# MTCars Exploration

*Carsten Ersch*

*16 February 2017*

## Summary

This short data analysis focusses on the question whether manual or automatic cars have a lower mpg. On a first look at the data this seems verified as automatic cars are shown to have lower mpg than manual cars. In a more detailed data analysis, however, it was shown that this difference can be solely explained by the weight and horse power.

## Exploratory Data Analysis

```
require(datasets);data("mtcars")
mtcars$Transmission <- ifelse(mtcars$am ==1 , "Manual","Automatic")
mtcars$am <- NULL
require(ggplot2);require(GGally)
```

The variables from the dataset are explained in the relevant help file.

As shown in the correlation plot in the appendix the mpg data is clearly correlated with some of the variables such as the number of cylinders, displacement, horsepower, rear axle ratio, etc. Some of these variables, however, are also correlated with each other which should be noted and taken into account during the subsequent analysis. What is also visible from the data below is that on a first look, the data for the automatic cars (shown in red) seem to have a clearly lower mpg as the manual cars and in some of the relationships between mpg and other variables data for manual and automatic cars are clearly distinguishable.

### Fitting linear regression models

From the exploratory data analysis, most relationships seemed linear which might justify using simple linear models.

Lets first inspect the difference between automatic and manual cars without including other variables. The result is shown below which shows that automatic cars have significantly more mpg than manual cars. Based on this first look at the data the difference between manual cars and automatic cars is in the order of magnitude around 10 mpg.

```
require(lme4)
fit <- lm(mpg ~ Transmission, data = mtcars)
summary(fit)$coeff
```

```
##                     Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)        17.147368   1.124603 15.247492 1.133983e-15
## TransmissionManual  7.244939   1.764422  4.106127 2.850207e-04
```

As shown in the correlation plot in the appendix, some other factors might impact on the relationship between transmission and mpg. From the data analysis above and logic behind the variables the horsepower and weight might be the most relevant ones. Lets fit one model including each and one with both. and comapre them to the model which did not take into account.

The results below (details in the appendix 2) show that once the horse power and the weight are included in the model, the effect of transmission is partially explained by these variables. Adding both of these variables can account for most of the variation in the data and the effect of transmission becomes insignificant. This last model therefore seems to be most relevant as it explains most of the variation without including the transmission which to some extend seems logical as the weight and horsepower should have a much larger effect on the fuel usage than the type of transmission.

```
require(knitr)
```

```
## Loading required package: knitr
```

```
## Warning: package 'knitr' was built under R version 3.3.2
```

```
fit3 <- lm(mpg ~ Transmission * hp , data = mtcars)
fit4 <- lm(mpg ~ Transmission  + wt, data = mtcars)
fit5 <- lm(mpg ~ Transmission  + wt + hp, data = mtcars)
anova(fit,fit4,fit3,fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ Transmission
## Model 2: mpg ~ Transmission + wt
## Model 3: mpg ~ Transmission * hp
## Model 4: mpg ~ Transmission + wt + hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 50.4908 9.893e-08 ***
## 3     28 245.43  1     32.89  3.7517    0.0629 .
## 4     28 180.29  0     65.14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the residuals plot shown in appendix 3 no clear patterns could be detected. Looking at the relative hat values and ddfit values it seems that especially the crysler imperial and Maserati Bora have a large impact on the final fit which seems logical given that these are pretty heavy and high horsepower cars with a very low mpg compared to the rest of the cars.

```
kable(as.data.frame(tail(sort(hatvalues(fit5)/mean(hatvalues(fit5))),5)))
```

| | tail(sort(hatvalues(fit5)/mean(hatvalues(fit5))), 5) |
|---|---:|
| Ford Pantera L | 1.782337 |
| Chrysler Imperial | 1.842595 |
| Cadillac Fleetwood | 1.879711 |
| Lincoln Continental | 2.180678 |
| Maserati Bora | 3.297574 |

```
kable(as.data.frame(tail(sort(abs(dffits(fit5))),5)))
```

| | tail(sort(abs(dffits(fit5))), 5) |
|---|---:|
| Lotus Europa | 0.5475738 |
| Maserati Bora | 0.7485464 |
| Fiat 128 | 0.8134942 |
| Toyota Corolla | 0.8688907 |
| Chrysler Imperial | 1.2378250 |

# Appendix 1 Full correlation plot

