Motor Trend MPG vs. Transmission Analysis

Executive Summary

In the following analysis, I analyze linear models for MPG of a car in the MTCars dataset. I try three nested linear models with predictors: am for the first model, am and cyl for the second, and am, cyl, and wt for the third. I include information about the anova analysis for these models below, but find that the third is the best according to that analysis. The resulting model gives the coefficient of am as 0.1765 with a confidence interval $(-2.495555, 2.8485408)$ containing zero and a t-test p value of 0.89334. Therefore, we should fail to reject the null hypothesis which here is $H_0 : \beta_{am} = 0$ (with the alternate hypothesis $H_a : \beta_{am} \neq 0$. Hence, we do not find a significant relationship between MPG and transmission. All code is collected in an appendix for ease of reading

Key finding and suggestions for further study: We fail to reject the null hypothesis for the selected linear model, namely $H_0 : \beta_{am} = 0$ finding a p value of $p = 0.89334 \gg 0.05$, and hence find that, given the current data, transmission does not appear to significantly affect MPG. To further investigate this problem one could for example hold weight approximately fixed by considering MPG or multiple models of cars with both transmission types included in the dataset.

Overview and Exploration of MTCars Dataset

The MTCars dataset is a set of 32 observations of 11 variables. The key variables for our analysis will turn out to be transmission type (am where automatic$= 0$ and manual$= 1$), number of cylinders (cyl $\in \{4, 6, 8\}$), and weight (wt in units of 1000 lbs) In Fig. 1 below, a series of boxplots is shown to compare each delivery method for a given dose level. Notice the inclusion of a simple linear regression of the data to indicates that the MPG is positively correlated with transmission type.

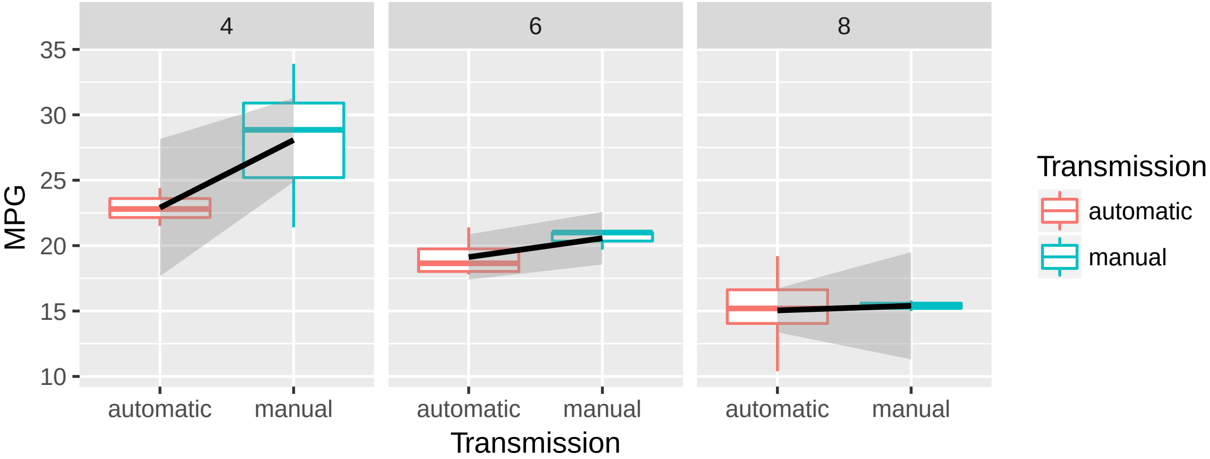


Figure 1: Boxplots of MPG given the transmission type for each cylinder number.

However, further exploratory analysis shown in Fig. 2 shows that the car weights are negatively correlated with the transmission type, so we might expect weight to be a confounding variable. An observation which infact is supported by the modeling in the next section.

