# Analysis of Mileage dependence on Transmission

*Unmesh Phadke*

*25 July,2015*

## Executive Summary

In this report we have analysed the mtcars dataset which was present in the 1974 Motor Trend Us magazine. The follwing questions about the data are answered: 1)Is an automatic or manual transmission better for MPG? 2)Quantifying the MPG difference between automatic and manual transmissions It is found that on an average cars perform well on manual transmission with about a differnce of 7 in mpg(But the multivariate regression model which accoubts for effects of other variables tells the real story and predicts only a difference ofabout 1.4).We fit a model by choosing appropriate variables by looking at their correlation with response variable.The model succeeds in explaining 85 % of the variance in the data. A Welch t test confirms better performance of manual transmission.

## Exploratory Data Analysis

To do some basic exploratory data analysis we plot the the mpg values depending upon their transmission type in a simple scatter plot as shown below.Referring to plot1 in appendix it can be observed than in general, that the average mileage appears to be lower for the automatic variance case and higher fot the manual transmission case. But, the variance for the manual transmission case appears to be higher.

## Modelling

While modelling it is important to choose the correct variables to use in the model. The response variable in this case is the mpg variable. Variables that would be used as the explanatory variables should in general be correlated with the response variable in order to fit a linear model. Also, no two explanatory variables should be correlated with each other as this may result in redundancy and may result in a bad model. Keeping these facts in mind, we now do some basic model searching.

First, we will do the model with only am as the explanatory variable as that is by default(our analysis is based on the am variable)

```
fit1<-lm(data=mtcars,mpg~factor(am))
summary(fit1)
```

```
## 
## Call:
## lm(formula = mpg ~ factor(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 ***
## factor(am)1    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
```

The coefficient for the am variable is 7.245. This means that the mean mpg for manual cars is larger than that for automatic cars by about 7.24. The standard errros for all the coefficients are low as compared to the coefficients themselves and this is reflected in the t-values which are greater than 1. The p-values are significantly lower than 0.05. The F-statistics represents degeree of correlation between the response variable and teh predictor variable. The F-statistic is giher than 1 with avery low P-value which implies that the mpg variable is correlated with the am variable and the linear model is not frivolous. The R-squared value though is nly about 35.98 percent. This means that the model explains only about 35.98 percent variance in the data and so it isnt quite a good model. So we should look at including some more predictor variables maybe.

To choose some other predictor variables we look at their correlation with the mpg variable. To do this we take a row from the correlation matrix. We also take a look at the variation inflation factors if all the variables were included in our model.

```
cor(mtcars)[1,];fit<-lm(data=mtcars,mpg~.)
```

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

Looking at the above result, the variables cyl,wt,disp,hp all have a high correlation with mpg(greater than 0.7). But we also need to look at the correlation among the predictor variables.

```
with(mtcars,cor(cyl,wt));with(mtcars,cor(cyl,disp));with(mtcars,cor(cyl,hp));with(mtcar
s,cor(wt,disp));with(mtcars,cor(hp,wt));with(mtcars,cor(disp,hp))
```

```
## [1] 0.7824958
```

```
## [1] 0.9020329
## [1] 0.8324475
```

```
## [1] 0.8879799
```

```
## [1] 0.6587479
```

```
## [1] 0.7909486
```

Note that cyl and disp and wt and disp are highly correlated with each other. So we wont include disp in our model as it is highly correlated with both cyl and wt and this will make our model redundant. So, now we construct a new model with the cyl,wt,hp and am(of course!) as the explanatory or the predicotr variables.

```
fit2<-lm(data=mtcars,mpg~am+cyl+wt+hp)
summary(fit2)
```

```
## 
## Call:
## lm(formula = mpg ~ am + cyl + wt + hp, data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.4765 -1.8471 -0.5544  1.2758  5.6608 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 36.14654    3.10478  11.642 4.94e-12 ***
## am           1.47805    1.44115   1.026   0.3142    
## cyl         -0.74516    0.58279  -1.279   0.2119    
## wt          -2.60648    0.91984  -2.834   0.0086 ** 
## hp          -0.02495    0.01365  -1.828   0.0786 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.509 on 27 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.8267 
## F-statistic: 37.96 on 4 and 27 DF,  p-value: 1.025e-10
```

Note that the coefficients for am and cyl have pretty high std errors. The t-values are also close to 1 and the P-value appreciably greater than 0.05. This means that the mean mpg for manual is higher than that for automatic by a number in the range (1.478-1.44,1.47+1.44)=(0.038,2.9). Though this interval doesnt contain zero, the lower bound is periliously close to zero ,so using this model we canot say firmly that manual outperforms automatic in mpg.

