Fuel Efficiency Analysis Using Linear Regression

Ravi Kumar Tiwari

14 June 2016

Introduction

In this analysis, I am going to build a linear regression model to predict the fuel efficiency of a car based on its characteristics such as its weight, horsepower, cylinder displacement.

Data Exploration

Load the required libraries

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.2.5
```

Load the dataset

```r
fuelData <- read.table("FuelEfficiency.csv", sep = ",", header = TRUE)
```

Look at the data set summary

```r
names(fuelData)
```

```
## [1] "MPG" "GPM" "WT"  "DIS" "NC"  "HP"  "ACC" "ET"
```

```r
dim(fuelData)
```

```
## [1] 38  8
```

```r
str(fuelData)
```

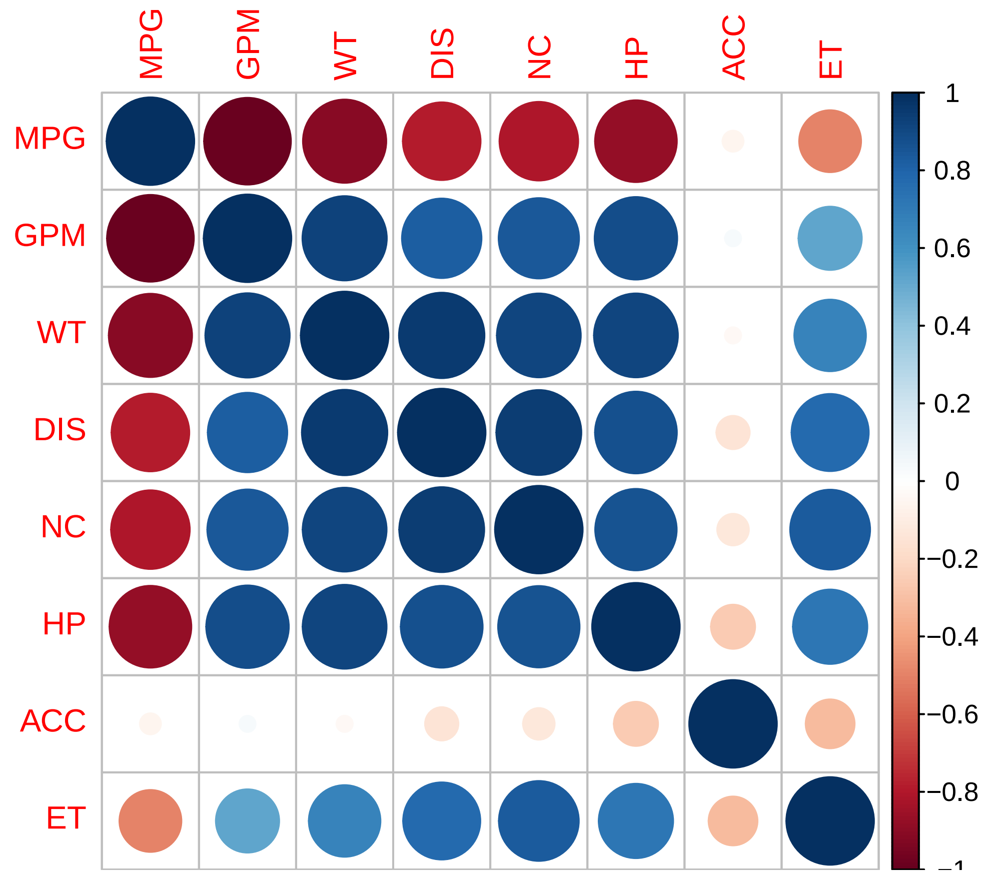
```
## 'data.frame':    38 obs. of  8 variables:
##  $ MPG: num  16.9 15.5 19.2 18.5 30 27.5 27.2 30.9 20.3 17 ...
##  $ GPM: num  5.92 6.45 5.21 5.41 3.33 ...
##  $ WT : num  4.36 4.05 3.6 3.94 2.15 ...
##  $ DIS: int  350 351 267 360 98 134 119 105 131 163 ...
##  $ NC : int  8 8 8 8 4 4 4 4 5 6 ...
##  $ HP : int  155 142 125 150 68 95 97 75 103 125 ...
##  $ ACC: num  14.9 14.3 15 13 16.5 14.2 14.7 14.5 15.9 13.6 ...
##  $ ET : int  1 1 1 1 0 0 0 0 0 0 ...
```

