Study of the relationship of MPG and Transmission type

Shallu Arora

March 11, 2018

Click here to go to project on Github

Executive Summary

This report aims to find answers to the following 2 questions that a magazine "Motor Trend" has:

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

Thus, in this report, we explore the effects of manual and automatic transmission on the fuel efficiency (measured in "miles per gallon") of a set of 32 automobile models from 1973-74. The data was extracted from the 1974 *Motor Trend* US magazine, and is available from the R Package *datasets* (version 3.4.1) The analysis of this data is composed of three parts: (i) *Exploratory Data Analysis*, in which the data is loaded, preprocessed, and subject to an initial graphical examination; (ii) *Regression Analysis*, in which a linear model is fit to the data. This part also contains discussions on model selection, validation (by residual analysis), and interpretation of the relevant regression coefficients; and (iii) *Appendix*, wherein the plots that are used to support the discussion throughout this report are presented.

Exploratory data analysis

```
library (datasets)
library(ggplot2)
data(mtcars) # Loading data
fc<-c(2,8:11)
for (i in 1:length(fc)){mtcars[,fc[i]]<-as.factor(mtcars[,fc[i]])}
levels(mtcars$am) <- c("Automatic","Manual")
```

To get an initial feel for the relationships between the variables - and, in particular, between **mpg** and **am** - it is interesting to observe the scatterplots produced by plotting each variable against all others, as well as the specific distribution of **mpg** values within each level of **am**. The plots are shown in the **Appendix**, figures 1 and 2.

Two facts are immediately clear from the plots: first, **mpg** tends to correlate well with many of the other variables, most intensely with **drat** (positively) and **wt** (negatively). It is also clear that many of the variables are highly correlated (e.g., **wt** and **disp**). Second, it seems like manual transmission models present larger values of **mpg** than the automatic ones.

Inference

At this point, let's test null hypothesis that difference in MPG of the automatic and manual transmissions is zero.

```
result <- t.test(mtcars$mpg ~ mtcars$am)
result$p.value # results hidden
```

Since the p-value is 0.00137, we reject our null hypothesis. Now, let's calculate the difference in MPG between the automatic and manual transmissions.

```
result$estimate # results hidden
```

The difference in estimate between the 2 transmissions is 7.24494 MPG that too in the favor of manual.

Regression Analysis

First, let's fit the full model as the following.

```
fullModel <- lm(mpg ~ ., data=mtcars)
summary(fullModel) # results hidden
```

This model has the Residual standard error as 2.833 on 15 degrees of freedom. And the Adjusted R-squared value is 0.779, which means that the model can explain about 78% of the variance of the MPG variable. However, none of the coefficients are significant at 0.05 significant level.

Now, lets use backward selection to select some statistically significant variables.

```
stepModel <- step(fullModel, k=log(nrow(mtcars)))
summary(stepModel) # results hidden
```

This model is "mpg ~ wt + qsec + am". It has the Residual standard error as 2.459 on 28 degrees of freedom. And the Adjusted R-squared value is 0.8336, which means that the model can explain about 83% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level.

Please refer to the **Appendix** figure 3. According to the scatter plot, it indicates that there appear to be an interaction term between "wt" variable and "am" variable, since automatic cars tend to weigh heavier than manual cars. Thus, we have the following model including the interaction term:

```
amIntWtModel<-lm(mpg ~ wt + qsec + am + wt:am, data=mtcars)
summary(amIntWtModel) # results hidden
```

This model has the Residual standard error as 2.084 on 27 degrees of freedom. And the Adjusted R-squared value is 0.8804, which means that the model can explain about 88% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level. This is a pretty good one.

Next, we fit the simple model with MPG as the outcome variable and Transmission as the predictor variable.

```
amModel<-lm(mpg ~ am, data=mtcars)
summary(amModel) # results hidden
```

It shows that on average, a car has 17.147 mpg with automatic transmission, and if it is manual transmission, 7.245 mpg is increased. This model has the Residual standard error as 4.902 on 30 degrees of freedom. And the Adjusted R-squared value is 0.3385, which means that the model can explain about 34% of the variance of the MPG variable. The low Adjusted R-squared value also indicates that we need to add other variables to the model.