But, considering the R-squared value of 85% almost, we can say that this model explains the variance in the mpg data though some minor improvements could be made. But its definitely better than our previous model which had only ‘am’ as the predictor variable. The F-statistic is very high and the corresponding P-value extremely low,which suggests that the correlation is good.

# Comparing the models

We now use ‘anova’ function to compare the two models.

```
anova(fit1,fit2)
```

```
## Analysis of Variance Table
## 
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ am + cyl + wt + hp
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)    
## 1     30 720.9                                  
## 2     27 170.0  3     550.9 29.166 1.274e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The P-value in the last column is very very small which is strong evidence that the second model is better than the first one.So we will probably choose the second model for this analysis.

# Diagnostics

We will plot some plots which will give some information about the residuals of our second model. For the plots refer to plot 3 in the Appendix.

From the above plots we can make the following deductions: 1> The points are randomnly distributed in the Residuals vs Fitted plot. 2. In the Normal Q-Q plot all the points fall almost along a line verifyinh that the residuals follow a normal distribution.

We will now perform some basic diagnostic tests to know what measures we could take to imporve the model.

```
dat <- dfbetas(fit2);tail(sort(dat[,2]),3);head(sort(dat[,2]),3)
```

```
## Chrysler Imperial          Fiat 128     Toyota Corona 
##         0.3430597         0.4211007         0.6795213 
```

```
##     Mazda RX4 Mazda RX4 Wag    Volvo 142E 
##    -0.3640506    -0.3251406    -0.2843548 
```

The dfbetas for a point gives the change in the coefficients when that point is not considered while fitting the model. From the model we observe that there is maximum increase in coefficient for ‘am’ if we remove the cars ‘Chrysler Imperial’ , ‘Fiat 128’ and ‘Toyota Corona’ and minimum three are ‘Mazda RX4’, ‘Mazda RX4 Wag’ and ‘Volvo142E’.

Now we find the maximum leverages.

```
tail(sort(hatvalues(fit2)),3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##           0.2770490           0.2846012           0.4661016
```

# An important Test.

We now need to test the hypothesis that the mean mpg for “Automatic” is greater than the mean mpg for “Manual”.