```r
summary(fuelData)
```

```
##       MPG              GPM              WT               DIS       
##  Min.   :15.50   Min.   :2.681   Min.   :1.915   Min.   : 85.0  
##  1st Qu.:18.52   1st Qu.:3.292   1st Qu.:2.208   1st Qu.:105.0  
##  Median :24.25   Median :4.160   Median :2.685   Median :148.5  
##  Mean   :24.76   Mean   :4.331   Mean   :2.863   Mean   :177.3  
##  3rd Qu.:30.38   3rd Qu.:5.398   3rd Qu.:3.410   3rd Qu.:229.5  
##  Max.   :37.30   Max.   :6.452   Max.   :4.360   Max.   :360.0  
##        NC              HP              ACC               ET        
##  Min.   :4.000   Min.   : 65.0   Min.   :11.30   Min.   :0.0000  
##  1st Qu.:4.000   1st Qu.: 78.5   1st Qu.:14.03   1st Qu.:0.0000  
##  Median :4.500   Median :100.0   Median :14.80   Median :0.0000  
##  Mean   :5.395   Mean   :101.7   Mean   :14.86   Mean   :0.2895  
##  3rd Qu.:6.000   3rd Qu.:123.8   3rd Qu.:15.78   3rd Qu.:1.0000  
##  Max.   :8.000   Max.   :155.0   Max.   :19.20   Max.   :1.0000
```

I am interested in knowing the relationship between MPG (Miles per Gallon) with rest of the variables. First of all, I look at the correlation of MPG with all other variables

```
corVal <- cor(fuelData)
corrplot(corVal)
```



```
corVal[,1]
```

```
##         MPG         GPM          WT         DIS          NC          HP 
##  1.00000000 -0.98079724 -0.90307083 -0.78604807 -0.80551105 -0.87128209
##         ACC          ET
## -0.05677359 -0.49816677
```

Unsuprisingly, GPM (Gallon per Miles) is very highly negatively correlated with MPG (Miles per Gallon) as by definition they are inverse of each other. I will therefore use only one of them in my model. I chose GPM because of its positive correlation with other variables which makes interpretation easy. Also, I leave out ACC variable in my model as it is uncorrelated with GPM.

Model building

```
lmModel <- lm(GPM ~ WT + DIS + NC + HP + ET, data = fuelData)
```

Model assessment

cross-validation

```
n <- nrow(fuelData)
diff <- vector(mode = "numeric", length = n)

for (i in 1:n){
  train <- fuelData[-i,]
  test <- fuelData[i,]

  model <- lm(GPM ~ WT + DIS + NC + HP + ACC + ET, data = train)
  yPredict <- predict(model, test)
  y <- test$GPM
  diff[i] <- yPredict - y
}