Modeling the Data

Now I model the data with a series of nested linear models:

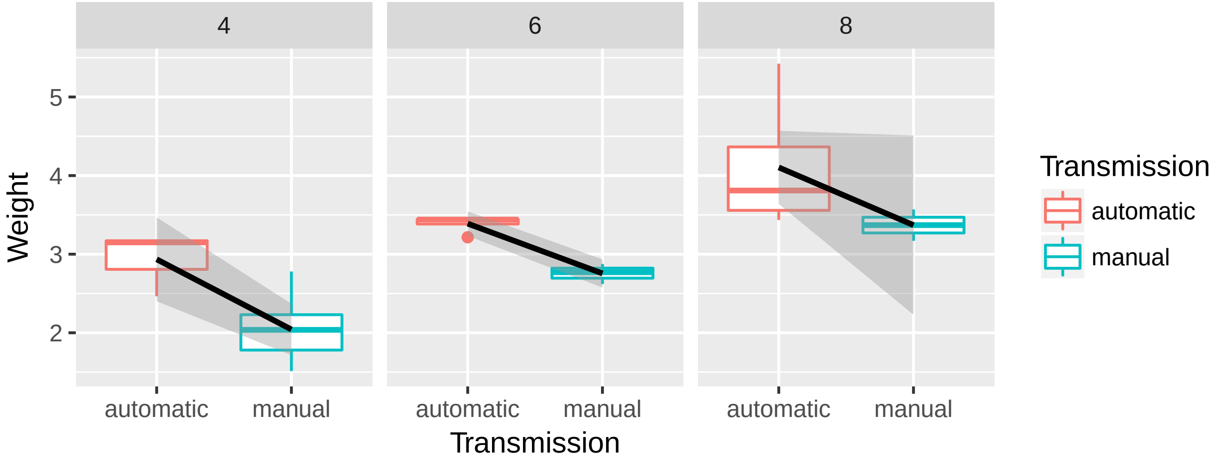


Figure 2: Boxplots of weight given the transmission type for each cylinder number. The negative correlation between weight and transmission type along with the positive correlation between MPG and transmission type indicates that weight may be a confounding variable.

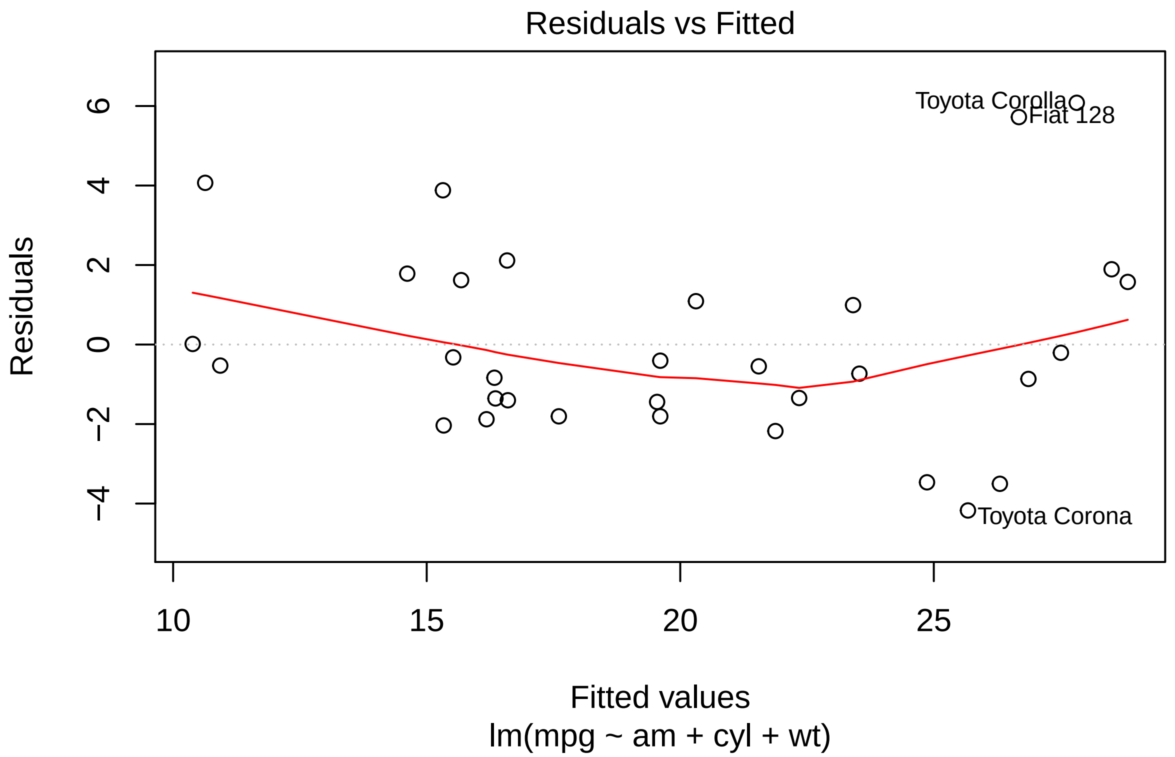
$$MPG_i = \beta_0 + \beta_{am} AM_i, \tag{1}$$
$$MPG_i = \beta_0 + \beta_{am} AM_i + \beta_{cyl} CYL_i, \tag{2}$$
$$MPG_i = \beta_0 + \beta_{am} AM_i + \beta_{cyl} CYL_i + \beta_{wt} WT_i. \tag{3}$$

