Gear Shift Type correlation to Fuel Efficiency from mtcars dataset

Odin Matanguihan

May 10, 2017

R Markdown

Instructions

"Is an automatic or manual transmission better for MPG"

"Quantify the MPG difference between automatic and manual transmissions"

```
#load the data
library(car)
library(dplyr)
library(ggplot2)
library(gridExtra)
data(mtcars)
```

The main item of interest in this data set is the effect between transmission type and fuel efficiency.

A simple boxplot can show that there is a noticeable difference in fuel efficiency between auto and manual transmission types.

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

Info regarding this data set can be found in this link.

We can also check the probability that this is just an accident of sampling.

```
t.test(mtcars$mpg~mtcars$am,conf.level=0.95)$p.value
```

```
## [1] 0.001373638
```

With such a low p-value, we reject the null hypothesis that there is no correlation. There is very little chance that it is mere coincidence.

There's a lot of confounding factors however. We can try to determine which factors have the most influence using the variance inflation factor.

```
vif(lm(mpg~., mtcars))
```

```
##       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
```

This shows that hp has the most effect, followed by wt. Intuitively however, we can see that many of these factors are dependent on each other, and as such, removing one or the other could have a significant impact. A pairwise plot can show that some are strongly correlated.

A step function would try several models and automatically determine a best fit.

```
best <- step(lm(mpg ~ ., mtcars),trace=0)
best$coefficients
```

```
## (Intercept)          wt        qsec          am
##    9.617781   -3.916504    1.225886    2.935837
```

The best model that came from the function shows that the best models for predicting efficiency(mpg) includes just wt, qsec, and am as the predictors. All this, however, is from the assumption that all the other columns are uncorrelated with each other. A basic understanding of mechanics should give us the intuition that many of these are correlated. Some would show up in the pairwise function plots, others do not due to the number of confounding variables.

Based on this intuition, we will add 2 additional variables. One is the weight to power ratio, a factor that affects acceleration. The other is the weight to qsec squared, the expected relationship on the assumption that power and distance is constant.

```
mtcars<-mutate(mtcars, hp.wt = hp/wt)
mtcars<-mutate(mtcars, wt.qsec = wt/(qsec^2))
step(lm(mpg ~ ., mtcars),trace=0)$coefficients
```

```
## (Intercept)          wt        qsec          am
##    9.617781   -3.916504    1.225886    2.935837
```

The inclusion of both ratios did not result in a better model than what we have previously.

While this linear model provides the best prediction we have for the efficiency(mpg), it would also be helpful to see how variance in one factor can affect another. This is where the variance inflation factor comes in.

```
vif(best)
```

```
##       wt     qsec       am
## 2.482952 1.364339 2.541437
```

What the variance inflation factor shows is that the transmission type has the most impact in modifying the relationship between mpg and the other factors. qsec has the least impact. While the small role of qsec is expected, it is also expected that wt would play the biggest role. As it is, am has a bigger impact, albeit not by a large margin.

Next we can do a residual analysis to see if there is any high leverage or high influence data.

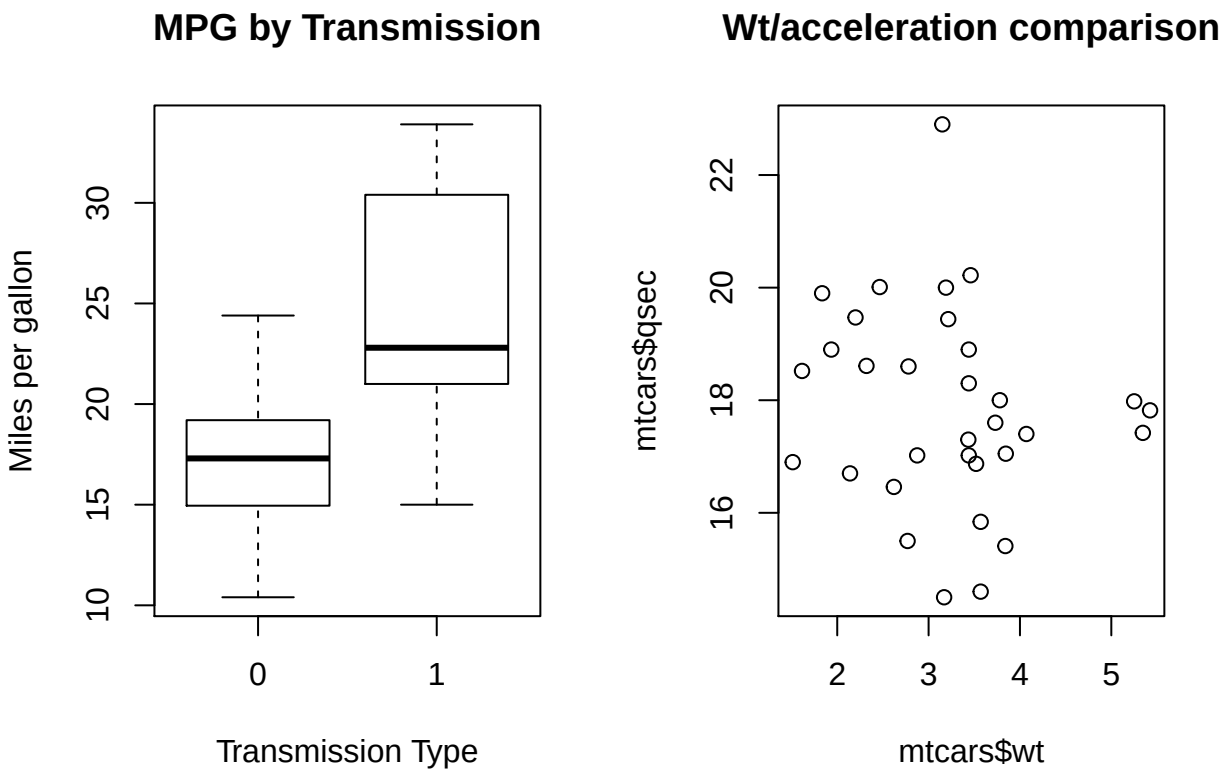
```
mtcars.resid <- resid(step(lm(mpg ~ ., mtcars),trace=0))
```

