Motor Trend Analysis

C. McBride

December 24, 2017

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Executive Summary

This report provides an analysis of the `mtcars` dataset and specifically looks for any significant association between the type of transmission (automatic or standard) a vehicle is equipped with and fuel efficiency as measured in miles/gallon. The methods of analysis were primarily used linear regression but also included exploratory plots, investigation of correlation to assist in feature selection, as well as statistical inference. The relevant R code, plots, and calculations can be found in the appendices. Results from the analysis show that while in the `mtcars` dataset fuel efficiency has an apparently significant association with transmission type, it is only marginal and not directly linked. When controlling for other factors such as vehicle weight, number of cylinders, and displacement the association between fuel efficiency and transmission type all disappears. These confounding factors are closely correlated with one another, which logically makes sense. The findings suggest that for reasons that aren't clear in the data, transmission type is also closely related to the three confounders. The highlights the primary limitation of the dataset and the analysis, a dataset with relatively few observations.

Motivating Questions

Motor Trend has asked that two motivating questions be addressed:

"Is an automatic or manual transmission better for MPG"

"Quantify the MPG difference between automatic and manual transmissions"

In order to better prepare ourselves to answer these questions, we'll start with a survey of the data and the features it describes for each observation.

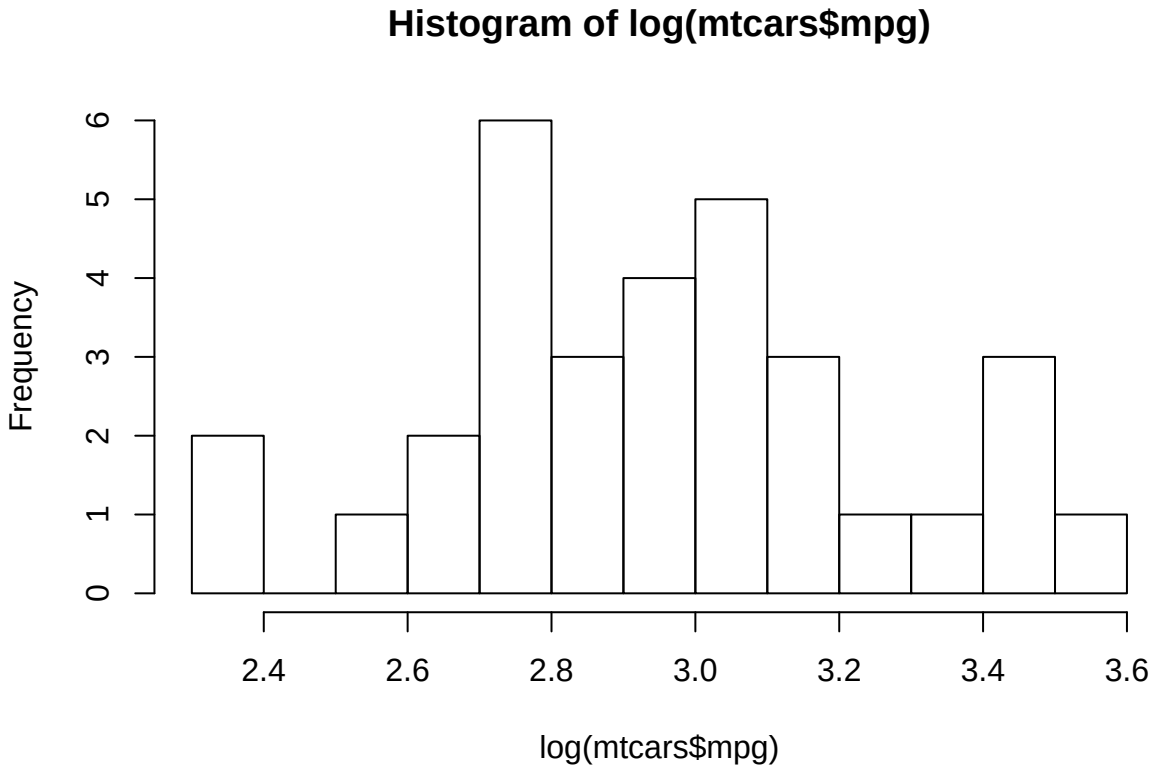
Exploratory Data Analysis

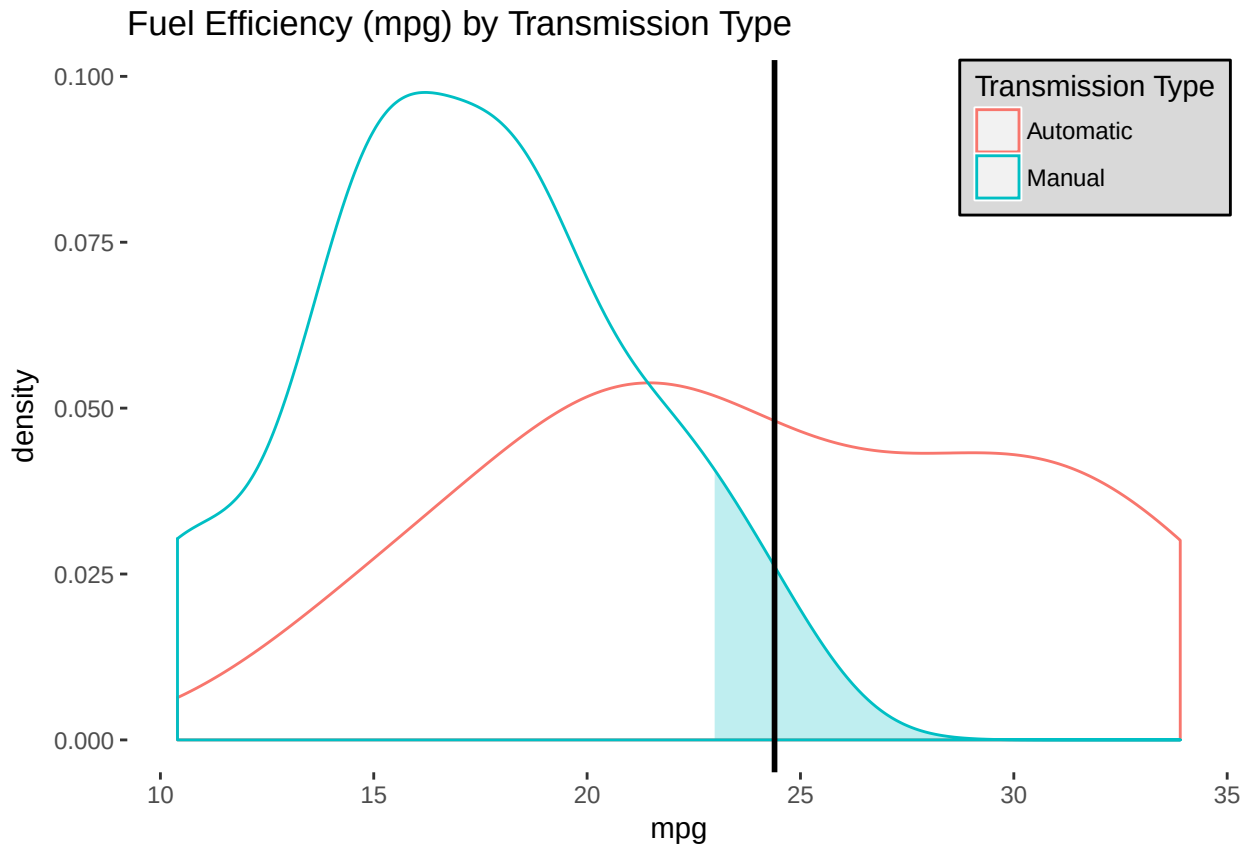
The `mtcars` dataset has 11 features describing 32 unique make and model combinations of vehicles. To get an idea of the interactions between the different features, and in particular look for signs of a correlation between transmission type and fuel efficiency, we'll start the exploratory analysis with a paired plot (*fig. 1*).