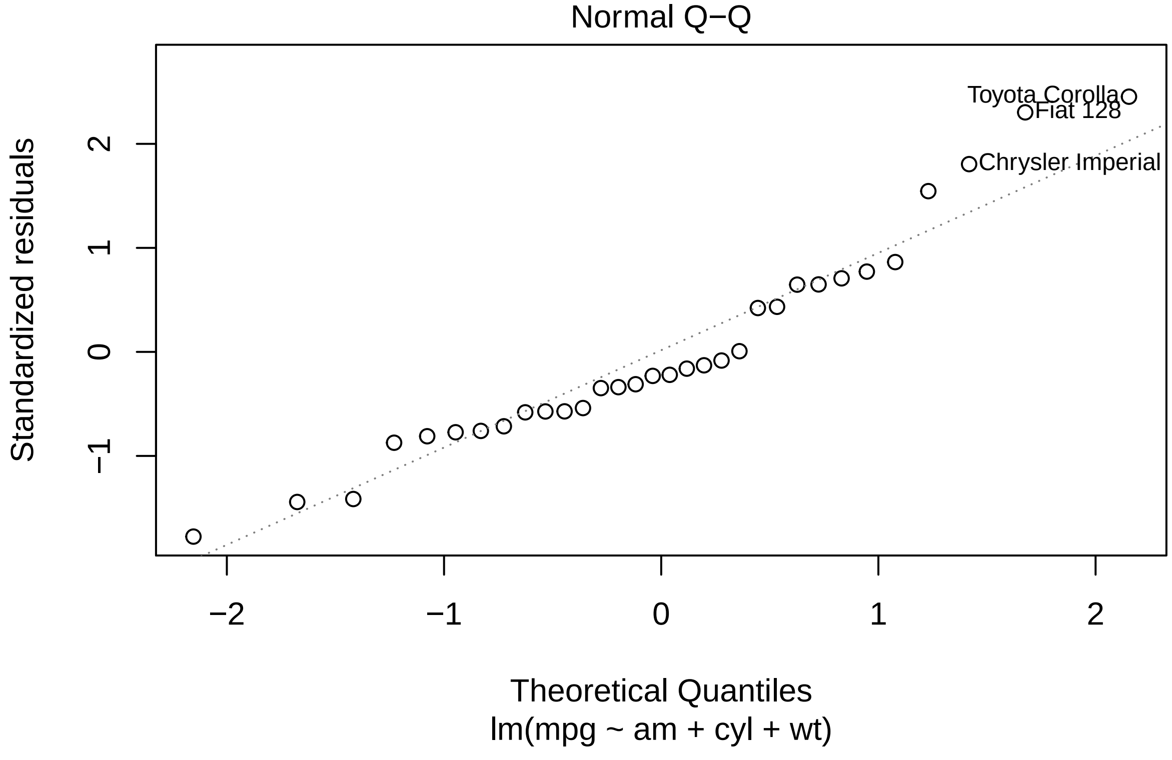
Additional nested linear models with more variables were explored, but the final was found to be the best by anova analysis and hence models with more predictors are not included here. Results of the anova analysis are shown below.

