Relationship Between Transmission Type and MPG

Tzachi Ezra Torf-Fulton

August 18, 2015

Relationship Between Transmission Type and MPG

Executive Summary

In this report we analyzed the relationship between a set of variables and miles per gallon (MPG), based on a data that was extracted from the 1974 Motor Trend US magazine (mtcars - https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html). We focused in particular on quantifying the MPG difference between automatic and manual transmissions and determine whether an automatic or manual transmission better for MPG.

Given the data we were able to find out that excessive heat is the main cause for fatalities, followed by tornado. The later though, is by far the main cause for injuries. When we look at property damage we find that flood is the main cause, where drought is the top reason for crop damage.

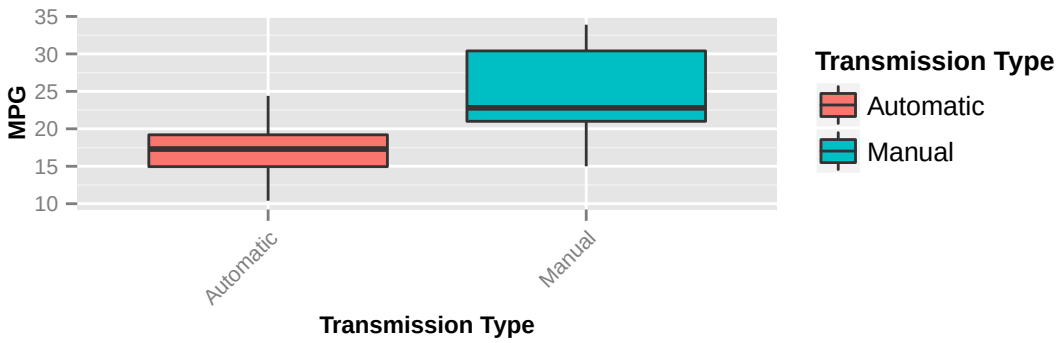
Data Processing

Loading libraries and preprocessing the data

```
library(ggplot2)
data(mtcars)
fmtcars<-mtcars
fmtcars$am <- factor(mtcars$am, labels=c("Automatic", "Manual"))
```

To get a sense of the relation between Transmission Type and MPG, we will start with a general boxplot of these two variables.

```
p <- ggplot(fmtcars, aes(factor(am), x=am, y=mpg, fill=am))
p + geom_boxplot() + labs(x="Transmission Type", y="MPG") +
    scale_fill_discrete(name="Transmission Type") +
    theme(axis.text.x = element_text(angle=45, hjust=1, vjust=1, size=8),
          axis.text.y = element_text(size=8), axis.title=element_text(size=9,face="bold"))
```



Looking at this plot alone might lead to conclusion that manual transmission has better mpg compare to automatic transmission. We will try to build a model based on this assumption, and see how well it performs and whether transmission type itself can be a good predictor on the mpg of a car.

Single Variant Regression

In this first model we will try to fit the transmission type as a predictor for a car's mpg.

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

```
##
## Call:
## lm(formula = mpg ~ am, data = fmtcars)
##
## 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 model we can say that in average, given that all other factors do not change, an automatic transmission underperform a manual one by 7.245 mpg. But looking carfuly we can see that the adjusted R-squared value is 33.85%, meaning that only 33.85% of the variation of the data can be explained by looking on transmission type alone.

This low adjusted R-squared value tells us that a model based on only on transmission type is not suffice to predict the mpg of a car.

Multi Variant Regression

In order to decide which predictors we should pick for our model, let's take a look at the correlation of mpg to each of the other variabls in the mtcars data set:

```
sort(cor(mtcars)[1,])
```

```
##         wt        cyl       disp         hp       carb       qsec
## -0.8676594 -0.8521620 -0.8475514 -0.7761684 -0.5509251  0.4186840
##       gear         am         vs       drat        mpg
##  0.4802848  0.5998324  0.6640389  0.6811719  1.0000000
```

From these results, it looks like there is a strong relationship between the weight of a car to its mpg. Also, from the graphs in Appendix-1, it is possible to see that for all the variables, besides qsec, there is a strong

separation in values between manual and automatic transmission. This makes qsec (1/4 mile time) a very interesting variable as it is seems non-dependant in the other variables (in constrast for example to cyl and wt which seem very correlated).

We will add wt and qsec to our base model and see how it impacts the base model:

```
fit1<-lm(mpg~am,data=fmtcars)
fit3<-update(fit1,mpg~am+wt)
fit5<-update(fit1,mpg~am+wt+qsec)
anova(fit1,fit3,fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + qsec
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 73.203 2.673e-09 ***
## 3     28 169.29  1    109.03 18.034 0.0002162 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the anova output, one can see that each of the extra variables is very significant (high RSS change, low P-value) to the model. It also means that we reject the null hypothesis that our new modal is similar to the single-variable model. Now that we belive we have a strong model we will check how well it performs.

```
multiFit <- lm(mpg~am+wt+qsec, data = fmtcars)
summary(multiFit)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec, data = fmtcars)
##
## 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
## amManual      2.9358     1.4109   2.081 0.046716 *
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## 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
```

Our new model has an adjusted R-squared value of 83.36% which is significantly higher than our basic model. It also has a siginificant lower p-value, 1.21e-11, which means that extreme results using this model are far

less common. Also, our multi-variant model residual plot (Appendix-2) shows that the points randomly dispersed around the horizontal axis, meaning a linear regression model is appropriate for the data. All these makes this multi-variant model a stronger thesis.

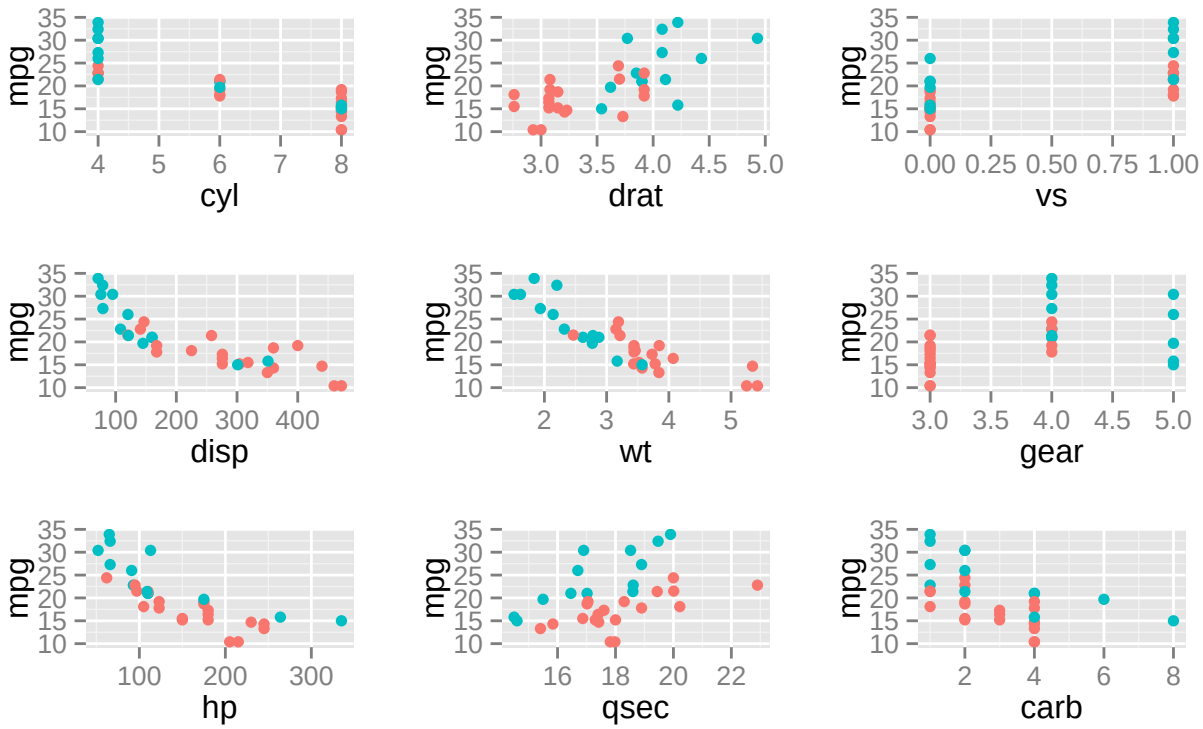
Regression Results Summary

As we take a look at the impact of transmission type on mpg using our last model, we can see that manual transmission type makes an average car mpg higher by only 2.9358 mpg. This is about 60% less our base analysis.

Appendix

Appendix-1

```
variables <- c("cyl","disp","hp","drat","wt","qsec","vs","gear","carb")
plots <- list()
for (i in 1:length(variables)) {
    p<-ggplot(fmtcars, aes_string((variables[i]), "mpg", col="am")) +
        geom_point() + theme(legend.position='none')
    plots[[i]]<-p
}
multiplot(plotlist = plots,  cols = 3)
```



The code for the multiplot function is taken from: http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/

Appendix-2

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