There appears to be a strong association between transmission type and fuel efficiency. However, is also strongly correlated with several other features (number of cylinders, displacement, and weight among others) that may better explain differences in fuel efficiency and therefore need to be addressed as confounders during modeling.

A density plot (*fig.2*) gives a better idea of how fuel efficiency measure are distributed by transmission type and shows that, at least marginally, the correlation is significant. We can further examine the correlation by constructing a linear model with miles/gallon as the response variable and transmission type as the predictor.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.43   19.20   20.09   22.80   33.90
```



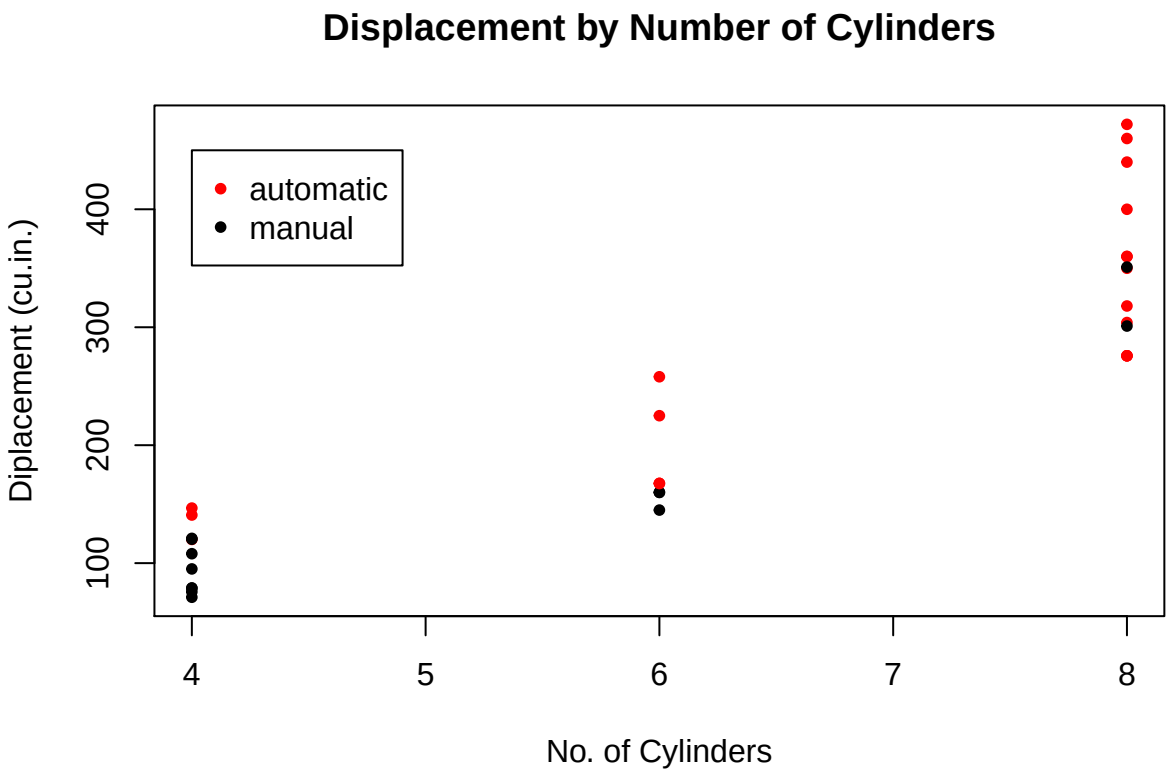


Analysis