mean(diff)
```

```
## [1] 0.003981948
```

```
RMSE <- sqrt(sum(diff^2)/length(diff))
RMSE
```

```
## [1] 0.3491357
```

```
summary(lmModel)
```

```
##
## Call:
## lm(formula = GPM ~ WT + DIS + NC + HP + ET, data = fuelData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60029 -0.20966  0.03059  0.21054  0.65544
##
## Coefficients:
```

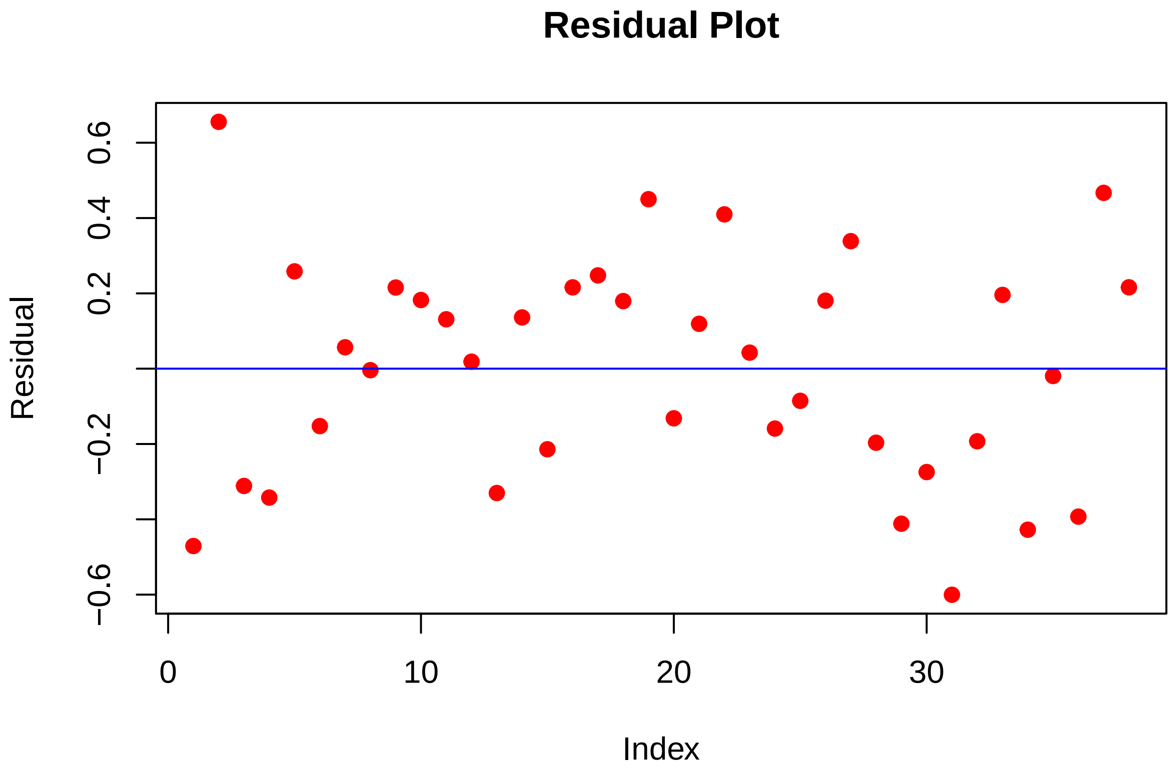
```
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -1.784938   0.416767  -4.283 0.000157 ***
## WT           1.160768   0.391573   2.964 0.005688 ** 
## DIS         -0.006481   0.002550  -2.542 0.016078 *  
## NC           0.447929   0.125356   3.573 0.001142 ** 
## HP           0.017668   0.005684   3.108 0.003931 ** 
## ET          -0.941913   0.272402  -3.458 0.001561 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3199 on 32 degrees of freedom
## Multiple R-squared:  0.9338, Adjusted R-squared:  0.9234 
## F-statistic: 90.23 on 5 and 32 DF,  p-value: < 2.2e-16
```

All the variables in my model are significant as can be seen from there p-values. Also, my model is able to explain 93.38% variation in the dependent variable which can be inferred from the model r-squared value of 0.9338. Let us know look at the residual to determine if there are any systematic error in our model.