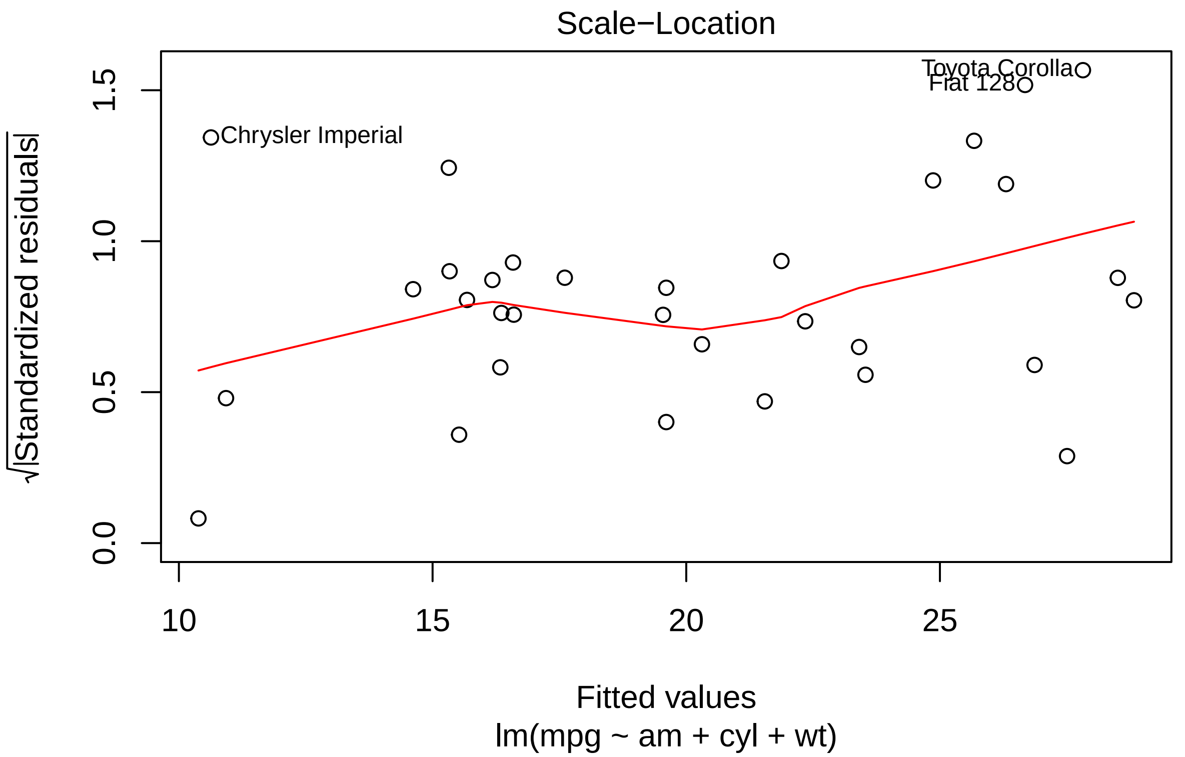
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 65.884 7.751e-09 ***
## 3     28 191.05  1     80.32 11.771  0.001886 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

