Linear Regression Project

Deenadayalamuthu and Mathew

Introduction about the dataset:

The data set auto-mpg.data contains information for 398 different automobile models. Information regarding the mpg, number of cylinders, displacement, horsepower, weight, acceleration, model year, origin and car name is given. Using this data set we have created two models of linear regression to predict the mpg.

Exploratory Data Analysis & Data Preparation:

The data set contains 5 continuous variables. Horsepower has got 6 missing values which are represented as '?' in the data set. For both single variable and multi-variable linear regressions we have taken mpg as dependent variable.

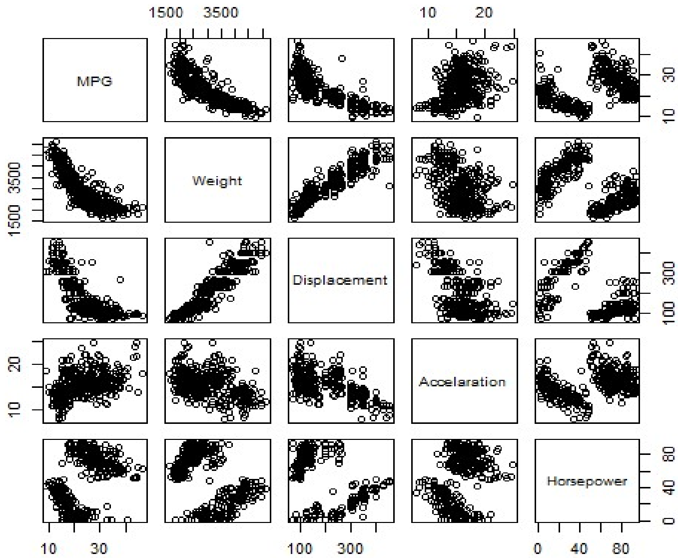
First we have given names for each column and then we have split the data into training set and test set. The training set contains first 300 observations from the original data set and test data set contains last 98 observations from the original data set.

```
names(cardata) <- c("MPG", "Cylinders", "Displacement", "Horsepower", "Weight",
                    "Accelaration", "ModelYear", "Origin", "CarName")

cardataTraining <- cardata[1:300, ]
cardataTest <- cardata[301:398,]
```

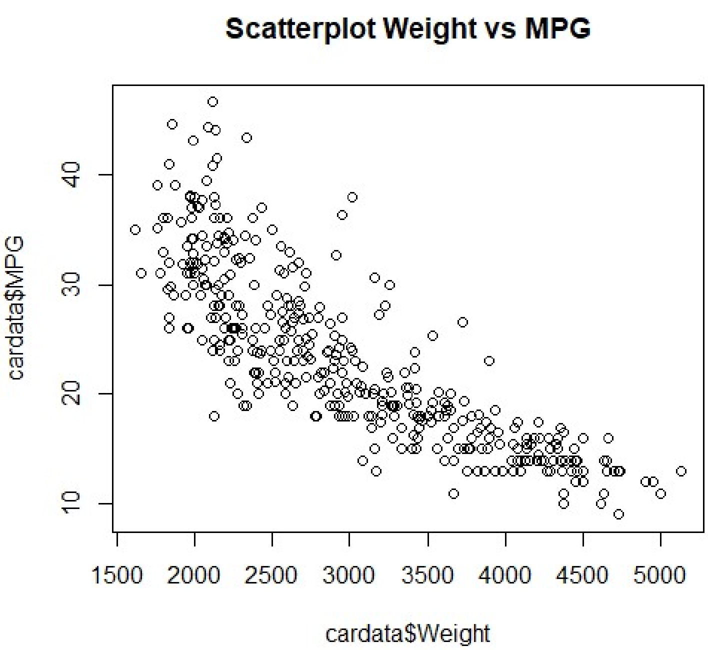
Before proceeding further we have analyzed the pairs chart for the continuous variables to understand the relationship between them.

```
pairs(~ MPG + Weight + Displacement + Accelaration + Horsepower, data=cardata)
```



