

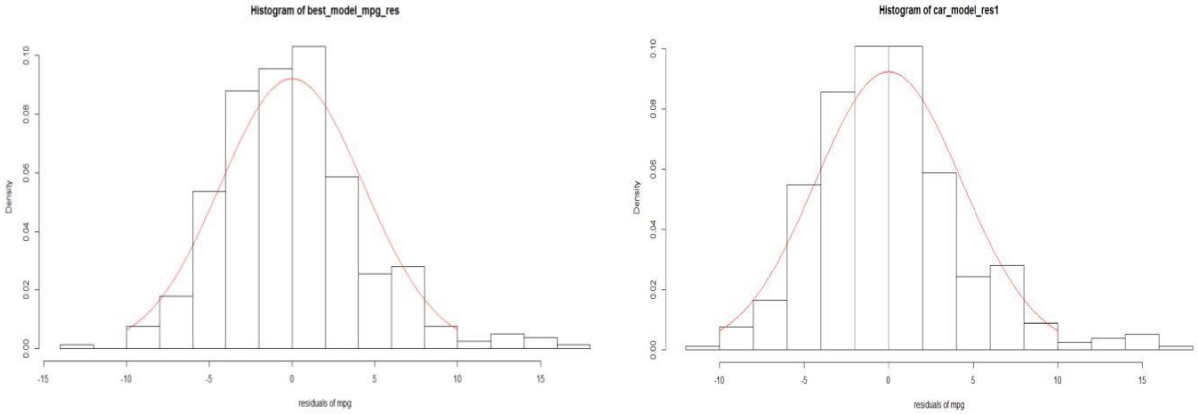
Residuals vs. the predictor variable



Absolute value of the residuals vs. the predictor variable

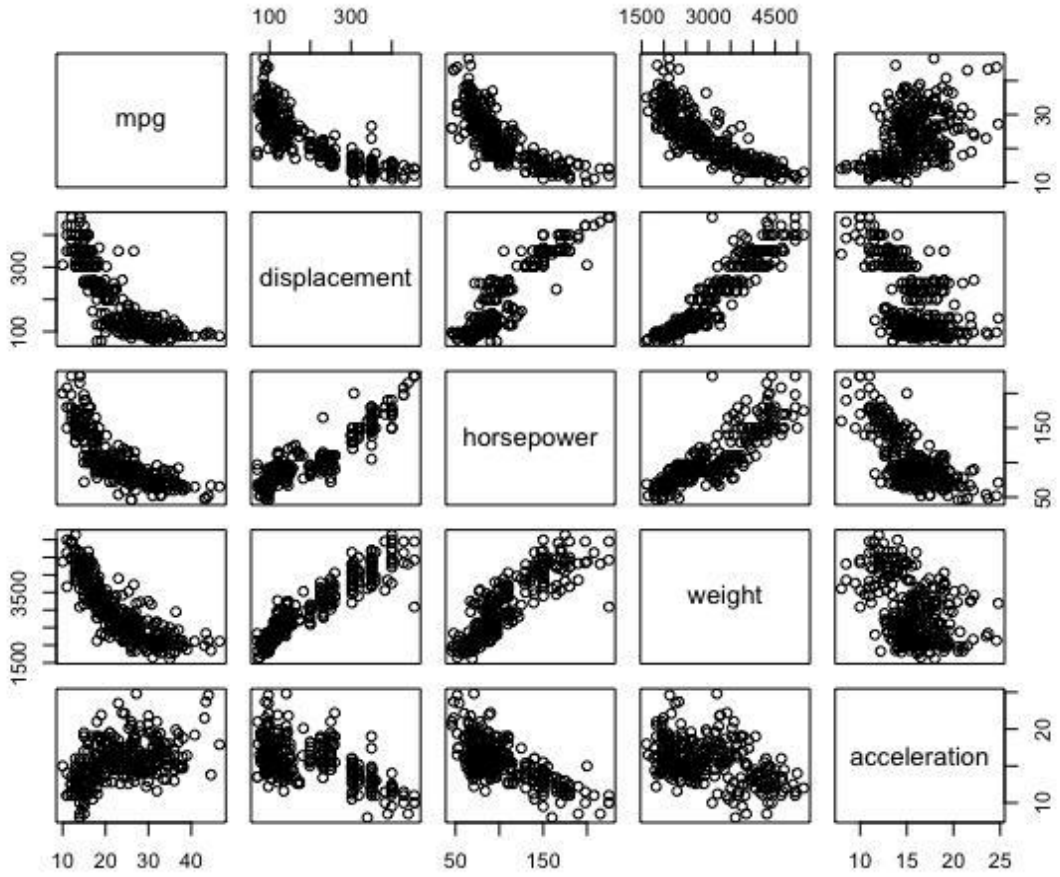


Histogram of the residuals

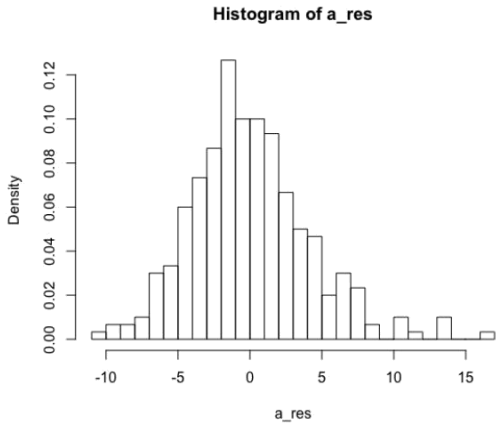


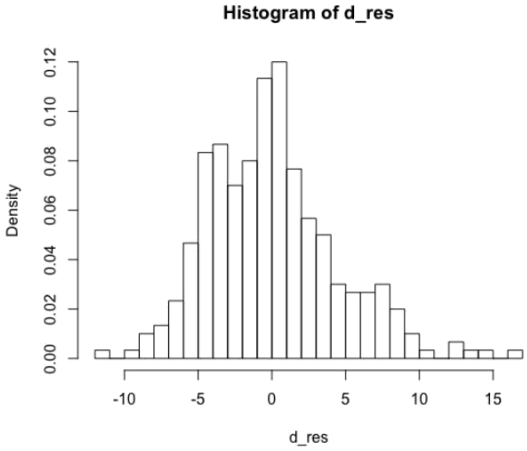
Question 2

For the second question, we had to build a model with more than one independent variable. We created a new matrix of 300 observations that were chosen randomly from our main data file to start building our model. After looking at the pairs of all continuous variables, we realized that the variable "acceleration" didn't have a linear correlation with our dependent variable mpg.



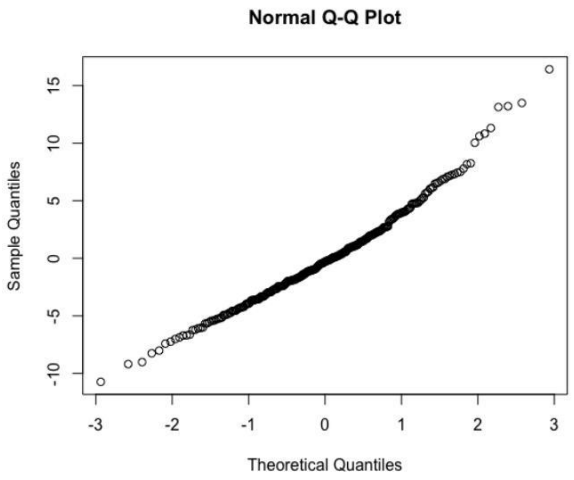