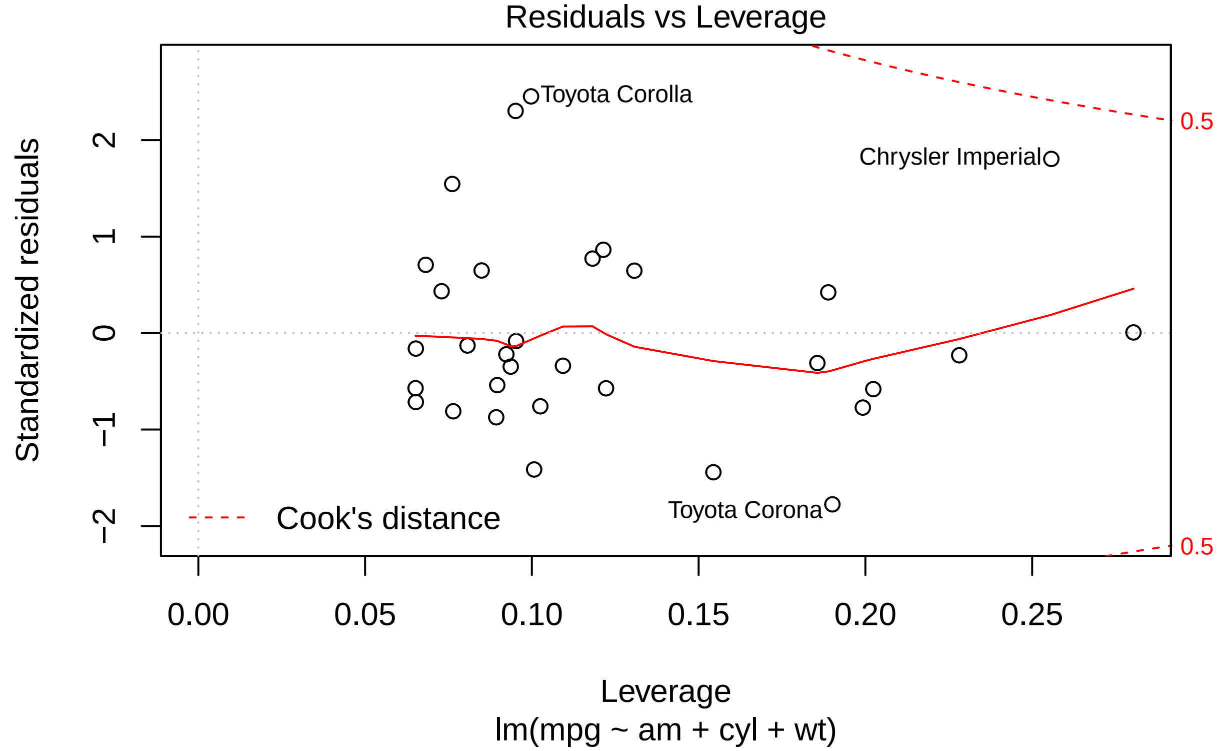
The anova analysis indicated model 3 is the best model of those presented. The model parameters are summarized below. Note that the standard error on the am coefficient indicates that its 95% confidence interval will contain zero so that the transmission type does not show a statistically significant effect on mpg. Finally, the remaining four plots are characterizations of our model. The first plot shows for example that our assumption of normally distributed residuals is fairly well met. This is even more evident in the Normal Q-Q plot shown as the second of the remaining four. The final plot seems to indicate that there are not data points with a particularly high leverage so we don't expect this to cause a problem for our model.

```
## $coefficients
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 39.4179334  2.6414573 14.9227979 7.424998e-15
## am           0.1764932  1.3044515  0.1353007 8.933421e-01
## cyl         -1.5102457  0.4222792 -3.5764148 1.291605e-03
## wt          -3.1251422  0.9108827 -3.4308942 1.885894e-03
```


Residuals vs Fitted
Residuals
6
4
2
0
−2
−4
Toyota Corolla
Fiat 128
Toyota Corona
10
15
20
25
Fitted values
lm(mpg ~ am + cyl + wt)


Normal Q−Q
Standardized residuals
2
1
0
−1
Toyota Corolla
Fiat 128
Chrysler Imperial
−2
−1
0
1
2
Theoretical Quantiles
lm(mpg ~ am + cyl + wt)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
15
20
25
Chrysler Imperial
Toyota Corolla
Fiat 128
Fitted values
lm(mpg ~ am + cyl + wt)


Residuals vs Leverage
Standardized residuals
−2
−1
0
1
2
0.00
0.05
0.10
0.15
0.20
0.25
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
0.5
0.5
Leverage
lm(mpg ~ am + cyl + wt)

Appendix

```
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(dev = 'pdf')
library(datasets)
library(ggplot2)
data("mtcars")
g<-ggplot(mtcars, aes(factor(am,labels=c("automatic","manual")), mpg))
g<-g+ geom_boxplot(aes(color=factor(am,labels=c("automatic","manual"))))
g<-g+ facet_grid( ~ cyl)+geom_smooth(method = "lm", se=TRUE, color="black", aes(group=2)
                                  ,level=0.95)
g+labs(x="Transmission",y="MPG",color="Transmission")

library(datasets)
library(ggplot2)
data("mtcars")
g<-ggplot(mtcars, aes(factor(am,labels=c("automatic","manual")), wt))
g<-g+ geom_boxplot(aes(color=factor(am,labels=c("automatic","manual"))))
g<-g+ facet_grid( ~ cyl)+geom_smooth(method = "lm", se=TRUE, color="black", aes(group=2)
                                  ,level=0.95)
g+labs(x="Transmission",y="Weight",color="Transmission")

mdl1<-lm(mpg~am,data=mtcars)

mdl2<-lm(mpg~am+cyl,data=mtcars)

mdl3<-lm(mpg~am+cyl+wt,data=mtcars)

print(anova(mdl1,mdl2,mdl3))
s<-summary(mdl3)
print(s[4])
plot(mdl3)
```