# Regression Models on Motor Trend

*Li Xiaowei*

# Executive Summary

This report is going to explore on the data set *Motor Trend* which is the *mtcars* in R. Two basic questions will be addressed in this report.

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions.

Techniques of exploratory data analysis, statistical inference and regression models will be mainly used in this report. Due to limitation of the report length, all code and plots will be put in the appendices. It's going to be slightly difficult to read.

# Which transmission is better?

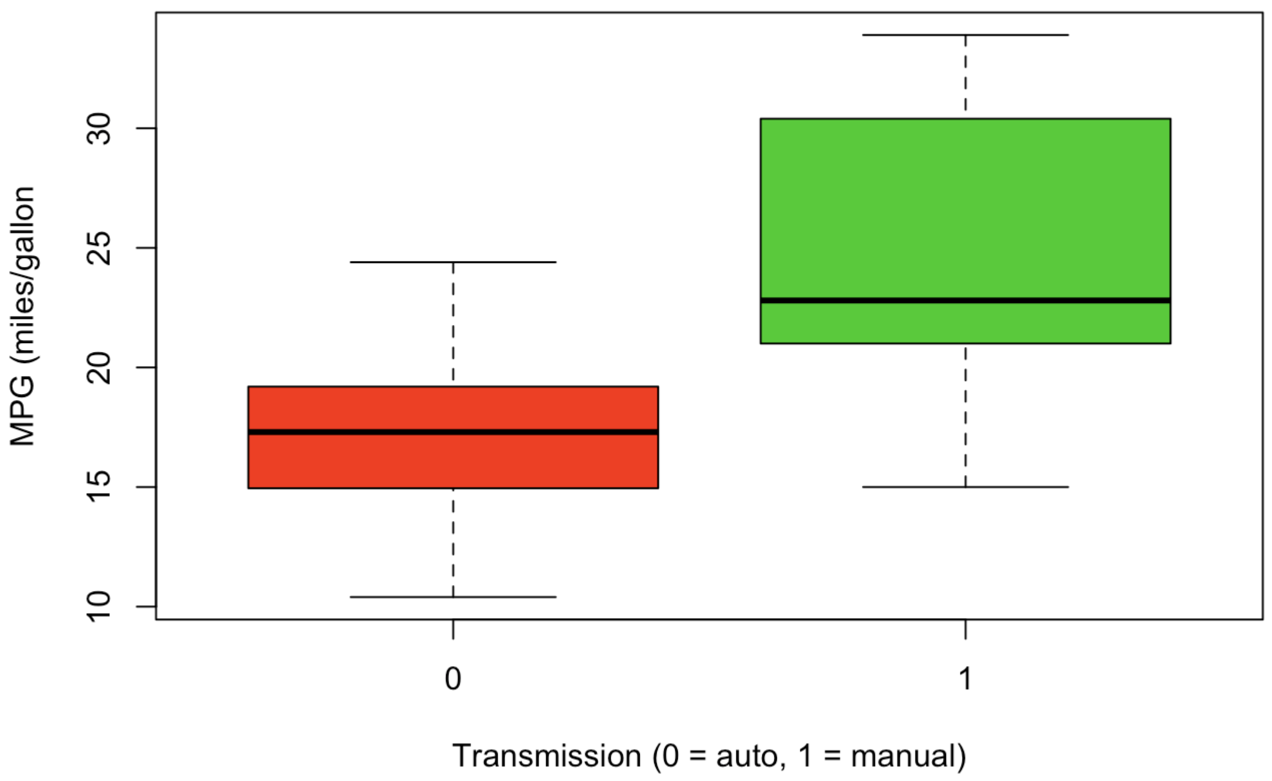
According to the documen from R. The *MPG* means miles per gallon. Auto or manual transmission is defined in *am* column, 0 means auto and 1 means manual.