The model confirms what was demonstrated in the density plot that the two features are signficantly associated. The model also quantifies the linear difference as β_1, the slope coefficient which says that on average automatic transmission vehicles get 7.24 fewer mpg compared to manual transmission cars. The model appears to be a good fit, but as the pairs plot demonstrated there are several potential confounders, and the correlation might be coincidental.

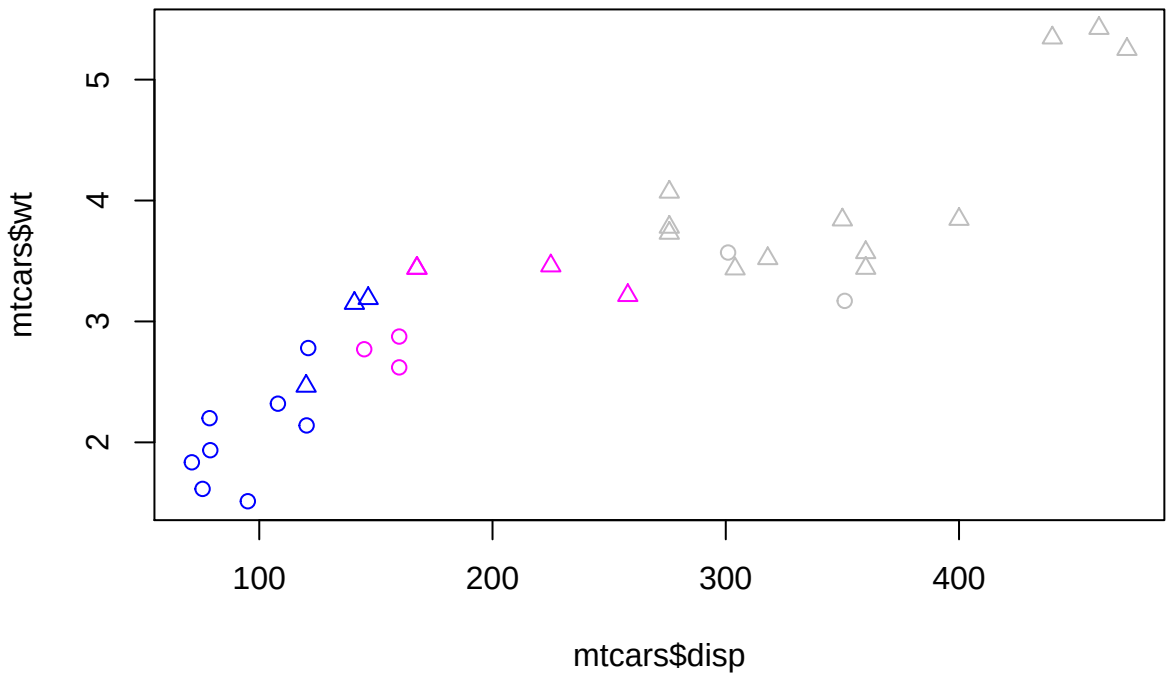
Two of the most closely correlated with transmission, number of cylinders and displacement, would intuitively be closely correlated since the latter is naturally a function of the former. A simple scatter plot (*fig. 3*) shows a clear pattern and confirms this intuition. While there are some automatic cars with small 4 cylinders and small displacements there are far more in with large diplacements and 8 cylinders, both categories correlated with low fuel efficiency.