```r
ggpairs(mtcars, mapping = aes(color = Transmission),
  upper = list( continuous = wrap("cor", size = 1.5, alignPercent = 1)))
```

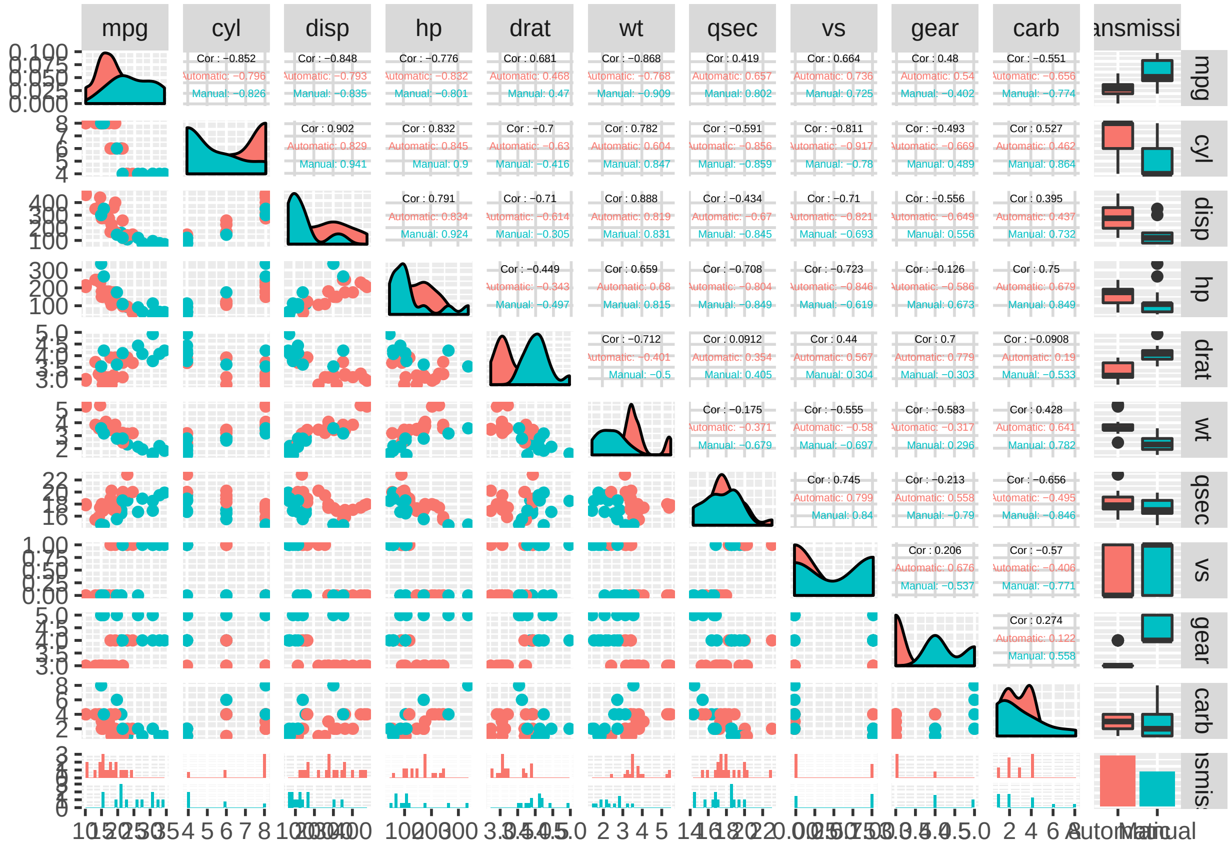


Figure 1: Correlation Plot for MPG Dataset

# Appendix 2 Model Details

```r
kable(summary(fit3)$coeff)
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 26.6248479 | 2.1829432 | 12.1967662 | 0.0000000 |
| TransmissionManual | 5.2176534 | 2.6650931 | 1.9577753 | 0.0602900 |
| hp | -0.0591370 | 0.0129449 | -4.5683758 | 0.0000902 |
| TransmissionManual:hp | 0.0004029 | 0.0164602 | 0.0244766 | 0.9806460 |

```r
kable(summary(fit4)$coeff)
```