Before running into any calculation, let's take a look at the data in text and plot.

```
library(datasets)
data(mtcars)
library(knitr)
kable(head(mtcars))
```

| | mpg | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mazda RX4 | 21.0 | 6 | 160 | 110 | 3.90 | 2.620 | 16.46 | 0 | 1 | 4 | 4 |
| Mazda RX4 Wag | 21.0 | 6 | 160 | 110 | 3.90 | 2.875 | 17.02 | 0 | 1 | 4 | 4 |
| Datsun 710 | 22.8 | 4 | 108 | 93 | 3.85 | 2.320 | 18.61 | 1 | 1 | 4 | 1 |
| Hornet 4 Drive | 21.4 | 6 | 258 | 110 | 3.08 | 3.215 | 19.44 | 1 | 0 | 3 | 1 |
| Hornet Sportabout | 18.7 | 8 | 360 | 175 | 3.15 | 3.440 | 17.02 | 0 | 0 | 3 | 2 |
| Valiant | 18.1 | 6 | 225 | 105 | 2.76 | 3.460 | 20.22 | 1 | 0 | 3 | 1 |

```
#plot(mtcars$mpg, col = mtcars$am + 2, xlab = "Car Index", ylab = "MPG (miles/gallon)")
#abline(h = tapply(mtcars$mpg, factor(mtcars$am), mean), col = c(2, 3))
#legend("topright", pch = 1, col = c(2, 3), legend = c("Auto", "Manual"))
boxplot(mpg ~ am, col = c(2, 3), xlab = "Transmission (0 = auto, 1 = manual)", ylab = "MPG (miles/gallon", data =
 mtcars)
```



The two horizontal lines are the means of *MPG* of each group. It's quite obvious from the plot to observe that manual transmission is better than auto, conditioning that we're not considering the other factors like number of cylinders or car weight. And let's prove it by performing hypothesis testing.

Assume the null hypotesis is that auto and manual transmission are equally efficient.

```
#autoMPG <- mtcars[mtcars$am == 0, ]$mpg
#manualMPG <- mtcars[mtcars$am == 1,]$mpg
#t.test(autoMPG, manualMPG, paired = FALSE)
t.test(mpg~am, data = mtcars, paird = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

The p-vale is 0.001 which is smaller than 0.05 so we'll reject the null hypothesis and conclude auto and manual transmission efficiency are different. Moreover, t value is -3.7671231 and the confidence interval of this t test is [-11.2801944, -3.2096842] which indicate that auto transmission is worse for MPG. Same conclusion as visual inspection previously. This addresses the first question at the beginning of this report.

However, the factors of number of engine cylinders, car weight and etcs were not considered during the above analysis. There are only 32 samples. Dividing them into more groups will cause very huge standard error of mean, eventually will become difficult to analyze.

# Quantify the MPG difference

There are multiple choices when performing regression model selection, ANOVA, VIF and Covariate model selection. I prefer the covariate model selection by applying the stepwise regression. It eliminates my effort of comparing different models manually.

Let's take a look at what will happen if I choose all columns (other than MPG) as regressors.

```
full.model <- lm(mpg ~ ., data = mtcars)
summary(full.model)$coef
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## am           2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

Look at those p-values, the regressors all look equally unsignificant. It's quite difficult to choose which regressor or not. We can try eliminate the regressor one by one. But as I mentioned, I prefer to leave it to computer and algorithm to determine for me.

```
reduced.model <- step(full.model, direction = "backward")
```

```
summary(reduced.model)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## 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 ***
## am            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
```

The chosen variables are *wt*, *qsec* and *am*. Usually *am* should be treated as categorical variable before performing the regression. However, this variable only has value of 0 and 1. There is no need to perform factoring beforehand.

Let's take a more detailed look at the coefficients. *wt* has negative coefficient while the other two variables have positive coefficients.

One unit increament on *wt* (1000 lb) will decrease MPG by 3.9 because car is heavier and less efficient. One unit increament on *qsec* (1/4 mile time) will increase MPG by 1.2 because the engine is less powerful and usually more efficient on MPG. One unit increament on *am*, in this report it means change form auto(0) to manual(1) will cause MGP increase by 2.9. This is the same observation as performing t test previously that manual transmission is more efficient. This also addresses the second question stated at the beginning of this report.

Before concluding this *reduced.model* is the final selection, let's do a comparison with the *full.model* and also plot the residual to ensure there is no pattern or correlation.

