Regression Models Course Project - Motor Trend

Rohit Padebettu

October 2, 2016

Executive Summary

In this exercise we try to explore, model and analyze the MTCARS dataset in order to quantify the relationship between mpg and the type of transmission. Although initial exploration of the data suggested that there might exist a significant difference in mpg observed in a vehicle depending on type of transmission (auto/manual), on further exploration and adjustment of the model for weight and initial acceleration of the car, we find that we are not able to quantify the difference in mpg due to the type of transmission

Loading and Exploring the Data

We begin by loading the libraries and exploring the data needed for the analysis. We factorize and label some of the data columns for later work.

```r
library(ggplot2)
library(car)
?mtcars

# Factoring and labelling some of the data
 mtcars$cyl <-as.factor(mtcars$cyl)
 mtcars$am <-as.factor(mtcars$am)
 mtcars$am <-relevel(mtcars$am,"1")
 translabs<-c("0" = "Auto","1" = "Manual")
```

We proceed to generate a box plot to visualize the relationship between `mpg` and type of transmission. We can see from the **Plot 1** in the **Appendix** that there seemingly exists a significant difference between distribution of mpgs for Manual Transmission vehicles (1) and Auto Transmission Vehicles(0)

We also explore the relationship between other variables like Weight and number of cylinders to better understand their impact on mpg. We show this in **Plot 2** of the **Appendix**

Fitting Models

In order to model the data to better predict the outcome `mpg` we use linear regression models. We specifically begin by using a full regression model where all the variables in the dataset `mtcars` are considered to be predictors.

```r
full <- lm(mpg ~ .,mtcars)
```

We then step through various regression models hierarchically, using the `step()` function backwards, dropping one variable at a time. Each time we assess the effect of the dropped variable via reduction in Akaike's Information Criteria (`AIC`). We tabulate this through the `ANOVA` analysis below

```r
search <-step(full,direction = "backward",trace = FALSE)
search$anova
```

```
##     Step Df     Deviance Resid. Df Resid. Dev      AIC
## 1         NA           NA        20   133.3235 69.66535
## 2 - drat  1  0.001646814        21   133.3251 67.66575
## 3 - gear  1  1.857511109        22   135.1826 66.10850
## 4   - vs  1  4.250437656        23   139.4330 65.09916
## 5 - carb  1  2.897542867        24   142.3306 63.75733
## 6 - disp  1  1.651140725        25   143.9817 62.12642
## 7  - cyl  2 16.084729691        27   160.0665 61.51530
## 8   - hp  1  9.219469347        28   169.2859 61.30730
```

The Best Fit model

From the above analysis, we determine the best model to be one where the predictors are Weight given by `wt`, Transmission given by `am` and time to 1/4 mile given by `qsec`. We model this by removing the intercept term to allow for direct interpretation of the coefficients.

```
mdl <-lm(mpg ~ wt+am+qsec-1 ,mtcars)
 summary(mdl)$coef
```

```
##       Estimate Std. Error   t value     Pr(>|t|)
## wt   -3.916504  0.7112016 -5.506882 6.952711e-06
## am1  12.553618  6.0573391  2.072464 4.754335e-02
## am0   9.617781  6.9595930  1.381946 1.779152e-01
## qsec  1.225886  0.2886696  4.246676 2.161737e-04
```

We also try to simplify the model further, which we detail in the **Appendix** under the **Other Models** section, but didn't proceed with them due to worsening `AIC`.

Interpretation of Coefficients

The coefficient estimates show that, all else equal

- An **increase in weight** of 1000lbs leads to **decrease in mpg of 3.9165**
- Auto Transmission vehicles have a lower but positive (9.6178) mean contribution to mpg but with a 95% confidence interval of (-4.64 to 23.87). *As this interval contains zero, it cannot be confirmed that Auto Transmission has a positive influence on mpg by itself*
- Manual Transmission vehicles have a higher positive (12.5536) mean contribution to mpg with a 95% confidence interval of (0.15 to 24.96). *However since this interval doesn't contain zero we can say with 95% confidence that influence on mpg of Manual Transmission is positive*
- Coefficient of `qsec` indicates that the mpg improves by 1.226 for every one sec increase in time required for first 1/4 mile i.e. **slower cars are better for mpg**

