Linear Regresssion - Predicting Miles Per Gallon

V2 Maestros

Contents

Problem Statement

The input data set contains data about details of various car models. Based on the information provided, the goal is to come up with a model to predict Miles-per-gallon of a given model.

Techniques Used

1. Linear Regression (multi-variate)
2. Data Imputation

Data Engineering & Analysis

```r
setwd("C:/Personal/V2Maestros/Modules/Machine Learning Algorithms/Linear Regression")

auto_data <- read.csv("auto-miles-per-gallon.csv")

str(auto_data)
```

Loading and understanding the dataset

```
## 'data.frame':    398 obs. of  8 variables:
##  $ MPG         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ CYLINDERS   : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ DISPLACEMENT: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ HORSEPOWER  : Factor w/ 94 levels "?","100","102",..: 17 35 29 29 24 42 47 46 48 40 ...
##  $ WEIGHT      : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...
##  $ ACCELERATION: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ MODELYEAR   : int  70 70 70 70 70 70 70 70 70 70 ...
##  $ NAME        : Factor w/ 305 levels "amc ambassador brougham",..: 50 37 232 15 162 142 55 224 242
```

```r
summary(auto_data)
```

```
##       MPG         CYLINDERS     DISPLACEMENT   HORSEPOWER      WEIGHT
##  Min.   : 9.0   Min.   :3.00   Min.   : 68   150    : 22   Min.   :1613
##  1st Qu.:17.5   1st Qu.:4.00   1st Qu.:104   90     : 20   1st Qu.:2224
##  Median :23.0   Median :4.00   Median :148   88     : 19   Median :2804
##  Mean   :23.5   Mean   :5.46   Mean   :193   110    : 18   Mean   :2970
##  3rd Qu.:29.0   3rd Qu.:8.00   3rd Qu.:262   100    : 17   3rd Qu.:3608
##  Max.   :46.6   Max.   :8.00   Max.   :455   75     : 14   Max.   :5140
##                                              (Other):288
##   ACCELERATION    MODELYEAR              NAME
##  Min.   : 8.0   Min.   :70   ford pinto    :  6
##  1st Qu.:13.8   1st Qu.:73   amc matador   :  5
##  Median :15.5   Median :76   ford maverick :  5
##  Mean   :15.6   Mean   :76   toyota corolla:  5
##  3rd Qu.:17.2   3rd Qu.:79   amc gremlin   :  4
##  Max.   :24.8   Max.   :82   amc hornet    :  4
##                              (Other)       :369
```

```r
head(auto_data)
```

```
##   MPG CYLINDERS DISPLACEMENT HORSEPOWER WEIGHT ACCELERATION MODELYEAR
## 1  18         8          307        130   3504         12.0        70
## 2  15         8          350        165   3693         11.5        70
## 3  18         8          318        150   3436         11.0        70
## 4  16         8          304        150   3433         12.0        70
## 5  17         8          302        140   3449         10.5        70
## 6  15         8          429        198   4341         10.0        70
##                        NAME
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

Data Cleansing

1. The ranges of values in each of the variables (columns) look ok without any kind of outliers
2. Horsepower is a number and R should have shown the quartiles like other numeric variables. It is being recognized as factor. Also, str() shows "?" as one of the values. So this means, the ? values should be imputed. We will replace the "?" with the mean value for Horsepower.

