Motor Trend Data Analysis

Alexander M Fisher

November 17 2020

Executive Summary:

In this study we have focused on a cars data set `mtcars`. The analysis taken and documented here, focuses on the relationship between transmission type, i.e. automatic or manual, and miles per gallon (MPG). The study aims to answer two questions,

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

Overall this report finds that yes manuals are slightly better with regards to MPG than automatics. MPG is dependent on other variables more significantly, and it is difficult to quantify the actual amount between types. The data does not overlap between types and is a quite small set which has to be taken into consideration. These results do find however, with respect to the data provided a 2.94 increase in mpg for manuals.

Load and Prepare Data:

First the data is loaded into R. The cyl and am varibales in the set are changed to factor variables as well. This can be seen in the summary provided with `str()`.

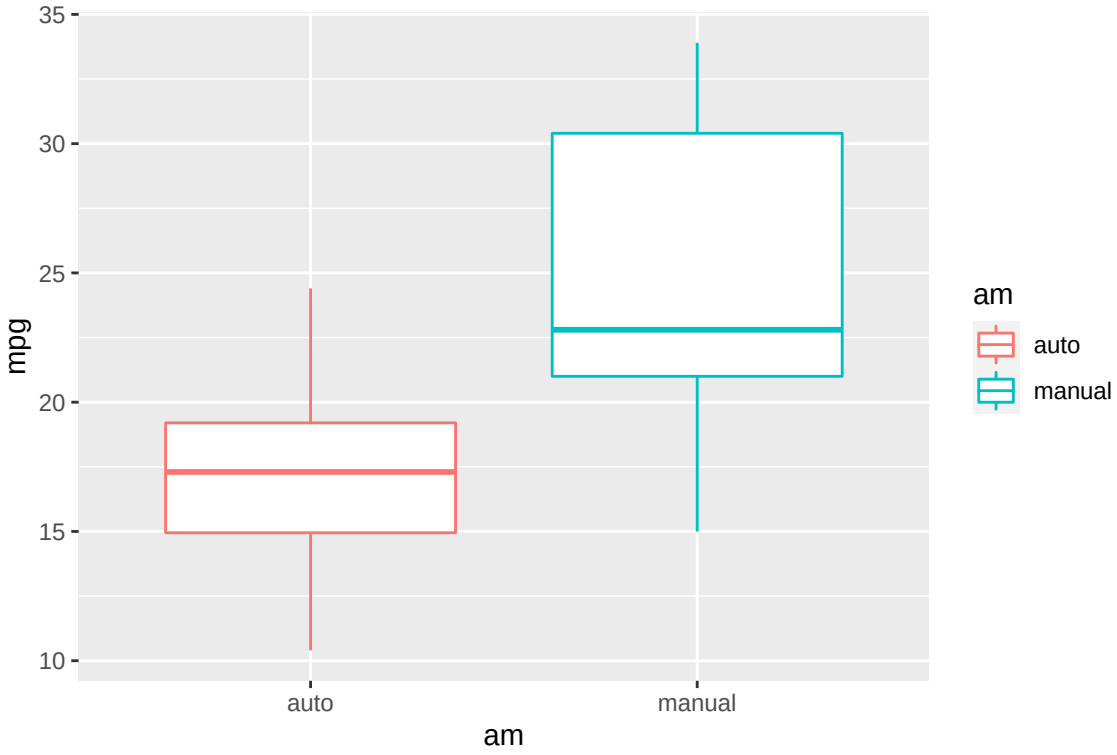
```
data("mtcars")
data <- mtcars
data$am <- factor(data$am, labels = c("auto","manual"))
data$cyl <- factor(data$cyl)
print(str(data))
```

```
## '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 : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ 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  : Factor w/ 2 levels "auto","manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
## NULL
```

Exploratory Analyis:

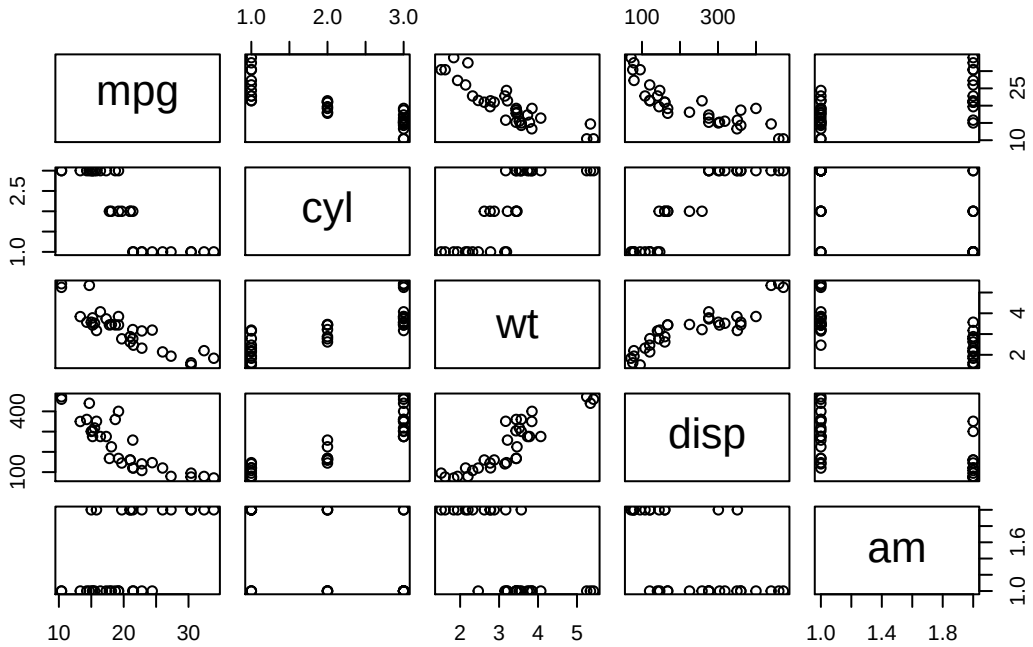
To get an initial look at the data and see the relationship between mpg and am, I have plotted a box plot.

```
library(ggplot2)
box_plot <- ggplot(data = data, aes(x = am, y = mpg)) +
                geom_boxplot(aes(colour = am))
print(box_plot)
```



This box plot clearly demonstrates that cars with a manual transmission have, or in this data set are related to cars with higher, mpg compared with cars that have automatic transmissions. This may not however mean there exists an underlying causal relationship between the two. There are quite a few other variables that are also correlated with mpg. Lets take a quick look using `plot()`.

```
plot(data[,c("mpg","cyl","wt","disp","am")])
```



It can be seen with mpg as outcome that cyl, wt, disp all are strongly negatively correlated with mpg, as expected. These variables are not independent either, as it is clear that in general heavier cars need larger engines, and larger engines have bigger displacement.

Model Selection:

In this section I will build some models. The first model is with just am as a predictor variable.

```
fit_1 <- lm(mpg~am, data = data)
summary(fit_1)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = data)
##
## 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 ***
## ammanual       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
```

From this we can see on average the mpg for automatics is 17.147, and for manuals there is a 7.245 increase on that. This result is significant (p_value < 0.05). The R-squared is poor as well. The initial standard error was (`sd(data$mpg) = 6`) quite high. After adding the am as a predictor the variance has been reduced,

however, a lot still remains. This model further demonstrates what we saw in the box plot, but for a better model more predictors will need to be added.