Finally, we select the final model.

```
anova(amModel, stepModel, fullModel, amIntWtModel)
```

We end up selecting the model with the highest Adjusted R-squared value, "mpg ~ wt + qsec + am + wt:am".

```
summary(amIntWtModel)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.723053  5.8990407  1.648243 0.1108925394
## wt          -2.936531  0.6660253 -4.409038 0.0001488947
## qsec         1.016974  0.2520152  4.035366 0.0004030165
## amManual    14.079428  3.4352512  4.098515 0.0003408693
## wt:amManual -4.141376  1.1968119 -3.460340 0.0018085763
```

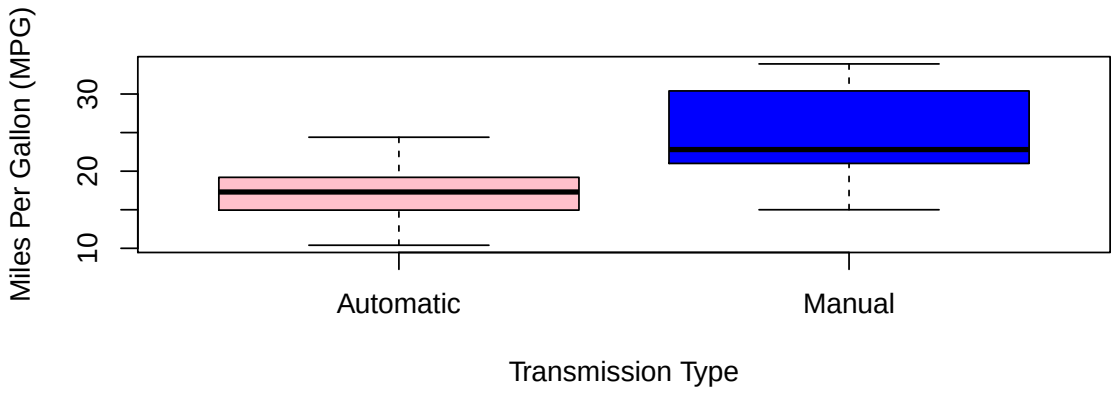
Thus, the result shows that when "wt" (weight lb/1000) and "qsec" (1/4 mile time) remain constant, cars with manual transmission add 14.079 + (-4.141)*wt more MPG (miles per gallon) on average than cars with automatic transmission. That is, a manual transmitted car that weighs 2000 lbs have 5.797 more MPG than an automatic transmitted car that has both the same weight and 1/4 mile time.

Residual Analysis and Diagnostics

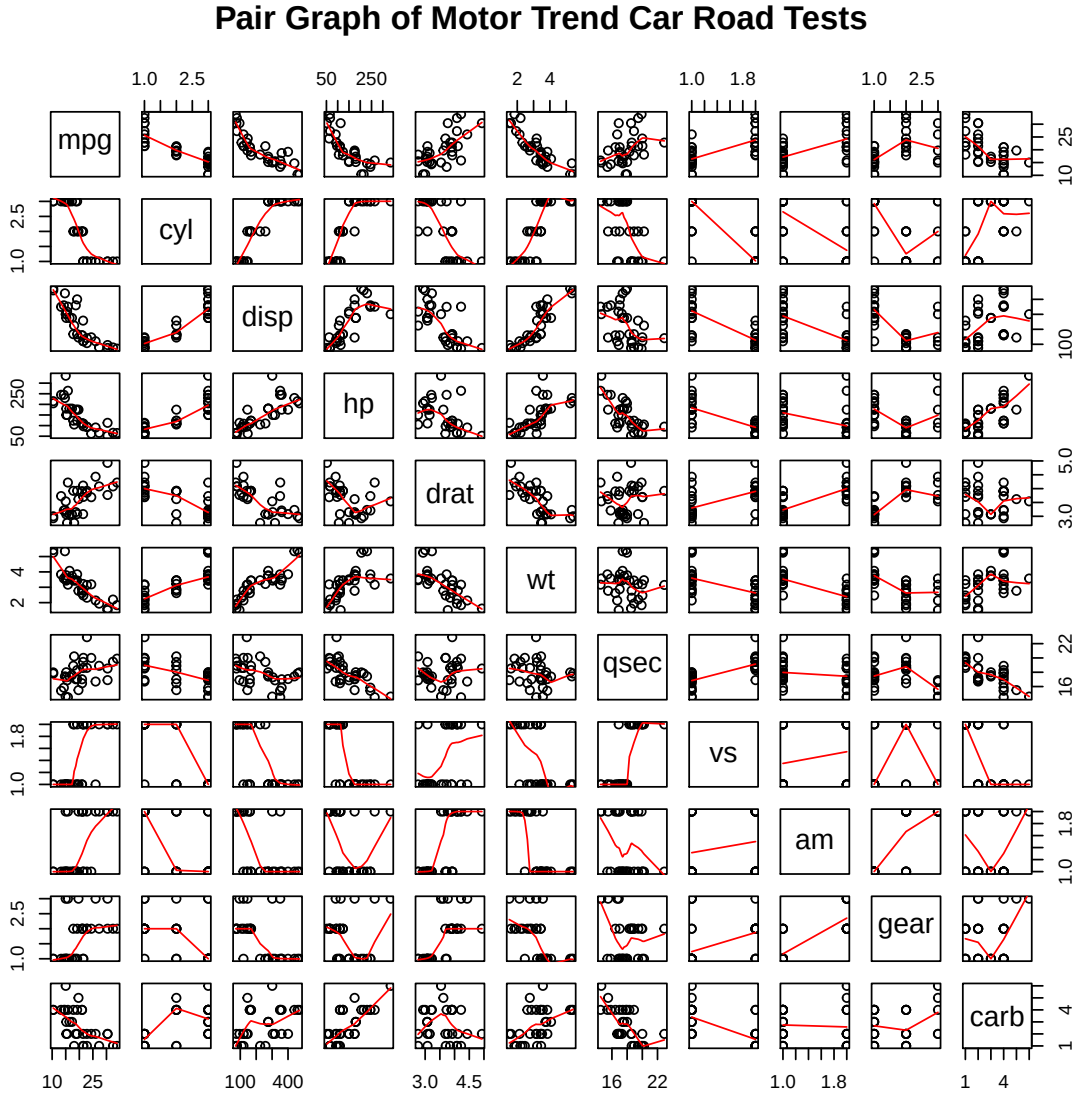
Please refer to the **Appendix** figure 4. According to the residual plots, we can verify the following underlying assumptions:
1. The Residuals vs. Fitted plot shows no consistent pattern, supporting the accuracy of the independence assumption.
2. The Normal Q-Q plot indicates that the residuals are normally distributed because the points lie closely to the line.
3. The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed.