```r
auto_data$HORSEPOWER <- as.numeric(auto_data$HORSEPOWER)
#as.numeric will have converted ? to NA
auto_data$HORSEPOWER[is.na(auto_data$HORSEPOWER)] <- mean(auto_data$HORSEPOWER, na.rm=TRUE)
summary(auto_data)
```

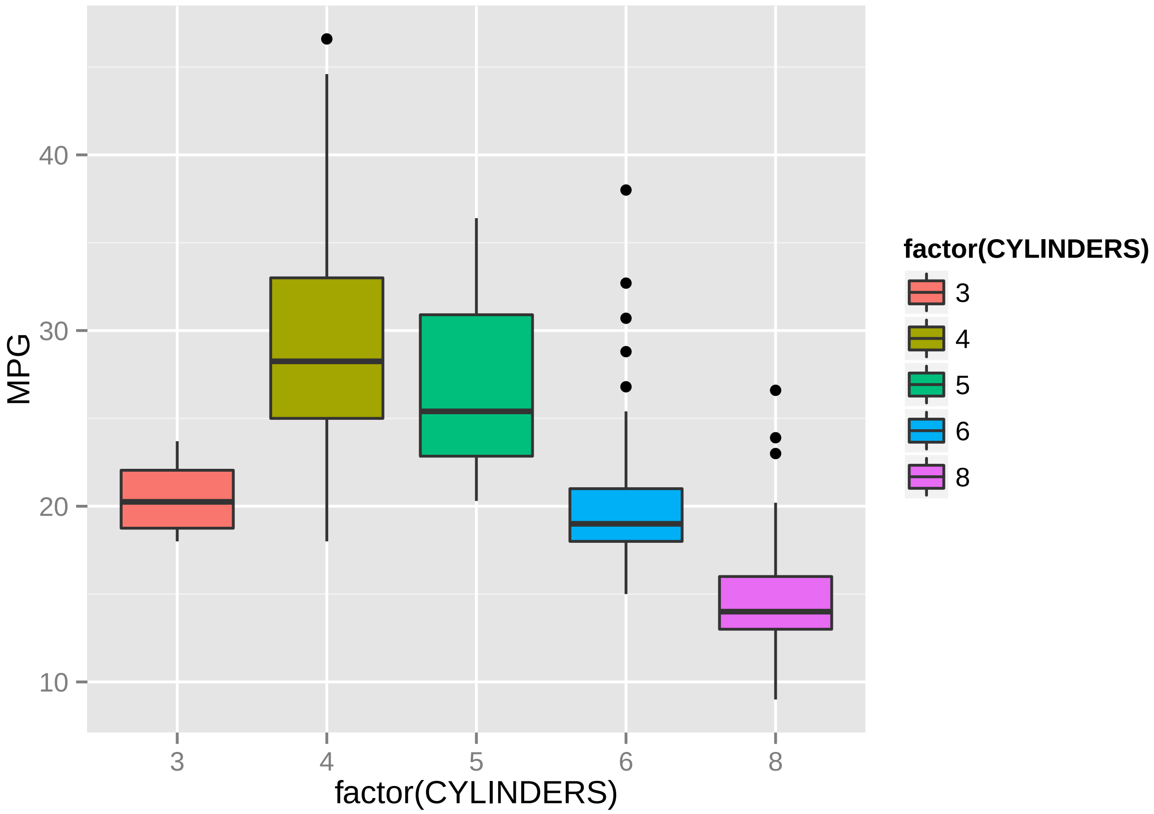
```
##       MPG         CYLINDERS     DISPLACEMENT   HORSEPOWER       WEIGHT
##  Min.   : 9.0   Min.   :3.00   Min.   : 68   Min.   : 1.0   Min.   :1613
##  1st Qu.:17.5   1st Qu.:4.00   1st Qu.:104   1st Qu.:26.0   1st Qu.:2224
##  Median :23.0   Median :4.00   Median :148   Median :60.5   Median :2804
##  Mean   :23.5   Mean   :5.46   Mean   :193   Mean   :51.4   Mean   :2970
##  3rd Qu.:29.0   3rd Qu.:8.00   3rd Qu.:262   3rd Qu.:79.0   3rd Qu.:3608
##  Max.   :46.6   Max.   :8.00   Max.   :455   Max.   :94.0   Max.   :5140
##
##   ACCELERATION    MODELYEAR              NAME
##  Min.   : 8.0   Min.   :70   ford pinto    :  6
##  1st Qu.:13.8   1st Qu.:73   amc matador   :  5
##  Median :15.5   Median :76   ford maverick :  5
##  Mean   :15.6   Mean   :76   toyota corolla:  5
##  3rd Qu.:17.2   3rd Qu.:79   amc gremlin   :  4
##  Max.   :24.8   Max.   :82   amc hornet    :  4
##                              (Other)       :369
```