```
# plot of displacement by number of cylinders
plot(mtcars$cyl, mtcars$disp, col=factor(mtcars$am), pch=20,
     main="Displacement by Number of Cylinders",
     ylab="Diplacement (cu.in.)", xlab="No. of Cylinders")
legend(4, 450, legend=c("automatic", "manual"), col = c("red", "black"), pch=20)
```



A similar pattern can be seen with weight. A plot of displacement by weight with points colored by number of cylinders shows how closely the three variables are correlated (*fig. 4*). All three variable track neatly together. For this reason, only the one that is most closely associated with fuel efficiency will be used. The correlations for weight, displacement, and number of cylinders respectively are -0.868, -0.848, and -0.852.

```
# plot of weight by displacement colored by number of cylinders
plot(mtcars$disp, mtcars$wt, col=mtcars$cyl, pch=as.integer(mtcars$am))
```

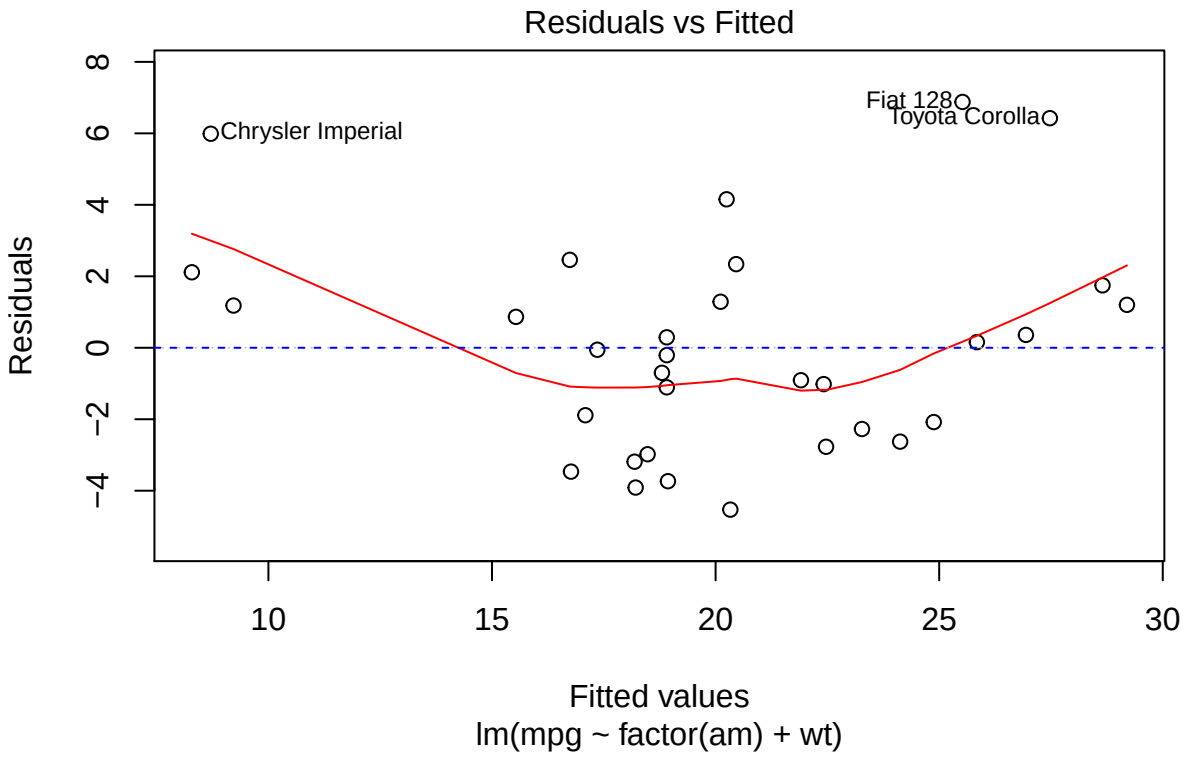


The intuition is that although it appeared to be strongly correlated to fuel efficiency in the naive model, transmission type will be insignificant when weight is added to the model.

```
# linear model that now includes weight as a controlling factor
fit2 <- lm(mpg ~ factor(am) + wt, data=mtcars)
```

The coefficient for automatic transmission is near zero (β_{AT}=0.0236152) indicating a neglible difference in the average fuel efficiency between automatic and manual transmission vehicles featured in the dataset when weight is accounted for. Based on the strong correlation that we established between weight and number of cylinders and displacement, we could expect similar results if we included either of the latter two in place of weight.

```
# residuals plots
par(mfrow=c(1,1))
plot(fit2, which=c(1))
abline(h=0, lty=2, col="blue")
```



There are a few candidate outliers with residuals greater than the coefficient β_1 we had in the naive model.

