Motor Trend Car Road Tests on MPG

Gordon CHAN

2015/9/24

Introduction

In this report we performed data analyse on the *mtcars* dataset to explore the relationship between a sets of variables with the miles per gallon (MPG), with the aim to answer 2 questions:

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions

Executive Summary

In this road test, 32 vehicles are tested on their Miles Per Gallon (MPG) and 11 variables are logged. From the data, a model ($mpg \sim wt + factor(cyl) + hp + am$) is fitted with MPG as the dependent variable, with 4 independent variables, weight, number of cylinders, horsepower and transmission identified.

R-packages and Dataset

The *mtcars* dataset is loaded. For the analysis the *Caret* package is used.

```
# Load libraries
library(caret)
```

```
## Loading required package: lattice
## Loading required package: ggplot2
```

```
library(GGally)
library(gridExtra)
# Load dataset
data(mtcars)
```

Exploratory Data Analysis

The *mtcars* dataset contained 11 variables of 11 models of cars tested.

```
# Dataset dimensions
dim(mtcars)
```

```
## [1] 32 11
```

From the documentation, the 11 variables are explained:

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

The pairs plot of the variables can be found in *Appendix 1*.

While **mpg** would be our dependent (outcome) variable, all others variables are potential independent variables. However, we have a special interst in **am** since it relates directly to the 2 question asked in the introduction.

Regression modelling

Linear multivariable regression was performed. A rough model is fitted with **mpg** against all other variables.

```
# Draft model1
modfit1 <- lm(mpg ~ factor(cyl) + disp + hp + drat + wt + qsec +
                  vs + am + factor(gear) + carb, data = mtcars)
```

From the summary (*Appendix 2*), we can see that only a handful of variables showed a meaningfully large coefficient. Automatic variable selection was then performed to select the relavant variables.

```
# Skeleton model
modfit0 <- lm(mpg ~ 1, data = mtcars)
# Automatic variable selection by step function
stepfit <- step(modfit0, scope=list(lower=modfit0, upper=modfit1), direction="forward", trace = 0)
summary(stepfit)
```

```
##
## Call:
## lm(formula = mpg ~ wt + factor(cyl) + hp + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.70832    2.60489  12.940 7.73e-13 ***
## wt           -2.49683    0.88559  -2.819  0.00908 **
## factor(cyl)6 -3.03134    1.40728  -2.154  0.04068 *
## factor(cyl)8 -2.16368    2.28425  -0.947  0.35225
## hp           -0.03211    0.01369  -2.345  0.02693 *
```

```
## am            1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

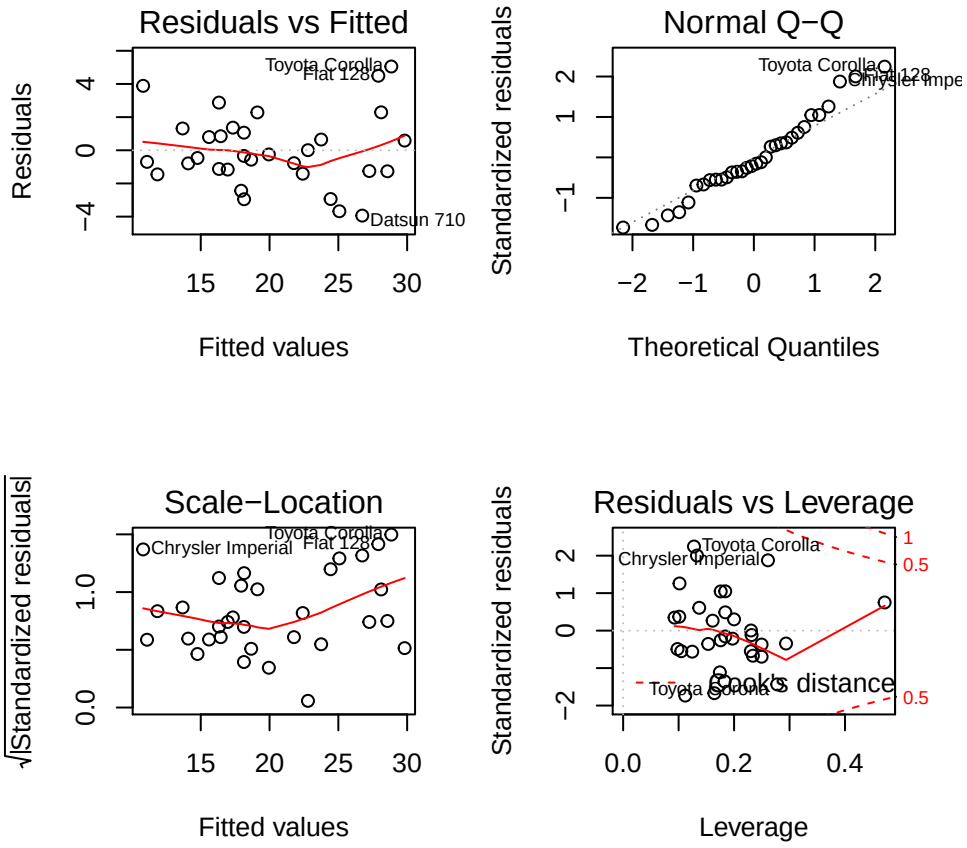
Since we wish to explore the effect of transmission (am) on the mpg, the fitted model is compared with one that has a single variable *am*. From the ANOVA result, we could see that the difference is very siginificant, hence the null hypothesis that the variables *wt*, *cyl*, and *hp* are not contributing to the variability of the model is rejected.