```r
library(ggplot2)
```

Exploratory Data Analysis

```
## Warning: package 'ggplot2' was built under R version 3.1.1
```

```r
ggplot(auto_data, aes(factor(CYLINDERS), MPG)) +
    geom_boxplot( aes(fill=factor(CYLINDERS)))
```



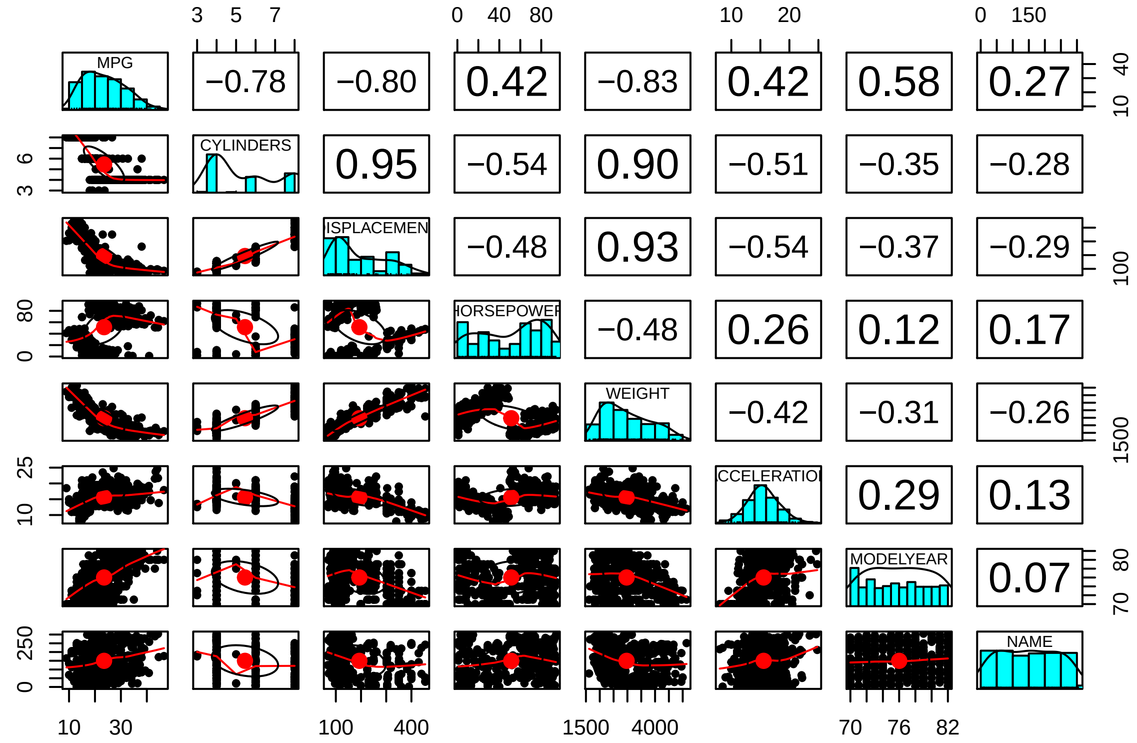
```
library(psych)
```

Correlations

```
## Warning: package 'psych' was built under R version 3.1.1
```

```
##
## Attaching package: 'psych'
##
## The following object is masked from 'package:ggplot2':
##
##     %+%
```

```
pairs.panels(auto_data)
```



Once you do correlations, it is important to find out domain (automobiles in this case) as to why they exist. In this example, we are trying to predict miles-per-gallon. The chart shows the Pearson correlation co-efficient (range -1 to + 1).

- Number of Cylinders has a high negative correlation to MPG (As Cylinders increase, MPG decreases). This is as expected.
- Same is the case with Displacement.
- The more the weight the less the acceleration. This also has a logical explanation
- Name has little correlation to MPG. True. This can be ignored.

Modeling & Prediction

```r
# ignore colume 8 - NAMES which is string. Regression works only on numbers.
lm_model <- lm( MPG ~ ., auto_data[,-8]  )
summary(lm_model)
```

