Is an automatic or manual transmission better for MPG

Stephen Yang

August 27, 2019

Executive Summary

Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

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

First, basic data processing would be used to convert the factor of am to the level of ""automatic" and "manual". Then we have to differentiate the effect of mpg from two different transmissions by boxplot and use t.test to prove the true difference effect between the two. Variance inflation provides some clues for giving up the unnecessary variables. The remaining variables are am, cyl, disp, wt, which finally are to be determined in model selection by anova.

Data Processing

Load the data from R "dataset". Originally the value of am are 0 and 1. We need to transform the binary form into factor.

```
library(ggplot2)
library(car)
```

```
## Loading required package: carData
```

```
data <- mtcars
head(data)
```

```
##                    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
```

```
data$am <- factor(data$am)
levels(data$am)<-c("automatic", "manual")
```

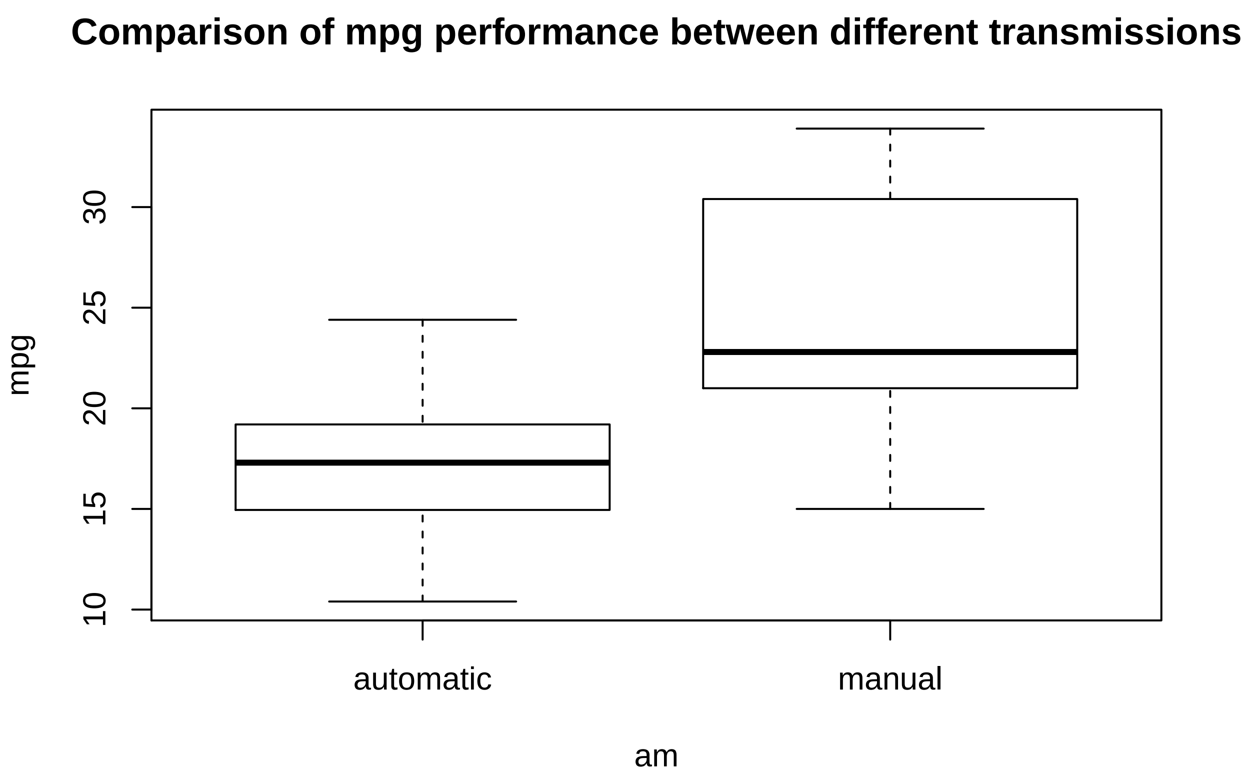
Exploratory analysis

Just take a look at the pvalue. To make sure if transmission type is important to engine performance. The boxplot is to demonstrate engine efficiency by different transmissions. It appears manual transmission shows about 7 mpg more than automatic transmission.

```r
t.test(data$mpg[data$am == "automatic"], data$mpg[data$am == "manual"])
```

```
##
##  Welch Two Sample t-test
##
## data:  data$mpg[data$am == "automatic"] and data$mpg[data$am == "manual"]
## 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 of x mean of y
##  17.14737  24.39231
```