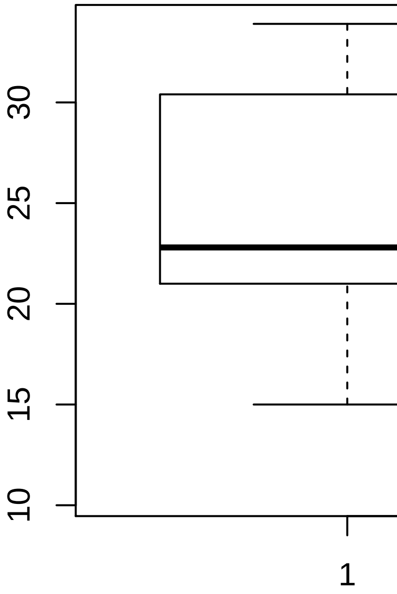
Diagnostics of the Model

We run a few standard diagnostics on the model we selected to not only ensure that the model doesn't violate any normality assumptions, but also to see if any specific outliers in the dataset are contributing to worsen the model. These **Diagnostic Plots** are listed in the **Appendix**

Appendix

Plot 1

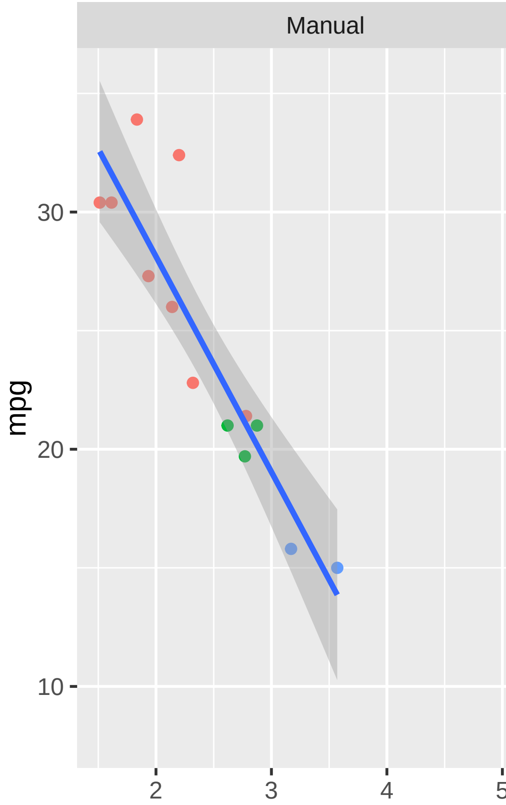
```
plot(mtcars$am,mtcars$mpg)
```



Boxplot of MPG distribution split by type of Transmission (Auto:0, Manual:1)

Plot 2

```
g<-ggplot(mtcars,aes(x=wt,y=mpg))+
     geom_point(aes(color = cyl))+
     geom_smooth(method = "lm") +
     facet_grid(~am, labeller = as_labeller(translabs) )
 print(g)
```



Relationship between MPG and Weight broken down by Transmission

Other Models

We considered a few other models simplifying the selected model further, but we rejected all of them because 1. They worsened the `AIC` and 2. They biased the remaining `wt` variable further leading us to beleive `wt` has a greater impact on `mpg`.

```
mdl2 <-lm(mpg ~ wt+am-1 ,mtcars)
extractAIC(mdl2)
```

```
## [1]  3.00000 75.21711
```

```
summary(mdl2)$coef
```

```
##      Estimate Std. Error   t value     Pr(>|t|)
## wt  -5.352811  0.7882438 -6.790807 1.867415e-07
## am1 37.297936  2.0856607 17.883032 3.326182e-17
## am0 37.321551  3.0546385 12.217993 5.843477e-13
```

```
mdl3 <-lm(mpg ~ wt+qsec-1 ,mtcars)
extractAIC(mdl3)
```

```
## [1]  2.00000 74.61398
```

```
summary(mdl3)$coef
```

```
##       Estimate Std. Error   t value     Pr(>|t|)
## wt   -4.222137  0.5171518 -8.164213 4.102362e-09
## qsec  1.878200  0.0968347 19.395936 1.588649e-18
```

```
mdl4 <-lm(mpg ~ am-1 ,mtcars)
extractAIC(mdl4)
```

```
## [1]   2.0000 103.6723
```

```
summary(mdl4)$coef
```

```
##     Estimate Std. Error  t value     Pr(>|t|)
## am1 24.39231   1.359578 17.94109 1.376283e-17
## am0 17.14737   1.124603 15.24749 1.133983e-15
```