```
# Baseline model
modfit.am <- lm(mpg ~ am, data = mtcars)
anova(modfit.am, stepfit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + factor(cyl) + hp + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If we look at the *Residual vs Fitted* plot of the model, we could see that the points are quite evenly and randomly scattered, while for the *Normal Q-Q* plot the points are lining quite neatly along the diagonal verifying that the residuals are randomly distributed. We could also observe some outliers in the higher end of mpg.

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



Question 1: Is an automatic or manual transmission better for MPG?

In regard to the vehical transmission, the 1st question is relatively straight forward. A t-test is performed to compare the mean mpg of automatic / manual transmission group. We can see that there is a very significant difference between the groups, where cars with **manual transmission** has a higher MPG than those with automatic transmission. A boxplot can be found in Appendix 3.

```
t.test(mpg ~ am, data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

Question 2: Quantify the MPG difference between automatic and manual transmissions

Although from the t-test performed above, one might be tempted to quantify the difference between the groups by a simple subtraction. This would be incorrect as MPG is also affected by other factors as well. Here we should refer to the fitted model. From the *am* coefficient, we can conclude that cars with manual transmission get **1.80921** more MPG than an automatic.

Conclusion

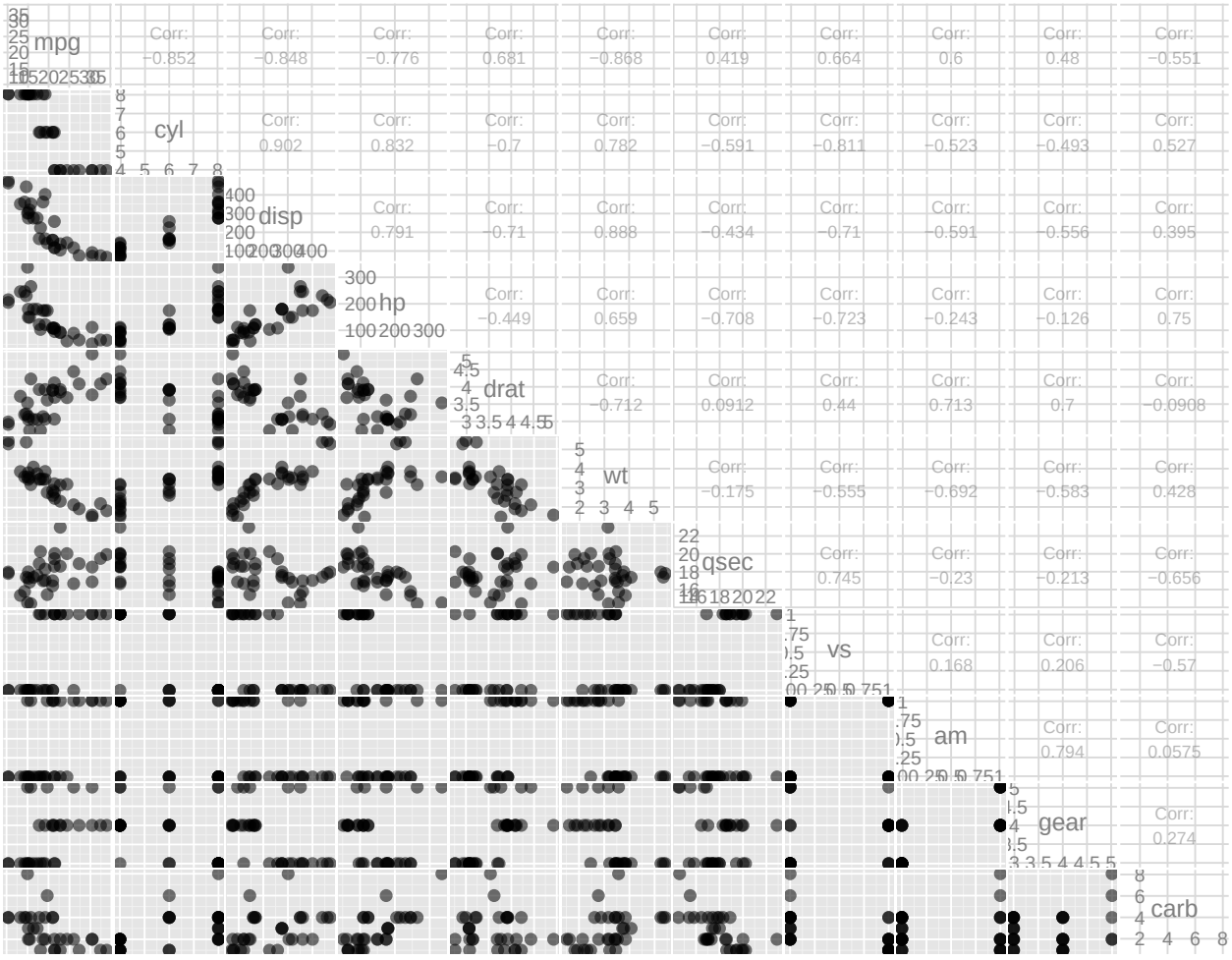
We have modelled the MPG of a given vehicle as $mpg \sim wt + factor(cyl) + hp + am$, in which it is affected by 4 variables:

1. Weight, a decrease in MPG of **2.49683** is expected per every increase of 1000lb in weight.