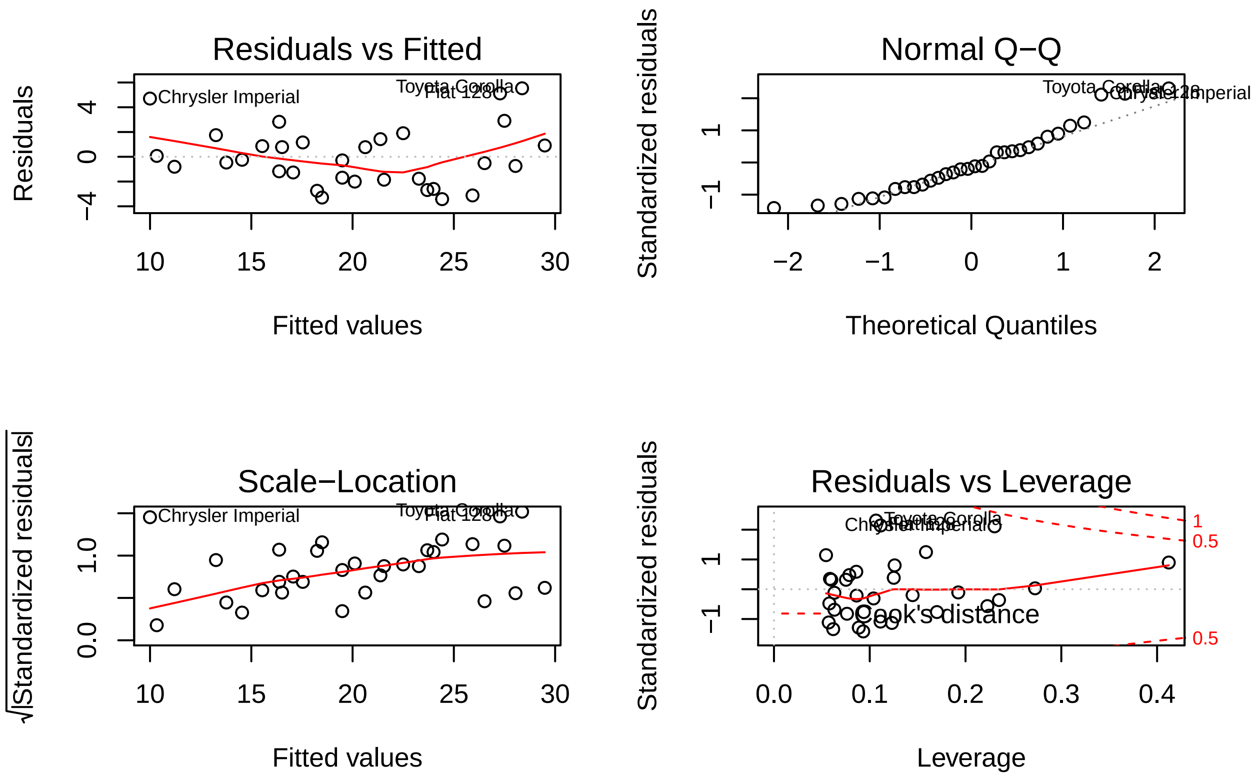
|  | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 37.3215513 | 3.0546385 | 12.2179928 | 0.0000000 |
| TransmissionManual | -0.0236152 | 1.5456453 | -0.0152786 | 0.9879146 |
| wt | -5.3528114 | 0.7882438 | -6.7908072 | 0.0000002 |

```r
kable(summary(fit5)$coeff)
```

|  | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 34.0028751 | 2.6426593 | 12.866916 | 0.0000000 |
| TransmissionManual | 2.0837101 | 1.3764202 | 1.513862 | 0.1412682 |
| wt | -2.8785754 | 0.9049705 | -3.180850 | 0.0035740 |
| hp | -0.0374787 | 0.0096054 | -3.901830 | 0.0005464 |

## Apendix 3 Residuals Plot

```r
par(mfrow = c(2, 2))
plot(fit5)
```



Residuals vs Fitted
Residuals
−4
0
4
Chrysler Imperial
Toyota Corolla
Fiat 128
10
15
20
25
30
Fitted values
Normal Q−Q
Standardized residuals
−1
1
Toyota Corolla
Chrysler Imperial
Fiat 128
−2
−1
0
1
2
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
0.0
1.0
Chrysler Imperial
Toyota Corolla
Fiat 128
10
15
20
25
30
Fitted values
Residuals vs Leverage
Standardized residuals
−1
1
Chrysler Imperial
Toyota Corolla
Fiat 128
1
0.5
0.5
Cook's distance
0.0
0.1
0.2
0.3
0.4
Leverage