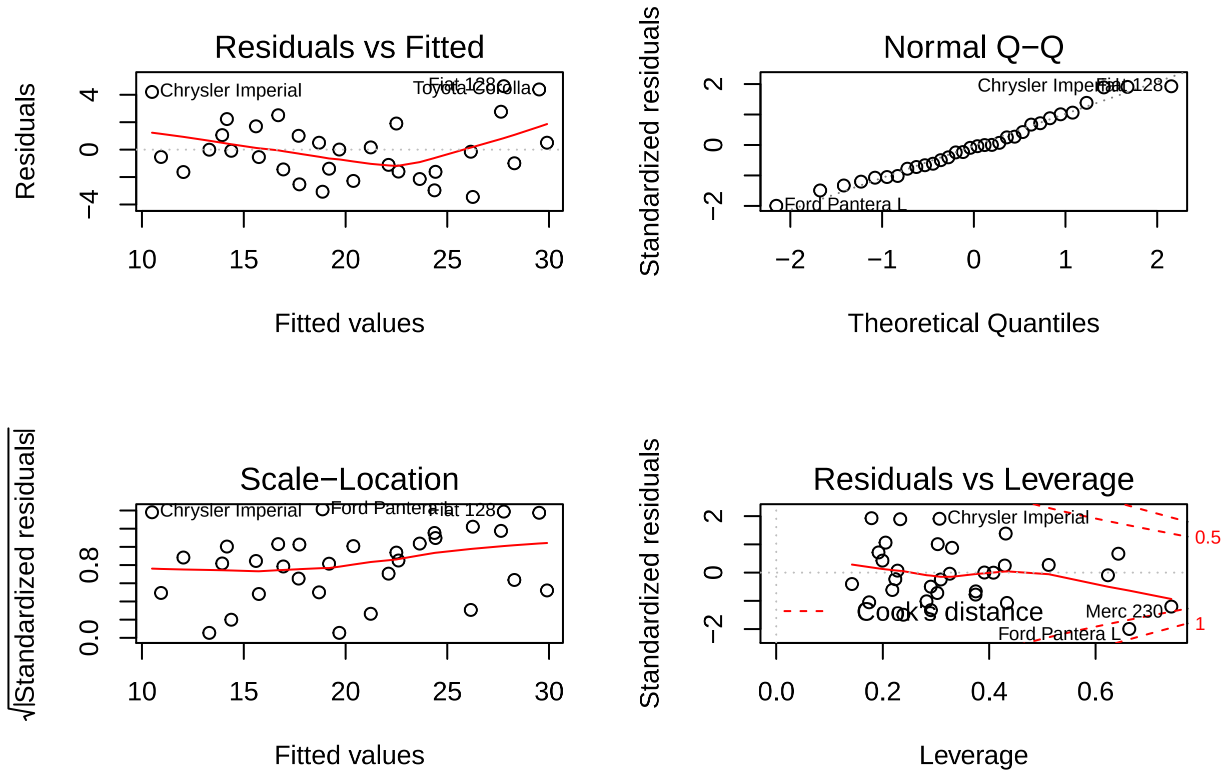
```r
boxplot(mpg ~ am, data, main = "Comparison of mpg performance between different transmissions")
```




```r
am_fit <- lm(mpg ~ am-1, data)
summary(am_fit)$coef
```

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

```r
fit <- lm(mpg ~ ., data)
par(mfrow = c(2,2))
plot(fit)
```



Model Selection

We get to check which variables are influential. One of the tools is variance inflation and it shows several variables are confounding ,which are am, cyl, disp, wt. Using anova analysis to check the confounding variables. So the nested model is a good way to see the sequential testing of coefficients. As we can see the sequential ANOVA in appendix indicates disp with low value of F statistics which isn't significant. So other variables should be included for final check.

```
test_fit <- lm(mpg ~ am-1, data)
vif(fit)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873
##        am      gear      carb
##  4.648487  5.357452  7.908747
```

```
fit1 <- lm(mpg ~ am+cyl+disp+wt-1, data)
summary(fit1)$coef
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## amautomatic 40.898313414 3.60154037 11.3557837 8.677574e-12
## ammanual    41.027378984 3.00859592 13.6367196 1.261570e-13
## cyl         -1.784173258 0.61819218 -2.8861142 7.581533e-03
## disp         0.007403833 0.01208067  0.6128661 5.450930e-01
## wt          -3.583425472 1.18650433 -3.0201537 5.468412e-03
```

Analysis and Conclusion

What if all the confounding variable are all included in the model? As we can see there's only slight difference over two transmissions. Manual transmission doesn't show better engine efficiency.

```
final_fit <- lm(mpg ~ am + cyl + wt -1, data)
sum_fit <- summary(final_fit)$coef
Interval <- sum_fit[2,1]+c(-1,1)*qt(.975,final_fit$df)*sum_fit[2,2]
```

Appendix supporting of model selection

```
test_fit1 <- lm(mpg ~ am + cyl-1, data)
test_fit2 <- lm(mpg ~ am + disp-1, data)
test_fit3 <- lm(mpg ~ am + wt-1, data)
anova(test_fit, test_fit1)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am - 1
## Model 2: mpg ~ am + cyl - 1
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 48.041 1.285e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(test_fit, test_fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am - 1
## Model 2: mpg ~ am + disp - 1
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 300.28  1    420.62 40.621 5.748e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(test_fit, test_fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am - 1
## Model 2: mpg ~ am + wt - 1
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 46.115 1.867e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
test_fit4 <- lm(mpg ~ am + cyl + wt-1, data)
test_fit5 <- lm(mpg ~ am + cyl + wt + disp-1, data)
anova(test_fit ,test_fit1 ,test_fit4, test_fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am - 1
## Model 2: mpg ~ am + cyl - 1
## Model 3: mpg ~ am + cyl + wt - 1
## Model 4: mpg ~ am + cyl + wt + disp - 1
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 64.4149 1.264e-08 ***
## 3     28 191.05  1     80.32 11.5085  0.002152 **
## 4     27 188.43  1      2.62  0.3756  0.545093
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```