The graph of the residuals can be found be found here. At first glance, it is not too obvious if there is any high leverage or high influence data plot. It might be quantifiable with more advanced skills in variance inflation factors. That is beyond the scope of this lesson.

Boxplot MPG by Transmission Type

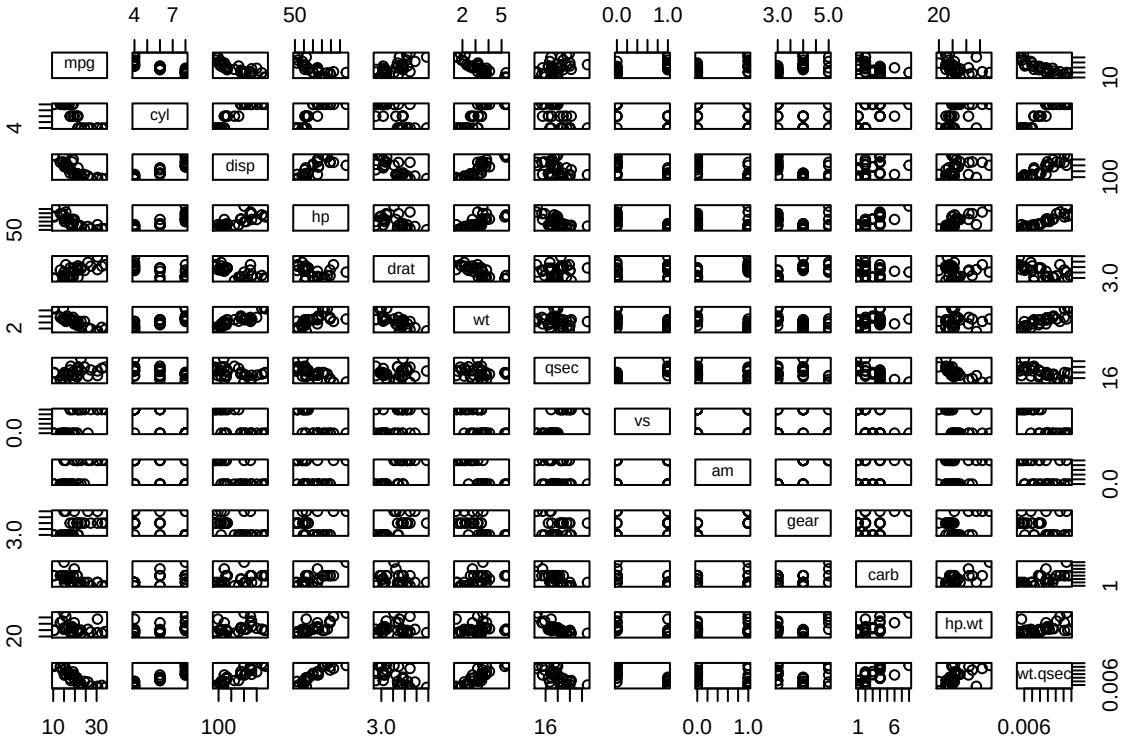
```
par(mfcol=c(1,2))

boxplot(mpg ~ am, data = mtcars, xlab = "Transmission Type", ylab = "Miles per gallon", main="MPG by Tra
plot(mtcars$wt, mtcars$qsec, main = "Wt/acceleration comparison")
abline(mtcars$wt, mtcars$qsec)
```



Pairwise plots of various factors

```
pairs(mpg ~ ., data = mtcars)
```



```
x <- mtcars$am == 1
p1 <- ggplot(mtcars[x,], aes(x=wt, y=qsec)) + geom_point(shape=1) + geom_smooth()
p2 <- ggplot(mtcars[!x,], aes(x=wt, y=qsec)) + geom_point(shape=1) + geom_smooth()
p3 <- ggplot(mtcars[x,], aes(x=wt, y=mtcars.resid[x])) + geom_point(shape=1) + geom_smooth()
p4 <- ggplot(mtcars[!x,], aes(x=wt, y=mtcars.resid[!x])) + geom_point(shape=1) + geom_smooth()
p5 <- ggplot(mtcars[x,], aes(x=qsec, y=mtcars.resid[x])) + geom_point(shape=1) + geom_smooth()
p6 <- ggplot(mtcars[!x,], aes(x=qsec, y=mtcars.resid[!x])) + geom_point(shape=1) + geom_smooth()
grid.arrange(p1, p2, p3, p4, p5, p6, ncol =2, nrow=3)
```