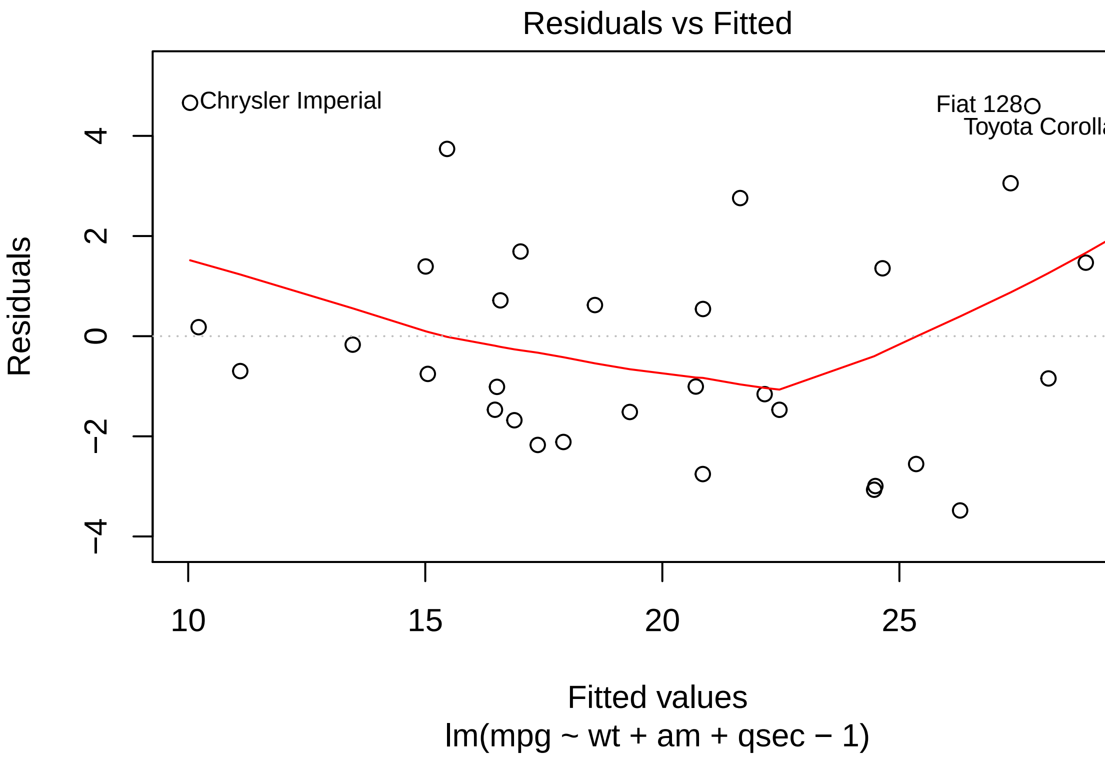
```
anova(mdl,mdl2,mdl3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + am + qsec - 1
## Model 2: mpg ~ wt + am - 1
## Model 3: mpg ~ wt + qsec - 1
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     28 169.29
## 2     29 278.32 -1  -109.034 18.034 0.0002162 ***
## 3     30 290.74 -1   -12.418  2.054 0.1628786
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diagnostic Plots

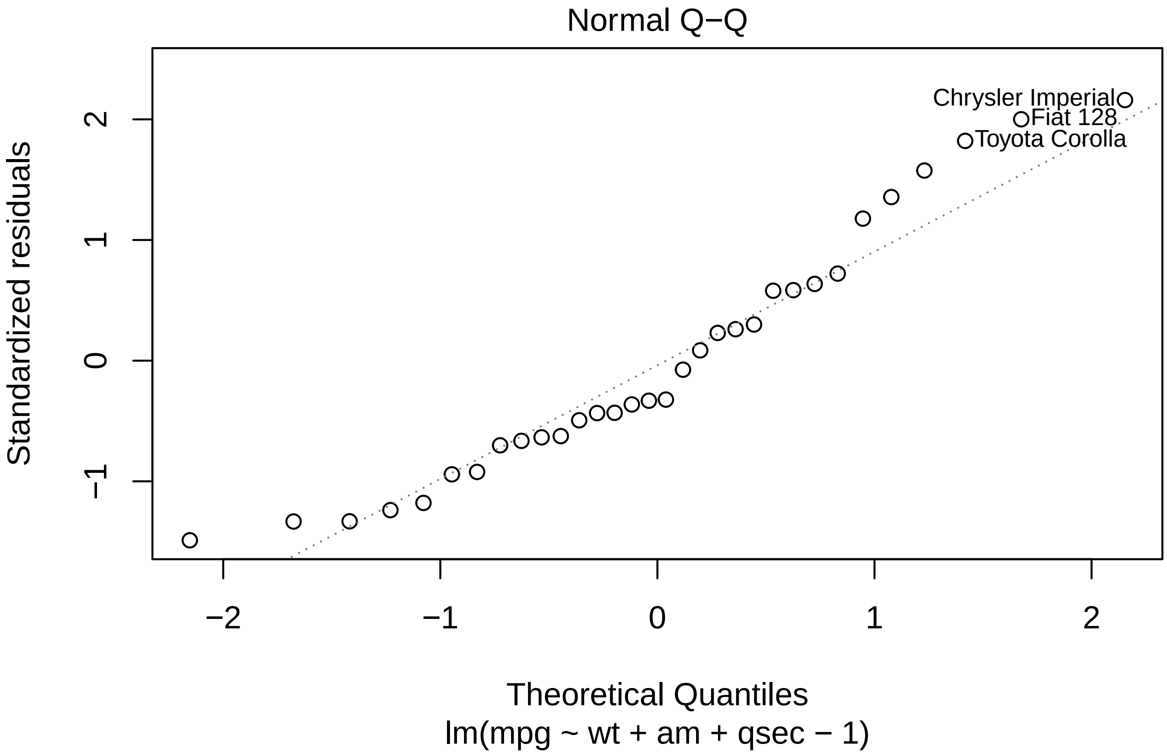
```