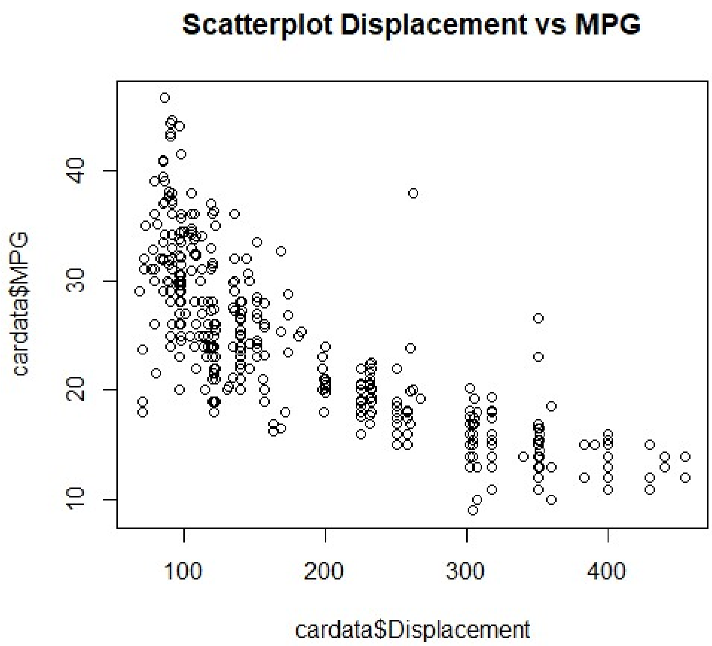
To understand the relationship better, we have calculated the correlation between MPG and other continuous variables weight, displacement, acceleration and horsepower. We have also made separate scatter plot for the same.

Weight vs MPG:



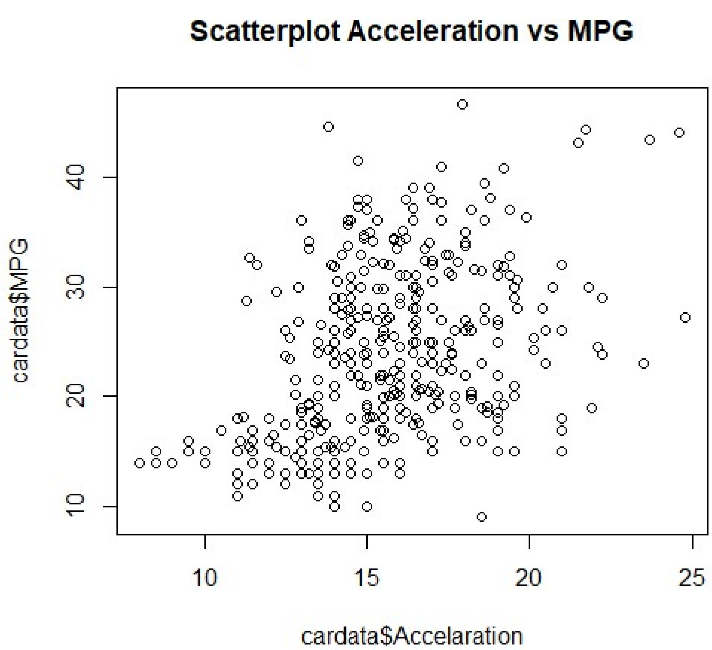
Correlation between Weight and MPG is -0.8317409. Correlation suggests that weight is good candidate for linear regression

Displacement vs MPG:



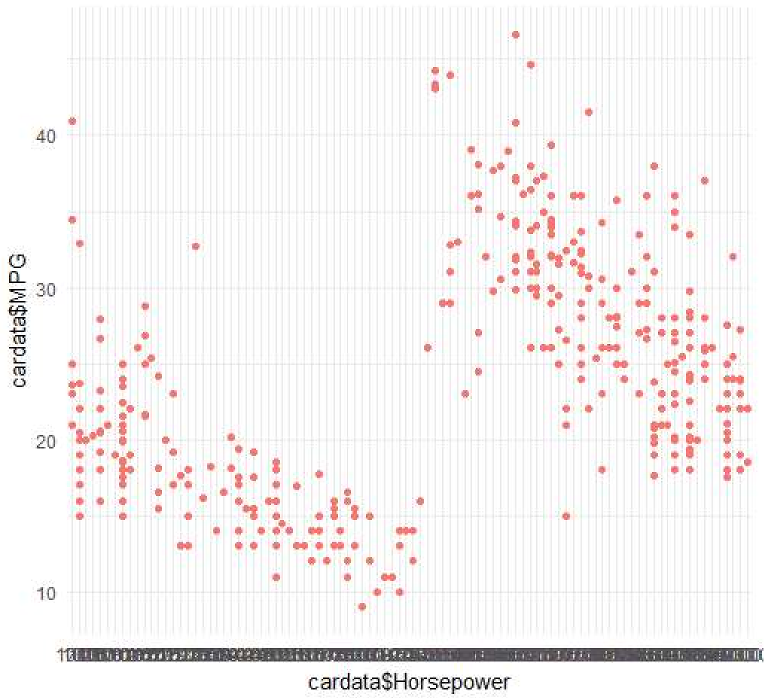
Correlation between Displacement and MPG is -0.8042028. Correlation suggests displacement may be likely candidate for linear regression. But from the scatterplot it looks like suitable for logistic regression.

