Regression Models Course Project

Dineshkumar Murugan

Saturday, May 23, 2015

Context

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

"Is an automatic or manual transmission better for MPG" "Quantify the MPG difference between automatic and manual transmissions"

Exploratory Data Analysis

Reading the Data

```r
data(mtcars)
```

Setting the Factors

```r
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
```

Getting summary of the Data

```r
summary(mtcars)
```

```
##       mpg        cyl        disp             hp             drat      
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760  
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080  
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695  
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597  
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920  
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930  
##        wt             qsec       vs             am     gear   carb  
##  Min.   :1.513   Min.   :14.50   0:18   Automatic:19   3:15   1: 7  
##  1st Qu.:2.581   1st Qu.:16.89   1:14   Manual   :13   4:12   2:10  
##  Median :3.325   Median :17.71                         5: 5   3: 3  
##  Mean   :3.217   Mean   :17.85                                4:10  
##  3rd Qu.:3.610   3rd Qu.:18.90                                6: 1  
##  Max.   :5.424   Max.   :22.90                                8: 1  
```

Testing for Normality

We will do two test

1. One with ad.test method from nortest package and checking if the p values is greater than 0.05
2. Another by plotting the dataset

```
library(nortest)
ad.test(mtcars$mpg)
```

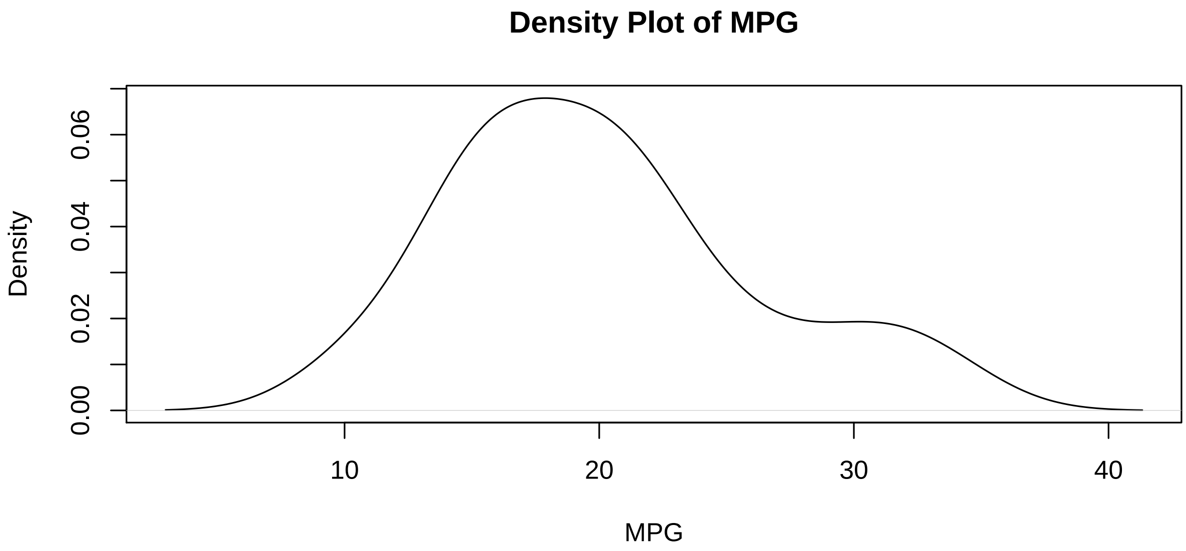
ad.test method

```
##
##  Anderson-Darling normality test
##
## data:  mtcars$mpg
## A = 0.5797, p-value = 0.1207
```

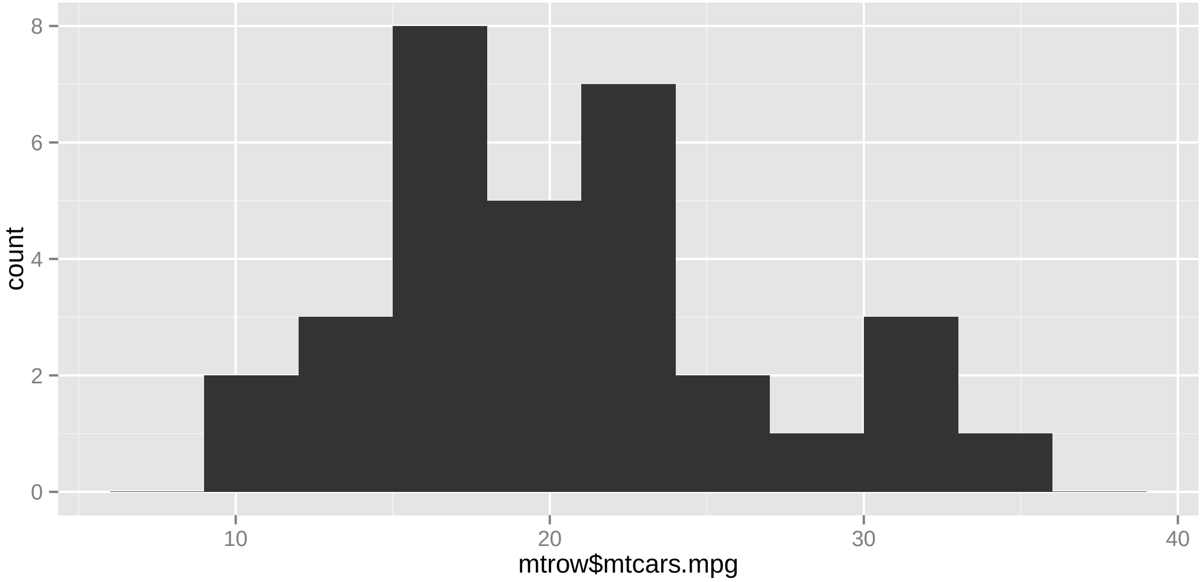
> p value is greater than 0.05

```
library(ggplot2)
d <- density(mtcars$mpg)
```

```
plot(d, xlab = "MPG", main ="Density Plot of MPG")
```



```
mtrow<-data.frame(mtcars$mpg)
ggplot(mtrow, aes(x=mtrow$mtcars.mpg)) + geom_histogram(binwidth=3)
```