```
# hat values of outliers
summary(hatvalues(fit2))
```

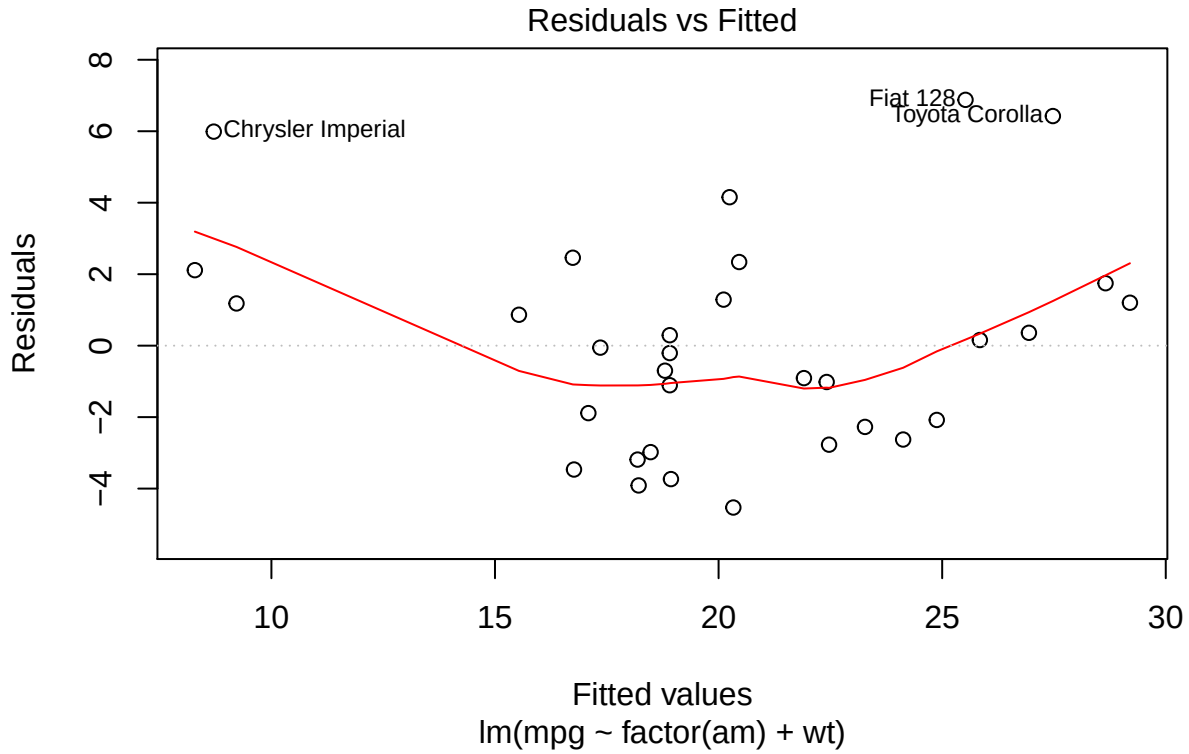
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.05264 0.05943 0.07861 0.09375 0.10236 0.22998
```

```
hatvalues(fit2)
```

```
##           Mazda RX4       Mazda RX4 Wag          Datsun 710
##          0.07975100          0.09086142          0.07745919
##      Hornet 4 Drive   Hornet Sportabout             Valiant
##          0.07249388          0.05963466          0.05880884
##          Duster 360           Merc 240D            Merc 230
##          0.05519265          0.07432731          0.07742913
##            Merc 280           Merc 280C          Merc 450SE
##          0.05963466          0.05963466          0.05850122
##          Merc 450SL         Merc 450SLC  Cadillac Fleetwood
##          0.05272952          0.05263956          0.19465076
## Lincoln Continental   Chrysler Imperial            Fiat 128
##          0.22997962          0.21345363          0.07980538
##         Honda Civic      Toyota Corolla       Toyota Corona
##          0.11794361          0.09840238          0.16269936
##    Dodge Challenger         AMC Javelin          Camaro Z28
##          0.05664215          0.05984920          0.05295890
##    Pontiac Firebird           Fiat X1-9       Porsche 914-2
##          0.05300656          0.09159169          0.08167767
```

```
##        Lotus Europa      Ford Pantera L        Ferrari Dino
##           0.12912997          0.11421877          0.08526688
##       Maserati Bora          Volvo 142E
##           0.16388758          0.08573819
```

```
plot(fit2, which=1)
```