```
t.test(df2$mpg,df1$mpg,paired=FALSE,var.equal=FALSE)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  df2$mpg and df1$mpg
## 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:
##   3.209684 11.280194
## sample estimates:
## mean of x mean of y 
##  24.39231  17.14737
```

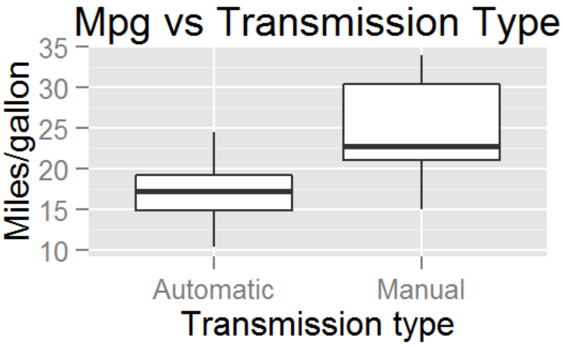
Note that the P-value is extremely small and hence we can say that we can reject the null Hypothesis that the two means are equal. Hence the two means are unequal and the difference between the means is an interval greater than zero.

# Inference

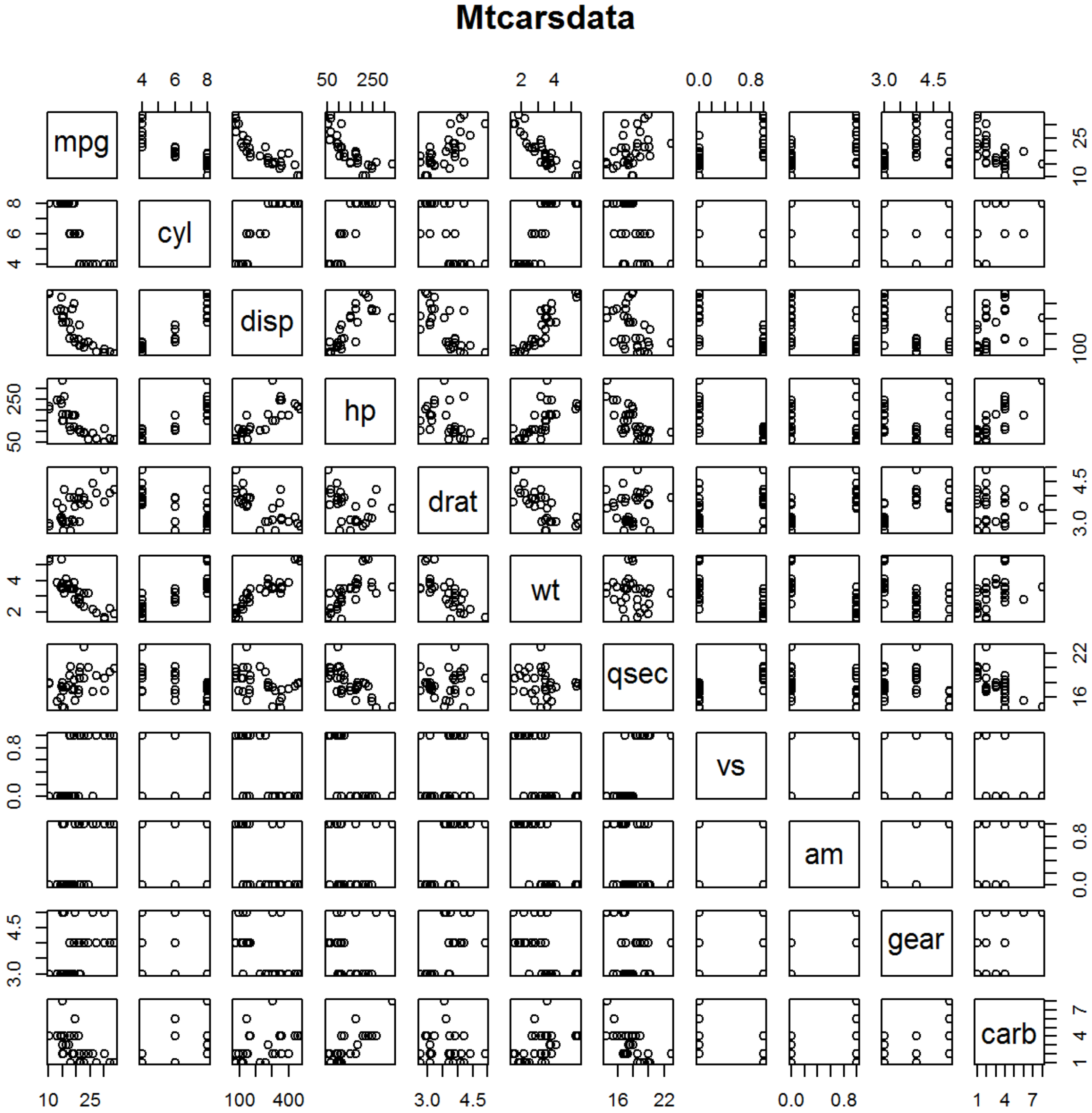
By using a linear model with the explanatory variables as ‘am’,‘cyl’,‘wt’,‘hp’ we can say that the mean mpg for manual transmission is better than the mean mpg for automatic transmission by an average of 1.48 approximately. Also, this model explains about 85 % of the variance in the data and so its apretty good model. For