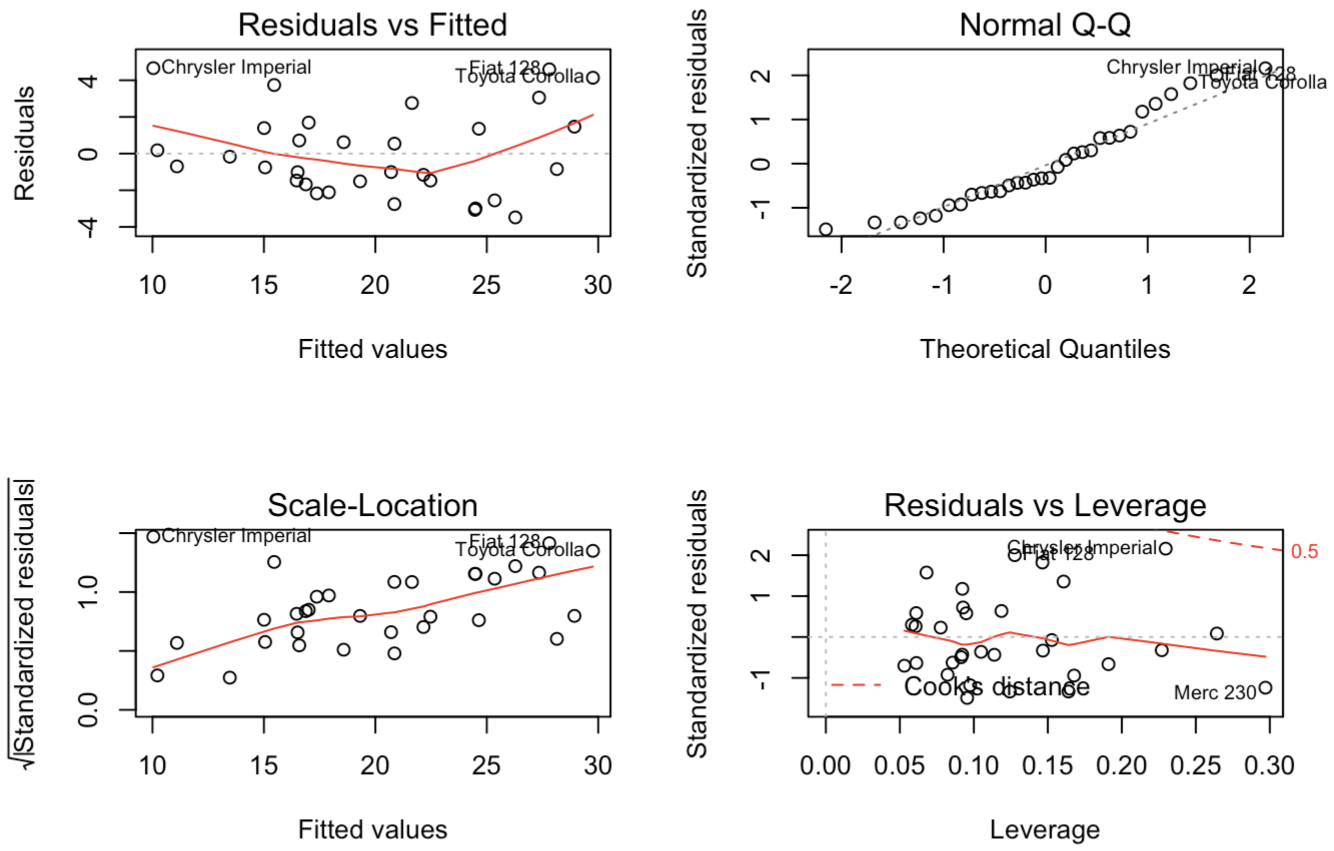
```
anova(reduced.model, full.model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + qsec + am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 169.29
## 2     21 147.49  7    21.791 0.4432 0.8636
```

The p-value is 0.86, greater than 0.05. So we failed to reject the null hypothesis that the coefficients of the additional variables in *full.model* is 0. Meaning, we can omit those additional variables by accepting the *reduced.model* is a better one.

The F-statistic is 52.75, much greater than 1, R-squared is 0.83 which is close to 1. These figures all agree the *reduced.model* is a "fitted" model. And lastly let's plot the model to perform some diagnostics.

```
par(mfrow = c(2, 2))
plot(reduced.model)
```



The diagnostic plots of an ideal model should have relatively flat red curve & random scatterplot for Residual and Scale plots , normality of errors should follow the diagonal line in the Q-Q plot. And lastly, smooth red line is close to the grey dotted line for the Leverage plot.

We can see that the *reduced.model* does not perform well on every plot, but still quite acceptable based on the criterias stated above.

# Conclusion

The p-value, t-value and confidence interval of the t test in the first part of this report all indicating manual transmission will have greater MPG. This conclusion is also proved by the regression model selected in the second part of this report as the coefficent of *am* is positive. The p-value, r-squared, F-statistic and diagnostic plots all imply the selected model is a "fitted" regression model.