Now using the `step()` function from the stats package starting from a full model (i.e. using all variables), we will get the 'best' model with a subset of variables. This is a stepwise algorithm using AIC as the model selection criterion.

```
best_fit <- step(lm(mpg~., data = data))
```

```
summary(best_fit)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ammanual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

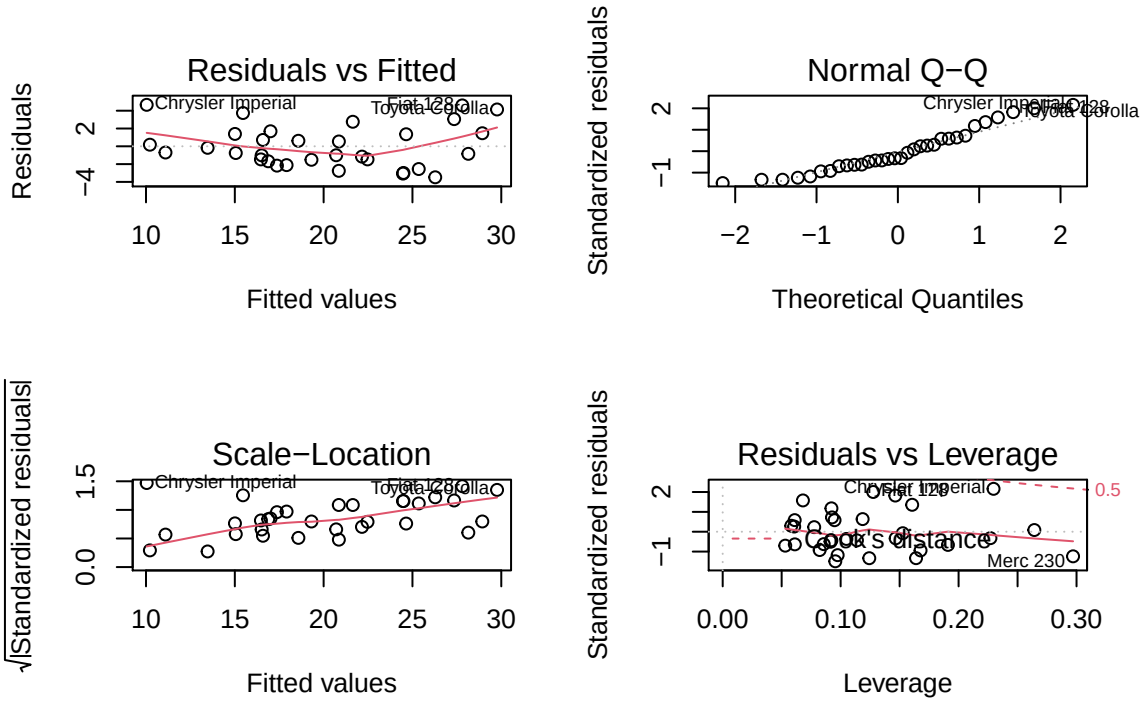
In this model the p-values are all significant. The standard error has been reduced again. In addition the R-squared value is good at 0.84, thus explaining 84% of the mpg variables variance. This is a good working model that isn't overly complicated and describes the relationships well. The model has 3 predictor variables, wt, qsec, and am. Holding other variables constant increasing wt by unit will decrease mpg by 3.9. Increasing qsec by 1 unit, will increase mpg by 1.2. Finally, it can be seen on average manual cars have 2.94 increase in mpg compared to automatics.

From here I will perform some residual plots.

Model Examination:

To examine the linear regression model derived in the above analysis residual plots will be done.

```
par(mfrow = c(2, 2))
plot(best_fit)
```



Overall the plots are reasonable and don't show any particular patterns. The Residuals vs. Fitted plot supports the assumption of independence. The distribution of residuals is normal as evidenced in the Q-Q plot.

Conclusion:

It can be concluded there is a difference in mpg for different transmission types. Manual are slightly better with a average 2.94 increase in mpg (standard error is 1.4). The results do however, indicate that there are more statistically significant predictors including wt, and qsec than am type when determining mpg. In addition with automatic cars improving and a wealth of new data, these results should be taken with a pinch of salt.