every increase in cylinder the mpg decreases by about 0.75. For every 1000 lbs increase in weight the mpg decreases by2.6. There is negligible change in mpg with change in the Hirse Power of the car.

# Appendix

# Plot 1


Mpg vs Transmission Type
Miles/gallon
35
30
25
20
15
10
Automatic
Manual
Transmission type

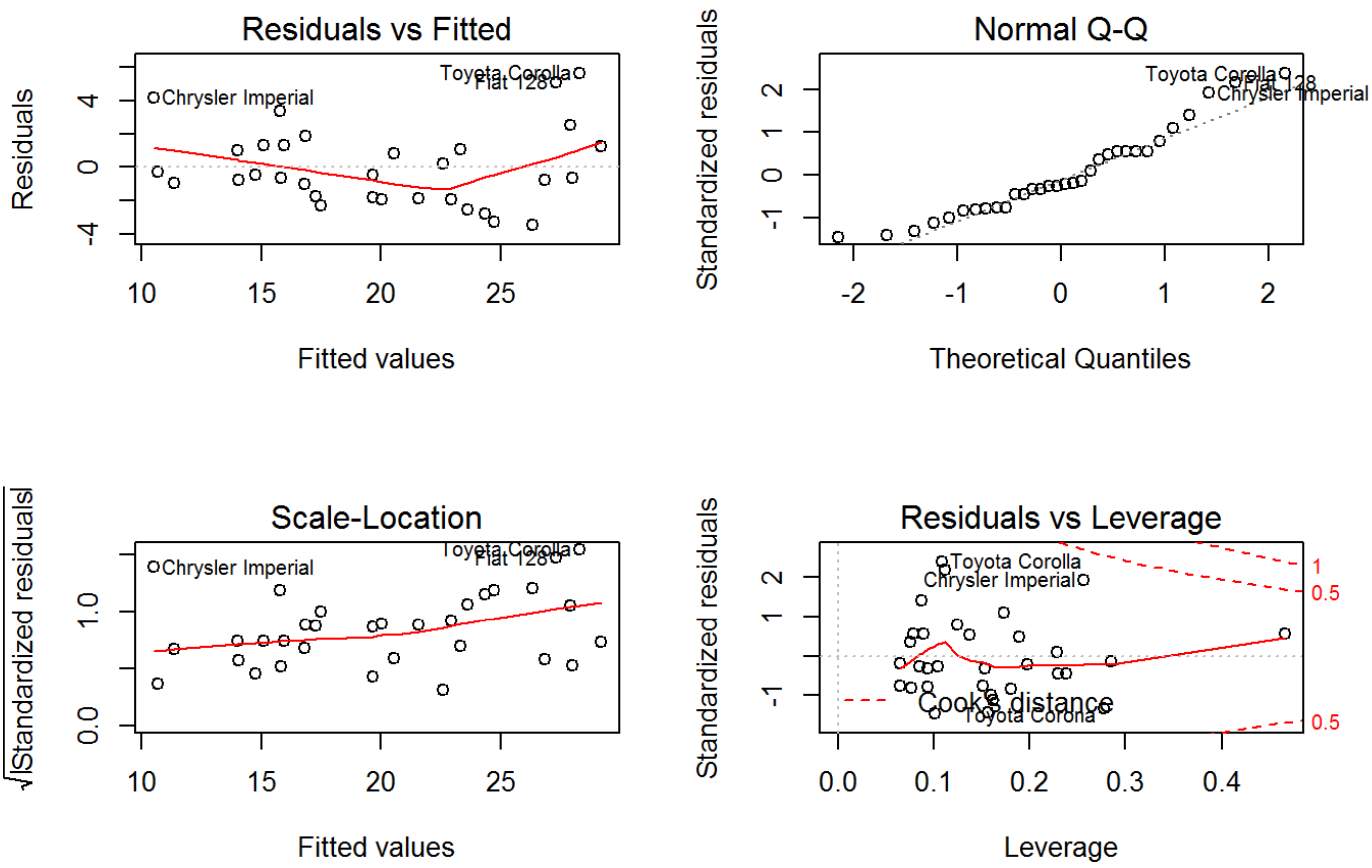
# Plot2


Mtcarsdata
mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb
4 6 8
50 250
2 4
0.0 0.8
3.0 4.5
10 25
100
3.0 4.5
16 22
0.0 0.8
1 4 7
10 25
100 400
3.0 4.5
16 22
0.0 0.8
1 4 7

# Plot3

# Diagnostics and Residuals Plot

```
par(mfrow=c(2,2));plot(fit2)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
-4
0
4
10
15
20
25
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial
-1
0
1
2
-2
-1
0
1
2
Scale-Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
0.0
1.0
10
15
20
25
Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5
-1
1
2
0.0
0.1
0.2
0.3
0.4