In order to verify that, we ran a regression with horsepower, displacement, weight and acceleration. Since acceleration had a large p-value (0.841) we decided not to include it into our model. Therefore, we were left with 3 optional variables for our model: acceleration, displacement and weight. Since we had to include at least two variables in our model we had 4 possible distinct models. We named them a, b, c and d. After running each one of them, we decided to compare the only two models that had the most significant p values: 'a' and 'd' (summary of all 4 models is in included in our script). We examined 4 elements: magnitude of p values, mean of each model's residuals (indicated in the script), distribution of the residuals and the normal QQ plot of the residuals.

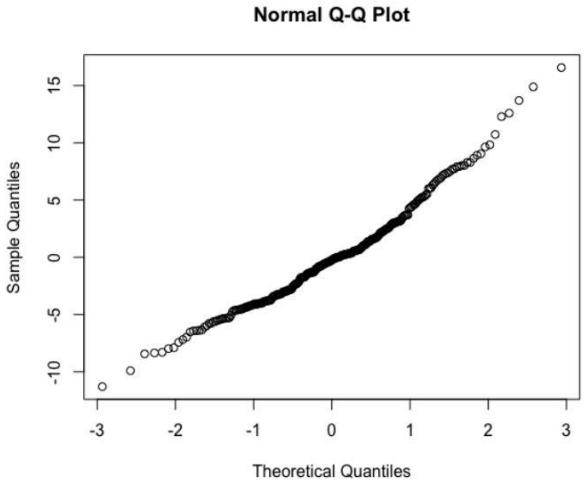




Model 'a':