Appendix: Figures

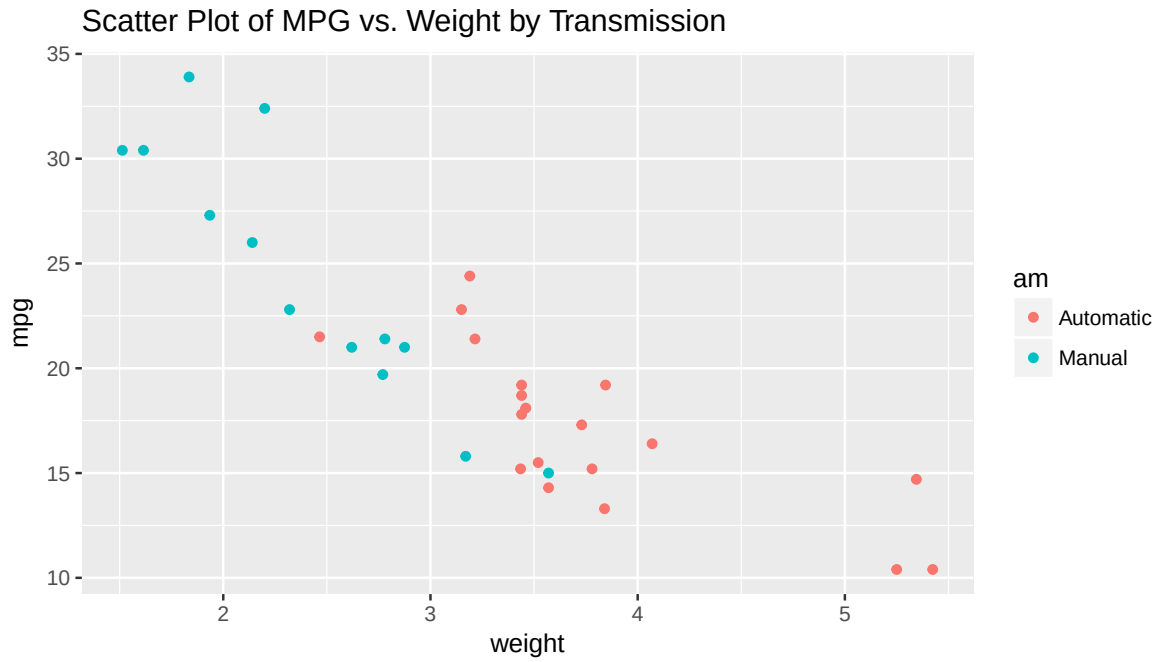
1. Boxplot of MPG vs. Transmission



2. Pair Graph of Motor Trend Car Road Tests



3. Scatter Plot of MPG vs. Weight by Transmission



4. Residual Plots

```
par(mfrow = c(2, 2))
plot(amIntWtModel,pch=16,lty=1,lwd=2)
```