> Plot looks to be normal distribution

Lets choose the predictors required for the model

Lets create a correlation matrix for all the predictors against **mpg**

```
data(mtcars)
sort(cor(mtcars)[1,])
```

```
##         wt        cyl       disp         hp       carb       qsec
## -0.8676594 -0.8521620 -0.8475514 -0.7761684 -0.5509251  0.4186840
##       gear         am         vs       drat        mpg
##  0.4802848  0.5998324  0.6640389  0.6811719  1.0000000
```

- **am** by default is included in the model
- **wt**, **cyl**, **disp**, and **hp** are highly correlated with **mpg**
- we also see that **cyl** and **disp** are highly correlated with each other, which we cannot use as a predictor

Performing Regression Analysis

Simple Linear Regression

```
Simplefit <- lm(mpg~am, data = mtcars)
summary(Simplefit)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

- on average, automatic cars have 17.147 MPG and manual transmission cars have 7.245 MPGs more
- we see that the Rˆ2 value is 0.3598
- This means that our model only explains 35.98% of the variance

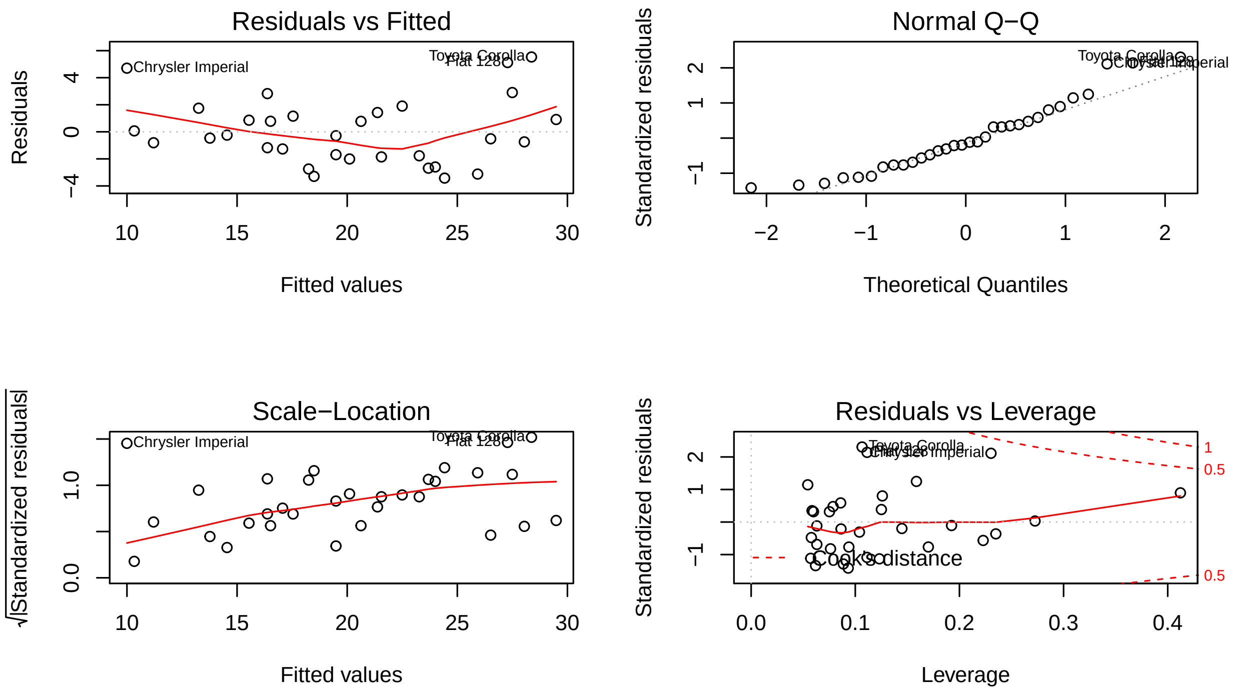
Multivariate Linear Regression

```
Multivariatefit <- lm(mpg~am + wt + hp, data = mtcars)
anova(Simplefit, Multivariatefit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     28 180.29  2    540.61 41.979 3.745e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p-value of 3.745e-09, we reject the null hypothesis and claim that our multivariate model is significantly different from our simple model. Lets check the residuals before we derive any conclution

```
par(mfrow = c(2,2))
plot(Multivariatefit)
```



They are normally distributed, so we can report on our final Model

```
summary(Multivariatefit)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## am           2.083710   1.376420   1.514 0.141268
## wt          -2.878575   0.904971  -3.181 0.003574 **
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

- This model explains over 83.99% of the variance.
- On average, manual transmission cars have 2.084 MPGs more than automatic transmission cars.