Motor Trend Car Road Tests Data Analysis

Yashu Seth

May 14, 2016

Executive Summary

Data Description

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

Data Format

Variable Details

- mpg Miles/(US) gallon
- cyl Number of cylinders
- disp Displacement (cu.in.)
- hp Gross horsepower
- drat Rear axle ratio
- wt Weight (1000 lbs)
- qsec 1/4 mile time
- vs V/S
- am Transmission (0 = automatic, 1 = manual)
- gear Number of forward gears
- carb Number of carburetors

Report Description

In this report I will analyze the mtcars data, and look for an appropriate relationship between the mpg (Miles/(US) gallon) variable and the am (Transmission (0 = automatic, 1 = manual)) variable and compare the difference between the two transmission types.

Data Manipulation

I will convert the columns in the data must be represented as factor variables.

```
data <- mtcars
data$cyl <- as.factor(data$cyl)
data$vs <- as.factor(data$vs)
data$am <- as.factor(data$am)
data$gear <- as.factor(data$gear)
data$carb <- as.factor(data$carb)
```

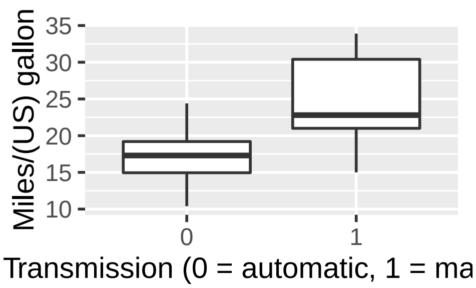
Exploratory Data Analysis

First I will make a boxplot to compare the two trasmission types.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

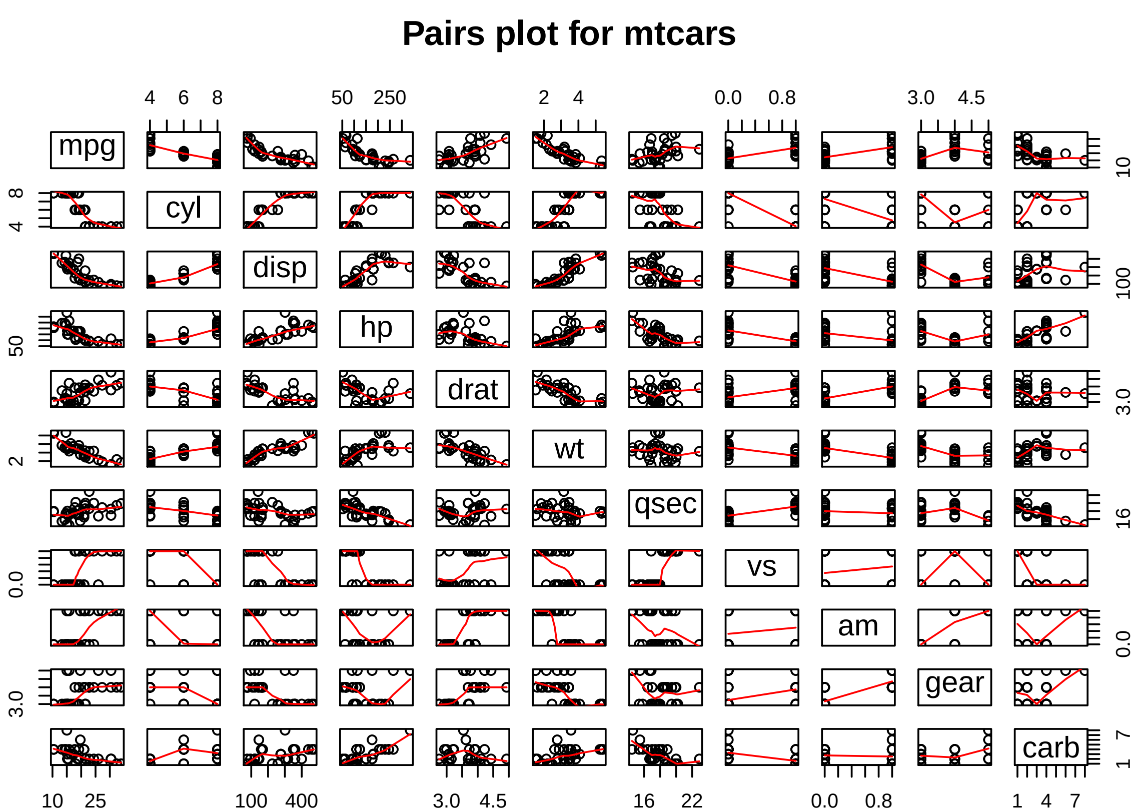
```
g <- ggplot(data, aes(x=am, y=mpg)) + geom_boxplot() +
    xlab("Transmission (0 = automatic, 1 = manual)") + ylab("Miles/(US) gallon")
g
```



This shows that manual transmission has higher values for mpg.

Now, I will make a pairs plot to infer the extent of correlation between all the variables.

```
pairs(mtcars, panel=panel.smooth, main = "Pairs plot for mtcars")
```



This shows that there exists a lot of corelation between some of the variables. We will exploit this fact later.