Acceleration vs MPG:



Correlation between Acceleration and MPG is 0.4202889. Correlation number suggests there is very little correlation between acceleration and MPG.

Horsepower vs MPG:



Horsepower data contains missing values. So we are not able to compute correlation between horsepower and MPG. But from the scatterplot it looks like horsepower is an unlikely candidate for linear regression.

From the above charts and correlation we chose weight as a predictor of MPG for the single variable linear regression as it has the most correlation and linearity out of the four continuous variables.

Correlation suggests weight and mpg are negatively correlated.

Building Single –Variable Linear Regression model:

cardataLinReg <- lm(MPG ~ Weight, data = cardataTraining)

Summary of model:

```
Call:
lm(formula = MPG ~ Weight, data = cardataTraining)

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
```

Predicting MPG based on the model:

Based on the model built we used the intercept, coefficients and weight from test data set we have predicted MPG.

```
predicted_y_sv_cardata <- cardataLinReg$coefficients[1] +
  cardataLinReg$coefficients[2] * cardataTest$Weight
```

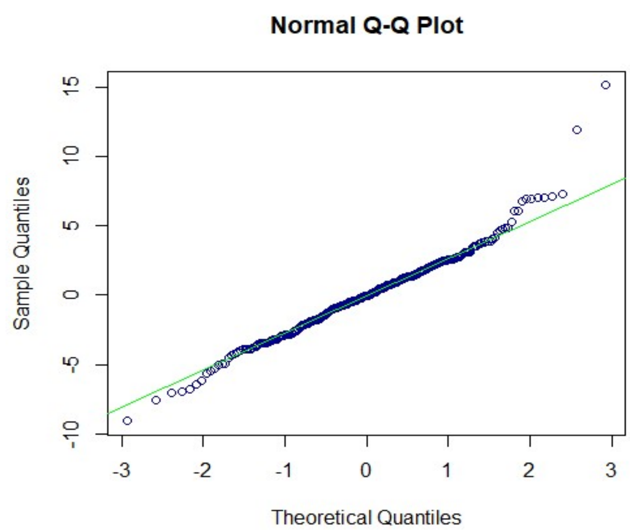
Calculating Model Error:

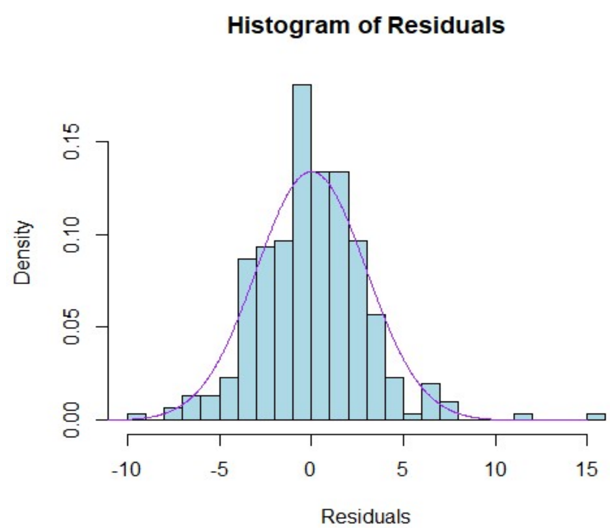
```
sv_modelerror <- cardataTest$MPG - predicted_y_sv_cardata
```

Histogram and QQ-Plot of Residuals

```
hist(cardataLinReg$residuals, breaks = 30, probability = T,
                              col = "light blue",
                              main = "Histogram of Residuals",
                              xlab = "Residuals")
x <- seq(-15, 15, length = 1000)
y <-dnorm(x, mean = mean(cardataLinReg$residuals),
               sd = sd(cardataLinReg$residuals))
lines(x,y,col="Purple")

qqnorm(cardataLinReg$residuals, col="Dark Blue")
qqline(cardataLinReg$residuals, col="Green")
```