```
##
## Call:
## lm(formula = MPG ~ ., data = auto_data[, -8])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.774 -2.409 -0.091  1.960 14.334
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.62e+01   4.27e+00   -3.78  0.00018 ***
## CYLINDERS    -1.02e-01   3.45e-01   -0.29  0.76821
## DISPLACEMENT  5.64e-03   7.22e-03    0.78  0.43531
## HORSEPOWER    1.02e-02   6.96e-03    1.47  0.14331
## WEIGHT       -6.85e-03   5.98e-04  -11.45  < 2e-16 ***
## ACCELERATION  7.55e-02   7.84e-02    0.96  0.33595
## MODELYEAR     7.60e-01   5.08e-02   14.96  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.44 on 391 degrees of freedom
## Multiple R-squared:  0.81,   Adjusted R-squared:  0.807
## F-statistic:  277 on 6 and 391 DF,  p-value: <2e-16
```

The model gives the Intercept and co-efficients required for the linear regression equation. The R-Squared value is .8, which is a very good fit for the problem.

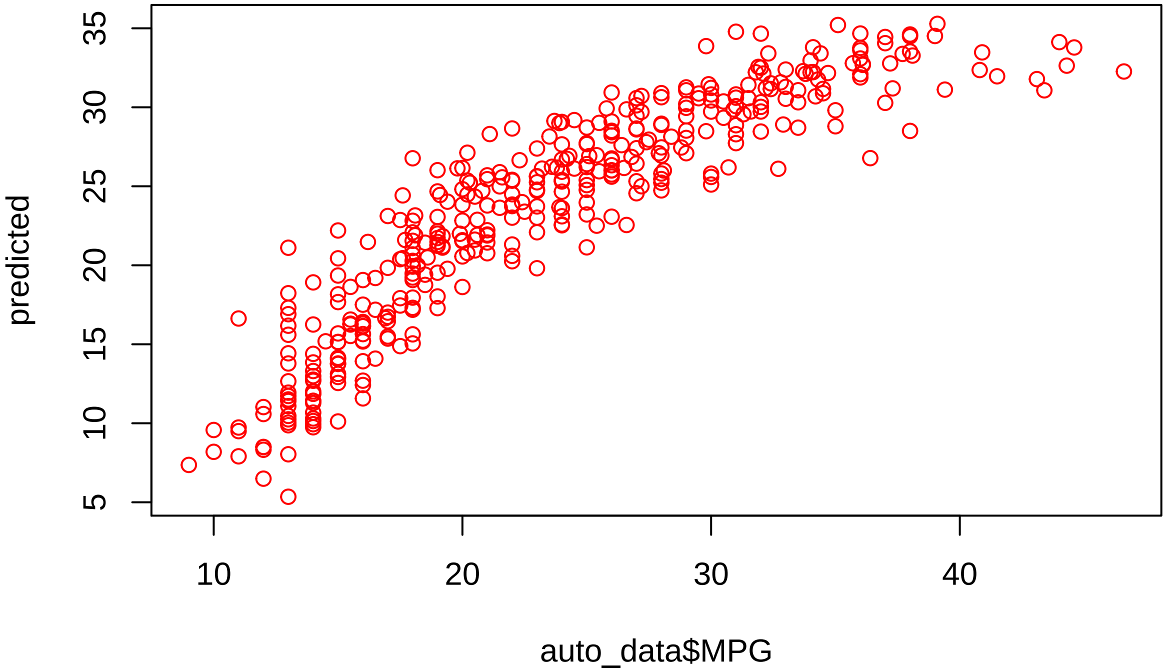
Testing

To test the accuracy of the equation, let us apply the equation on the same data set and predict the MPG for each record. Since we already know the actual value, let us compare the predicted value with the actual value and see the conformance/ error in the prediction.

```
predicted <- predict.lm(lm_model, auto_data)
summary(predicted)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.35   18.70   24.80   23.50   29.10   35.30
```

```
#plot prediction vs actuals
plot( auto_data$MPG, predicted, col="red")
```



```
#find correlation between prediction and actuals
cor(auto_data$MPG, predicted)
```

```
## [1] 0.8999
```

The plot of prediction vs actual follows a diagnal straight line, which means this is very good prediction. The correlation co-efficient is also very high, which again means very good prediction.

Conclusions

The model built can predict MPG with an accuracy of about 90% (based on the correlation co-efficient)