Regression Analysis

```
fit_am <- lm(mpg~am, data)
summary(fit_am)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = data)
##
## 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 ***
## am1            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
```

This shows that there is an increase of 7.245 mpg in manual transmission as compared to the automatic transmission when only the am variable is considered.

The adjusted R-squared value is 0.3385 which means almost 33.8% of the variance is explained by this model suggesting that this is not a very good fit. Hence we need to include more variables.

Now lets fit a model with all the variables.

```
fit_all <- lm(mpg~., data)
summary(fit_all)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
```

```
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## am1          1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

This has a residual error of 2.833 on 15 degrees of freedom. The adjsuted R-squared value is 0.779 which means that 77.9% of the variation in mpg variable is explained by this model.

Now, we will observe the pairs plot once again to see and try to remove some of the unwanted variables.

Lets see how this model behaves.

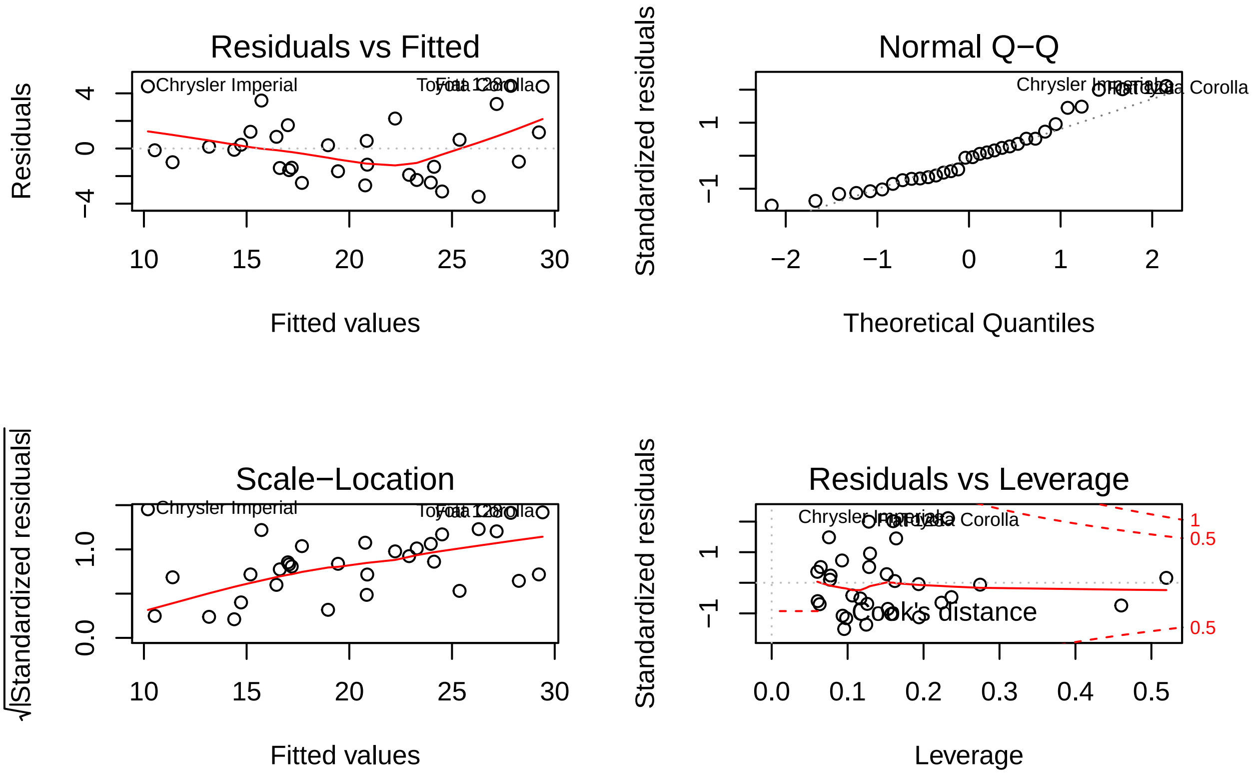
```r
fit_part <- lm(mpg ~ hp + qsec + wt + am , data)
summary(fit_part)
```

```
##
## Call:
## lm(formula = mpg ~ hp + qsec + wt + am, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4975 -1.5902 -0.1122  1.1795  4.5404
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.44019    9.31887   1.871  0.07215 .
## hp          -0.01765    0.01415  -1.247  0.22309
## qsec         0.81060    0.43887   1.847  0.07573 .
## wt          -3.23810    0.88990  -3.639  0.00114 **
## am1          2.92550    1.39715   2.094  0.04579 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.435 on 27 degrees of freedom
## Multiple R-squared:  0.8579, Adjusted R-squared:  0.8368
## F-statistic: 40.74 on 4 and 27 DF,  p-value: 4.589e-11
```

By considering the variables hp, qsec, wt, am as regressors we get 83.6% explanation of the variance from the model.

Let us make a residual plot of this model -

```r
par(mfrow = c(2,2))
plot(fit_part)
```



There is no specific pattern visible here.

Analyzing the coefficients of am variable in this model, we can say that there is increase of 2.92 in mpg when there is manual transmission present as compared to automatic transmission keeping hp, qsec and wt constant.

Hence a manual transmission is better for mpg.