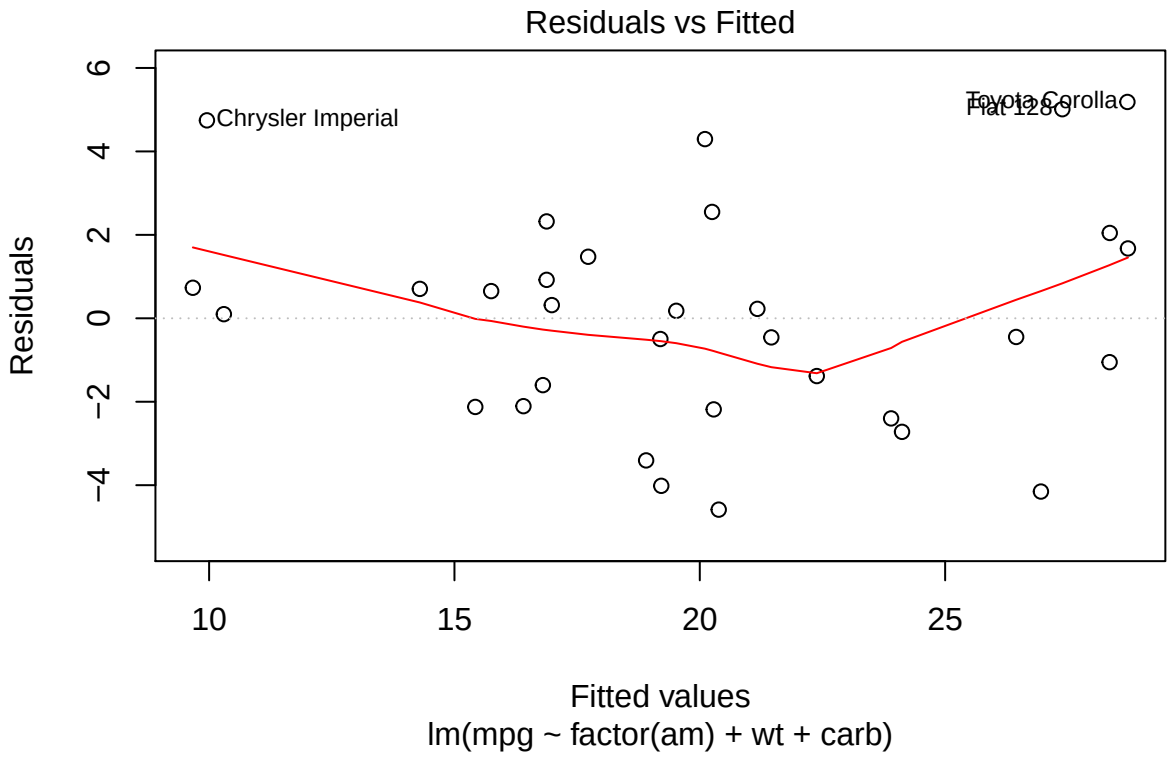

```
mtcars[c("Chrysler Imperial","Fiat 128", "Toyota Corolla"),]
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Chrysler Imperial 14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128          32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Toyota Corolla    33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
```

```
mean(mtcars$wt)
```

```
## [1] 3.21725
```