plot(mdl, which = 1)
```



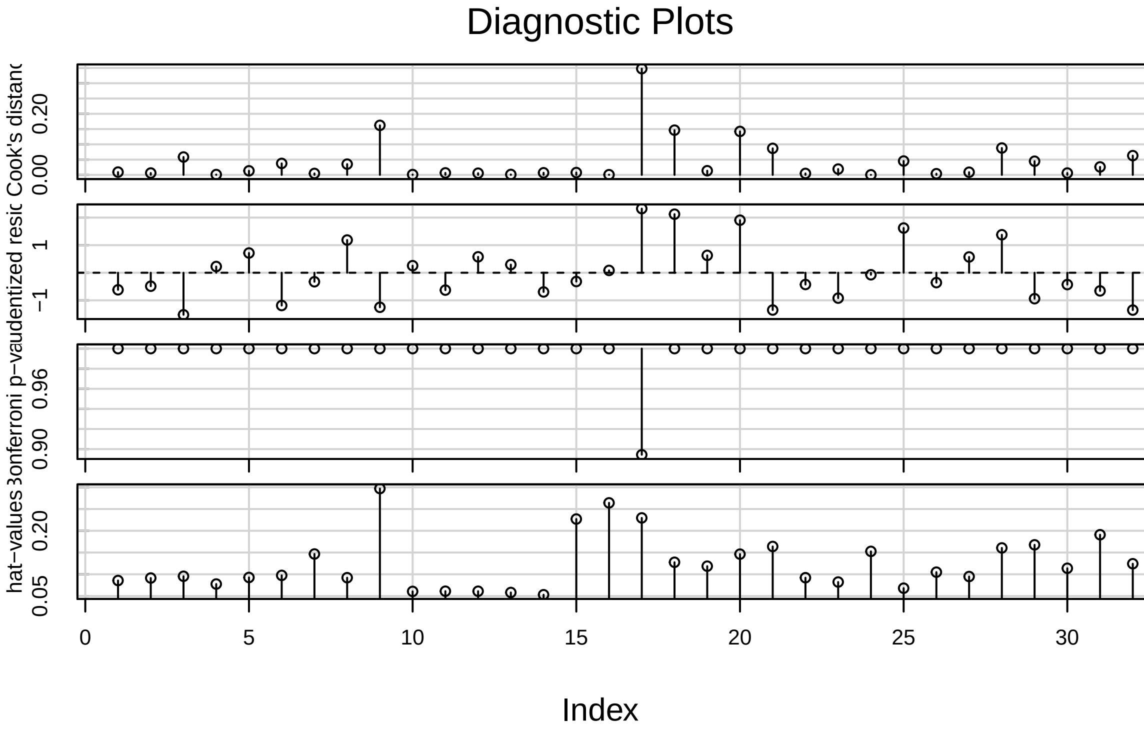
Residuals vs Fitted Plot

```
plot(mdl, which = 2)
```



Q-Q Plot

```
influenceIndexPlot(mdl)
```



Influence Index Plot