2. Number of cylinders, compared with those with 4 cylinders, a decrease in MPG of **3.03134** is expected for vehicle with 6 cylinders, and a decrease of **2.16368** for vehicle with 8 cylinders.
3. Horsepower, a decrease in MPG of **0.03211** is expected per every increase of 1 gross horsepower.
4. Transmission, an increase in MPG of **1.80921** is expected for manual vehicles.

Appendicies

Appendix 1: Pairs plot of variables in *mtcars*

```
# Pairs plot of mtcars
ggpairs(mtcars,
        title = "Variables of mtcars dataset",
        size = list(corSize = 10, size = 2),
        alpha = 0.8,
        axisLabels = "internal")
```



Appendix 2: Coefficient of fitted models

```
summary(modfit1)$coef
```

```
##                   Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept)   15.09261548 17.13627433  0.8807408 0.38946336
## factor(cyl)6  -1.19939698  2.38736481 -0.5023937 0.62116357
## factor(cyl)8   3.05491692  4.82986776  0.6325053 0.53459525
## disp           0.01256810  0.01774024  0.7084518 0.48726645
## hp            -0.05711722  0.03174603 -1.7991927 0.08789210
## drat           0.73576811  1.98461241  0.3707364 0.71493502
## wt            -3.54511861  1.90895437 -1.8570997 0.07886857
## qsec           0.76801287  0.75221895  1.0209964 0.32008122
## vs             2.48849171  2.54014636  0.9796647 0.33956206
## am             3.34735713  2.28948094  1.4620594 0.16006890
## factor(gear)4 -0.99921782  2.94657533 -0.3391116 0.73824498
## factor(gear)5  1.06454635  3.02729599  0.3516492 0.72897110
## carb           0.78702815  1.03599487  0.7596834 0.45676696
```

Apendix 3: Boxplot of MPG vs Transmissions

```
ggplot(data=mtcars, aes(am, mpg))+
    geom_boxplot(aes(fill=as.factor(am)))+
    labs(title ="Boxplot of Mean MPG vs Transmission", x = "Transmission", y = "MPG")+
    scale_fill_discrete(name = "Transmission", labels = c("Automatic", "Manual"))
```