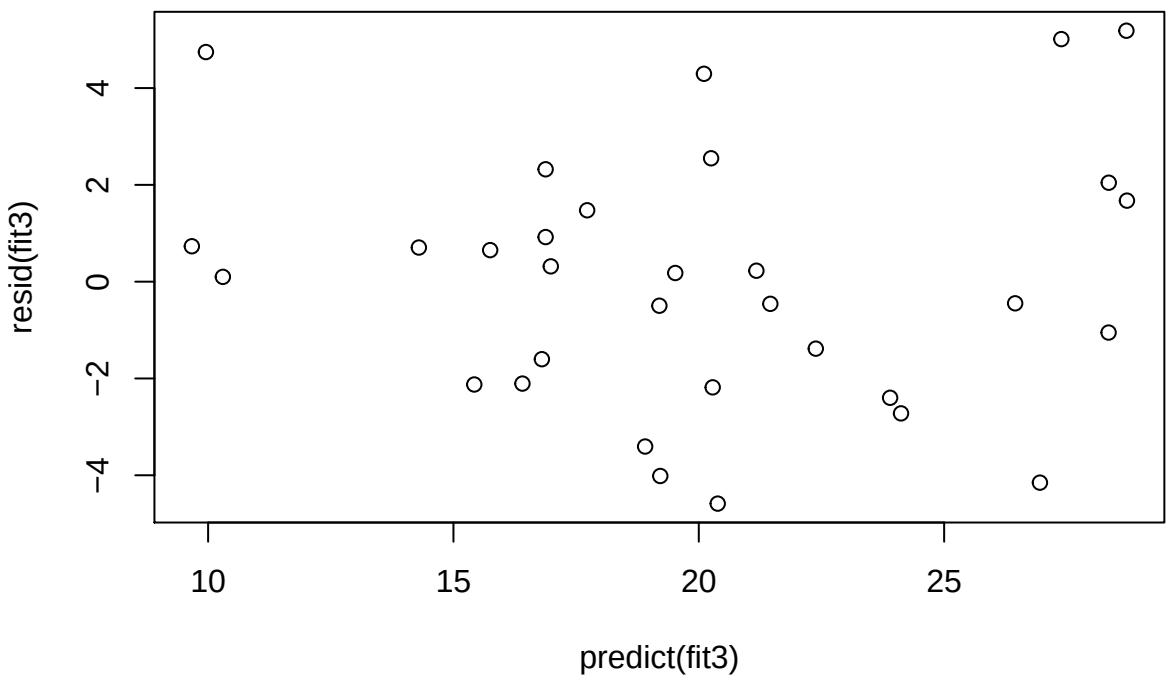
```
fit3 <- lm(mpg ~ factor(am) + wt + carb, data=mtcars)
plot(fit3, which=1)
```




```
hatvalues(fit3)
```

```
##           Mazda RX4       Mazda RX4 Wag          Datsun 710
##          0.09236266          0.09410456          0.14599799
##      Hornet 4 Drive   Hornet Sportabout             Valiant
##          0.09046384          0.06096586          0.09386557
##          Duster 360           Merc 240D            Merc 230
##          0.10722676          0.07464351          0.07812952
##            Merc 280           Merc 280C          Merc 450SE
##          0.12533925          0.12533925          0.05922123
##          Merc 450SL         Merc 450SLC  Cadillac Fleetwood
##          0.05493581          0.05394405          0.21330189
## Lincoln Continental   Chrysler Imperial            Fiat 128
##          0.26037387          0.23816260          0.13538376
##         Honda Civic      Toyota Corolla       Toyota Corona
##          0.11936061          0.12290083          0.16352598
##    Dodge Challenger         AMC Javelin          Camaro Z28
##          0.05954212          0.06110238          0.08168940
##    Pontiac Firebird           Fiat X1-9       Porsche 914-2
##          0.06848175          0.12335651          0.08750227
##        Lotus Europa      Ford Pantera L        Ferrari Dino
##          0.13270414          0.11426900          0.22406466
##       Maserati Bora          Volvo 142E
##          0.40567003          0.13206834
```

```r
plot(predict(fit3), resid(fit3))
```



Conclusions

In response to the motivating questions posed by *Motor Trend*, within the context of the `mtcars` dataset and without controlling for other variables, automatic transmission is associated with worse gas mileage not better. The difference was captured in our first, naive linear model which calculated the difference in means to be 24.3923077mpg, that is on average automatic transmission vehicles in the dataset have mileage measures over 7mpg lower than manual cars. As was shown, though, these results are irrelevant, because the several inter-correlated factors better explain the differences in gas mileage. When including a representative factor from this group in the linear model there is almost no apparent relationship between fuel efficiency and transmission type. The final model was further improved by including factors the exceptional cases of what were likely high-performance cars.

Confidence Intervals