```
lmModel$residuals
```

```
##            1            2            3            4            5 
## -0.470927459  0.655436830 -0.311383042 -0.342257578  0.258414656 
##            6            7            8            9           10 
## -0.152442164  0.056812345 -0.003957604  0.215433309  0.182340433 
##           11           12           13           14           15 
##  0.131162779  0.018586186 -0.330198506  0.136109584 -0.214007960 
##           16           17           18           19           20 
##  0.215927990  0.247755893  0.179470880  0.450026314 -0.131731022 
##           21           22           23           24           25 
##  0.119100574  0.409768104  0.042591264 -0.159021512 -0.085348333 
##           26           27           28           29           30 
##  0.180611206  0.338630011 -0.196615286 -0.411639153 -0.274519767 
##           31           32           33           34           35 
## -0.600287762 -0.192792725  0.195876157 -0.427598425 -0.019376464 
##           36           37           38 
## -0.392865198  0.466854336  0.216061110
```

```
plot(lmModel$residuals, ylab = "Residual", col = "red", pch=19, main = "Residual Plot")
abline(h=0, col = "blue")
```

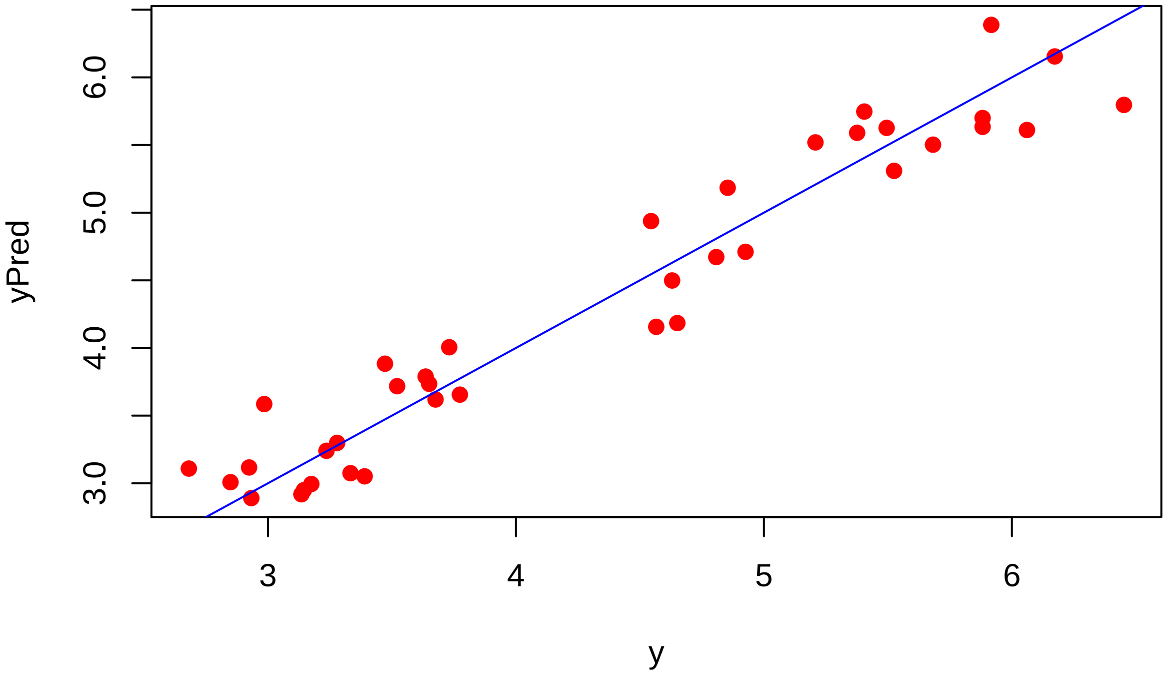


In the Residual Plot shown above, the residuals are randomly scattered around the y = 0 line. This tells that there is no systematic error in the model.

Model Prediction

I am going to make prediction using our model on the training data itself. It should be noted that prediction error on the training data is optimistic estimated of the error. For more accurate estimate of the error we should calculate error on the test data (the data that was not used in the model building). Since, I has only 38 observations, I chose not to divide the data into train and test set.

```
newData <- subset(fuelData, select = -c(MPG, ACC))
yPred <- predict(lmModel, newdata = newData)
y <- fuelData[,2]
plot(y, yPred, col = "red", pch = 19)
abline(a = 0, b = 1, col = "blue")
```



The above plot shows that the relationship between the predicted and the actual value is linear with slope 1. This once again clearly shows that the linear regression model is able to describe out data very well.