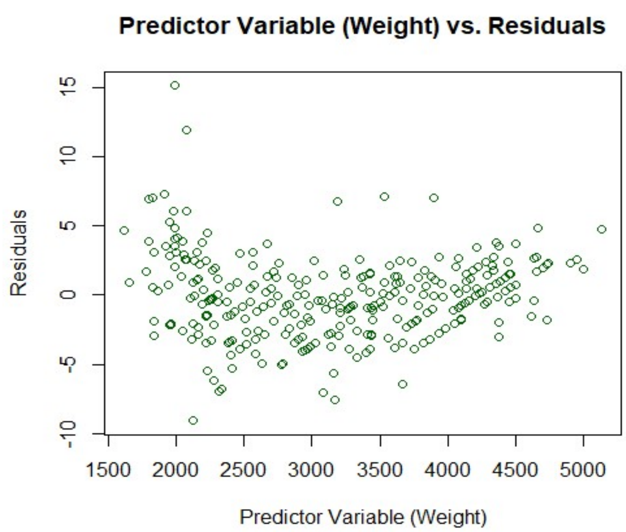


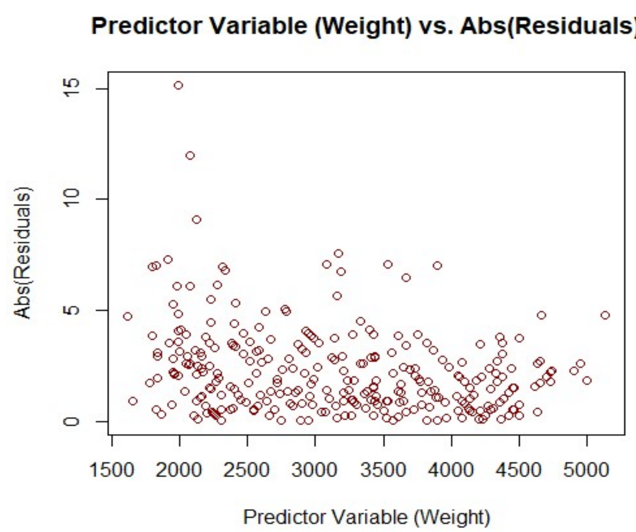


From the qq plot and the histogram of residuals it is clear that the residuals are normally distributed. Therefore it is proved that model is fitting exactly as it is expected.

Plots for Residuals:

```
plot(cardataTraining$Weight, cardataLinReg$residuals,
    col = "Dark Green",
    main = "Predictor Variable (Weight) vs. Residuals",
    xlab = "Predictor Variable (Weight)",
    ylab = "Residuals")
plot(cardataTraining$Weight, abs(cardataLinReg$residuals),
    col = "Dark Red",
    main = "Predictor Variable (Weight) vs. Abs(Residuals)",
    xlab = "Predictor Variable (Weight)",
    ylab = "Abs(Residuals)")
```





Building Multi-variable Linear Regression:

Before building a multi variable linear regression we normalized the continuous variables weight, displacement, acceleration to get all the variables in same scale.

```
cardata$norm_disp <- (cardata$Displacement -
                       mean(cardata$Displacement))/sd(cardata$Displacement)
cardata$norm_wt <- (cardata$Weight -
                       mean(cardata$Weight))/sd(cardata$Weight)
cardata$norm_acc <- (cardata$Accelaration -
                      mean(cardata$Accelaration))/sd(cardata$Accelaration)
```

Then we have split the dataset into training data set and test data set

```
mod_cardata_train <- mod_cardata[1:300,]
mod_cardata_test <- mod_cardata[301:398,]
```

Then we built 4 different models using different combinations of explanatory variables to predict MPG and explored the summary of those model to choose better combination of explanatory variables.

```
multi_val_reg1 <- lm(MPG ~ norm_wt + norm_disp, data = mod_cardata_train)
multi_val_reg2 <- lm(MPG ~ norm_wt + norm_disp + norm_acc, data =
mod_cardata_train)
multi_val_reg3 <- lm(MPG ~ norm_disp + norm_acc, data = mod_cardata_train)
multi_val_reg4 <- lm(MPG ~ norm_wt + norm_acc, data = mod_cardata_train)
```

Summary of multi – variable linear regression:

```
Call:
lm(formula = MPG ~ norm_wt + norm_disp, data = mod_cardata_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.4362 -1.8464 -0.1621  1.6470 15.2024 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  21.8453     0.1743 125.347   <2e-16 ***
norm_wt      -4.1929     0.4465  -9.390   <2e-16 ***
norm_disp    -1.1827     0.4458  -2.653   0.0084 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.963 on 297 degrees of freedom
Multiple R-squared:  0.7793,	Adjusted R-squared:  0.7778 
F-statistic: 524.5 on 2 and 297 DF,  p-value: < 2.2e-16
```

We chose the first model (Weight & Displacement as explanatory variable) because out of the four models this is the model with more significance for the explanatory variables.