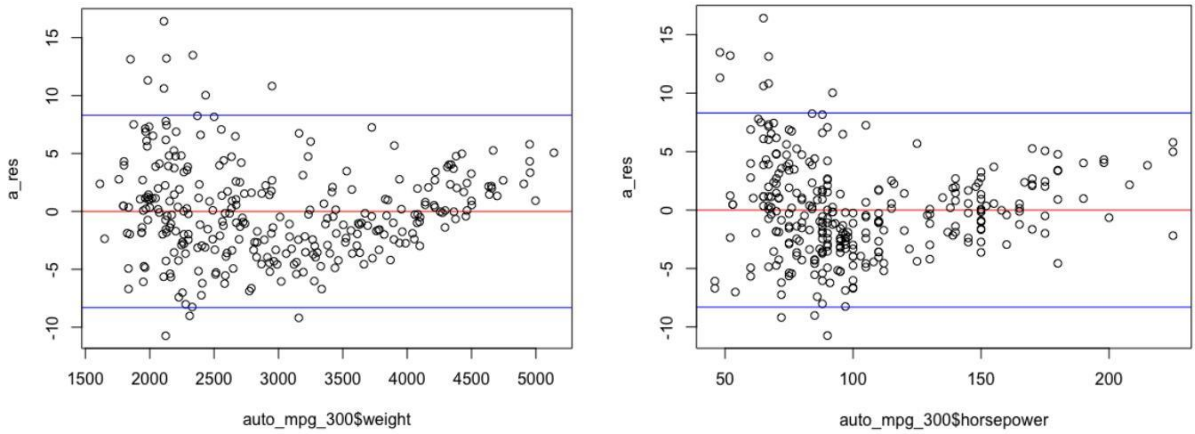
Model 'd':





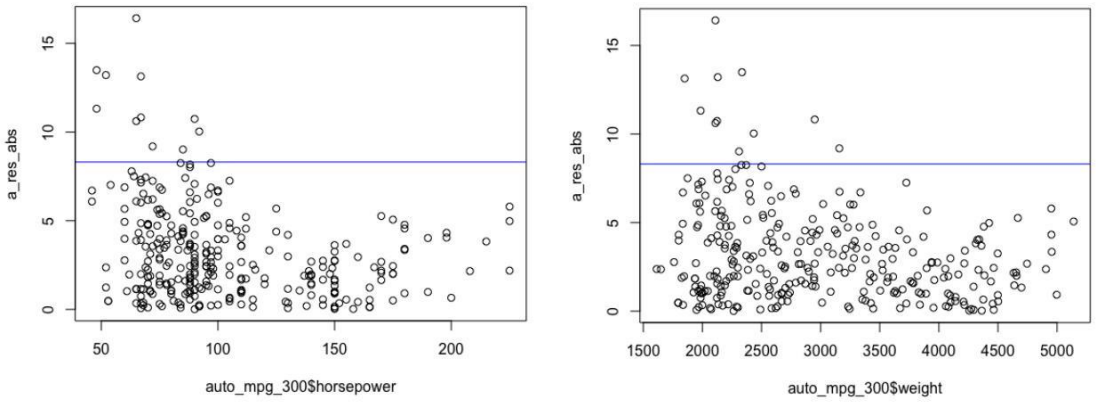
Model 'a' had smaller p-values, closer mean to 0, more normally shaped and a straighter normal QQ plot than model 'd'. Thus, we decided to choose model 'a' as our best model.

(a) Residuals vs. predictor variables



We can see that for both plots, most of the points fall within the region of 2 standard deviations (denoted as the blue lines).

(B) Absolute value of the residuals vs. the predictor variable



The absolute value of the residual simply gives us a mirroring approach of the graphs from part (a), which illustrates that a small part of the residuals is out of the 2 standard deviations bound (all point above the blue line).

(c) histogram of the residuals (was indicated in the comparison)

For the second question we simply used the coefficients of model 'a' to predict the estimated value of the observations that weren't included within the first 300 randomly chosen observations. We discovered that the expected value of the **estimated** mpg values was **23.03** vs. **23.326** of the **actual** mpg values, which gives us a sufficiently accurate prediction.