Predicting MPG based on the model:

```
pred_y_mv_cardata1 <- multi_val_reg1$coefficients[1] +
    multi_val_reg1$coefficients[2]*mod_cardata_test$norm_wt +
    multi_val_reg1$coefficients[3]*mod_cardata_test$norm_disp
```

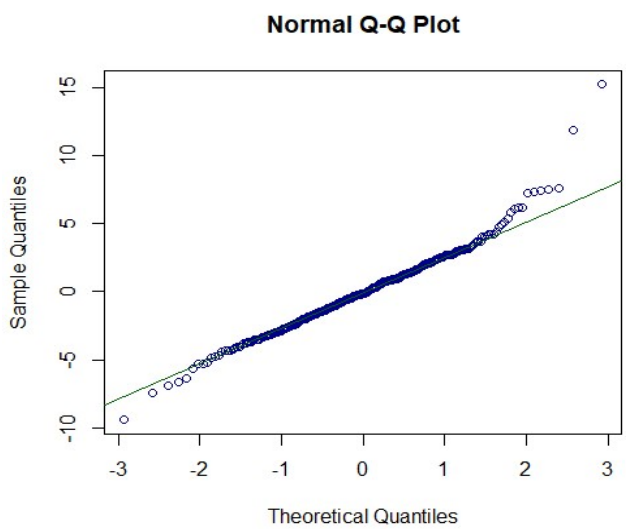
Calculating Model Error:

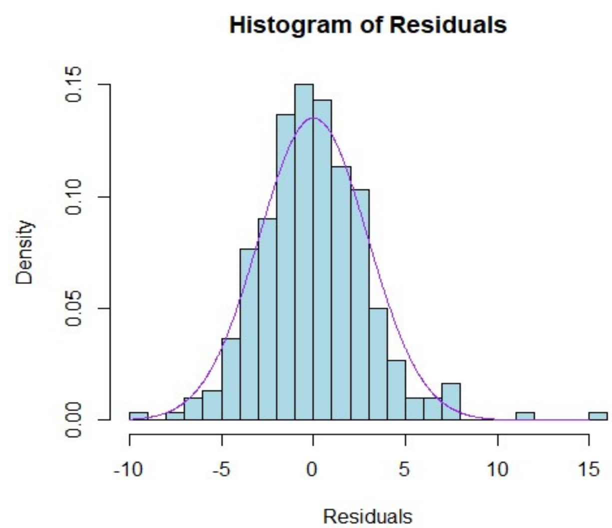
```
mv_modelerror <- mod_cardata_test$MPG - pred_y_mv_cardata1
```

Histogram and QQ-Plot of Residuals:

```
hist(multi_val_reg1$residuals, breaks = 30, probability = T, main="Histogram of
Residuals",     xlab = "Residuals", col = "light blue")
x <- seq(-10, 15, length = 1000)
y <- dnorm(x, mean = mean(multi_val_reg1$residuals),sd =
sd(multi_val_reg1$residuals))
lines(x,y,col="Purple")

qqnorm(multi_val_reg1$residuals, col="Dark Blue")
qqline(multi_val_reg1$residuals, col="Dark Green")
```





From the qq plot and the histogram of residuals it is clear that the residuals are normally distributed. Therefore it is proved that model is fitting exactly as it is expected.

Plots for Residuals:

```
plot (mod_cardata_train$norm_wt, multi_val_reg1$residuals,
    col = "Dark Blue", main = "Residuals vs. Weight",
    xlab = "Weight", ylab = "Residuals")

plot (mod_cardata_train$norm_disp, multi_val_reg1$residuals,
    col = "Dark Blue", main = "Residuals vs. Displacement",
    xlab = "Displacement", ylab = "Residuals")

plot(mod_cardata_train$norm_wt, abs(multi_val_reg1$residuals),
    col = "Dark Blue", main = "Abs. Residuals vs. Weight",
    xlab = "Weight", ylab = "Abs. Residuals")

plot(mod_cardata_train$norm_disp, abs(multi_val_reg1$residuals),
    col = "Dark Blue", main = "Abs. Residuals vs. Displacement",
    xlab = "Displacement", ylab = "Abs. Residuals")
```