Regression Models Course Project

Report of the analysis on the MPG difference between automatic and manual transmissions

Kenny Ong

June 11, 2017

Executive Summary

This course project ("project") is part of John Hopkins University Data Science Specialization Course 7 via Coursera (Online).

Problem Statement

You work for Motor Trend, a magazine about the automobile industry. 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:

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

Work Performed - Exploratory Data Analysis

The dataset can be found from the dataset package. Firstly, load the dataset and undertand the variables.

```
library(datasets)
data(mtcars)
?mtcars
```

According to documentation the data was extracted from 1974 Motor Trend US Magazine, and comprises fuel consumption and other performance information in respect of 32 car manufacturers. Consistent to the initial exploratory on the dataset, it appears that the dataset contains 32 observations and 11 variables.

We also performed other exploratory analsis on the dataset using boxplot to summarize the the dataset by miles per gallon (mpg). The output of the boxplot analysis can be found on Appendix 1 of this report.

```
boxplot(mpg ~ am, data = mtcars, xlab = "Transmission type", ylab = "Miles per gallon")
```

It appears from the boxplot analysis that cars with automatic transmission regardless of manufacturer have lower mpg than cars with manual transmission.

We also performed pairwise analysis using scatter plot among all the 11 variables on the dataset as presented on Appendix 2 of this report. It appears that the weight (wt) is very negatively correlated with mpg, meaning that the heavier the car, the lower its mpg.

```
pairs(mtcars)
```

Work Performed - Statistical Inference Analysis

Before we performed a regression modeling on the dataset, we performed analysis using t-test to confirm that the fuel efficiency (mpg) for automatics and manual transmission cars are different by statistics.

```
t.test(mtcars$mpg ~ mtcars$am)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg by mtcars$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
```

From the t-test analysts, it appears that the p-value computed is 0.001374. That proves that the fuel efficiency of cars (mpg) for automatic and manual transmission cars are indeed different. That is because if the p-value from the test is less than (<0.05) it means the the difference being tested is significant.

Work Performed - Regression Modeling

We first built a base model to predict mpg based on only the transmission type (am) as predictor.

```
baseModel <- lm(mpg ~ am, data = mtcars)
```

The result shows that the model is inadequate as the calculated R-squared value is 0.34, indicating that only 34% of the variance can be explained.

We then adopted the Backward Stepwise Regression method to design our desired model. We first created a full model called fullModel, and performed the backward stepwise regression method using the fullModel and called it stepModel.

```
fullModel <- lm(mpg ~ ., data = mtcars)
stepModel <- step(fullModel, direction="backward", k=2, trace=0)
```

When building our best model, we manually removed and added variables to evetually decide the best models to be built on quater mile time (qsec), weight (wt) and transmission type (am). We call our model bestModel. The summary of the best model is presented at Appendix 3 of this report.

```
bestModel <- lm(mpg ~  qsec + wt + am, data = mtcars)
```

Finally, we performed variance analysis (ANOVA) to test our best model. The analysis of variance shows that p-value is much lower than 0.05, hence we can conclude that our best model is statisticaly significant. The ANOVA results are presented at Appendix 4 of this report.

```
anova(fullModel, bestModel)
```

Work Performed - Residual Diagnostics

Please find the residual diagnostics plot of our best model at Appendix 5 of this report.

```
testModel <- lm(mpg ~ qsec + wt + am, data = mtcars)

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

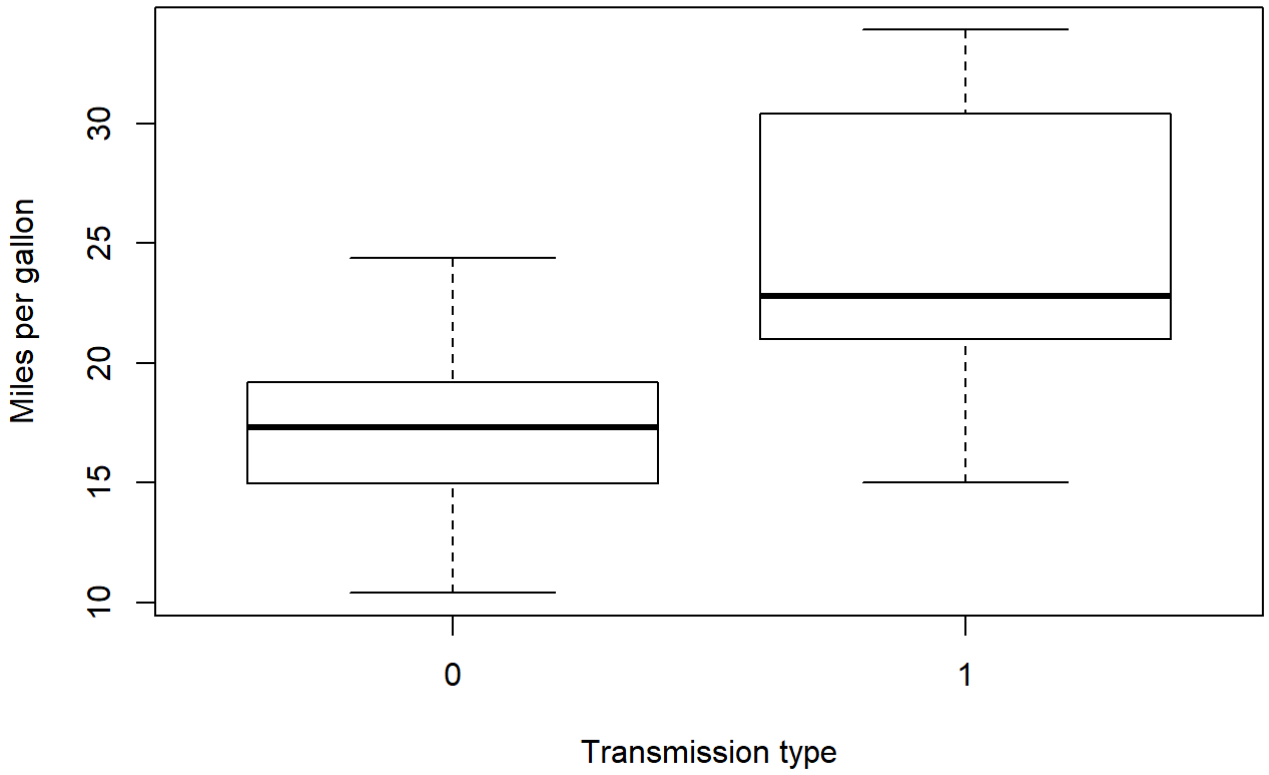
Conclusion

Based on our analysis, the p-value computed from t-test is 0.001374<0.05, we reject the null hypothesis that there is no difference in MPG influenced by other variables such as the transmission type (am).

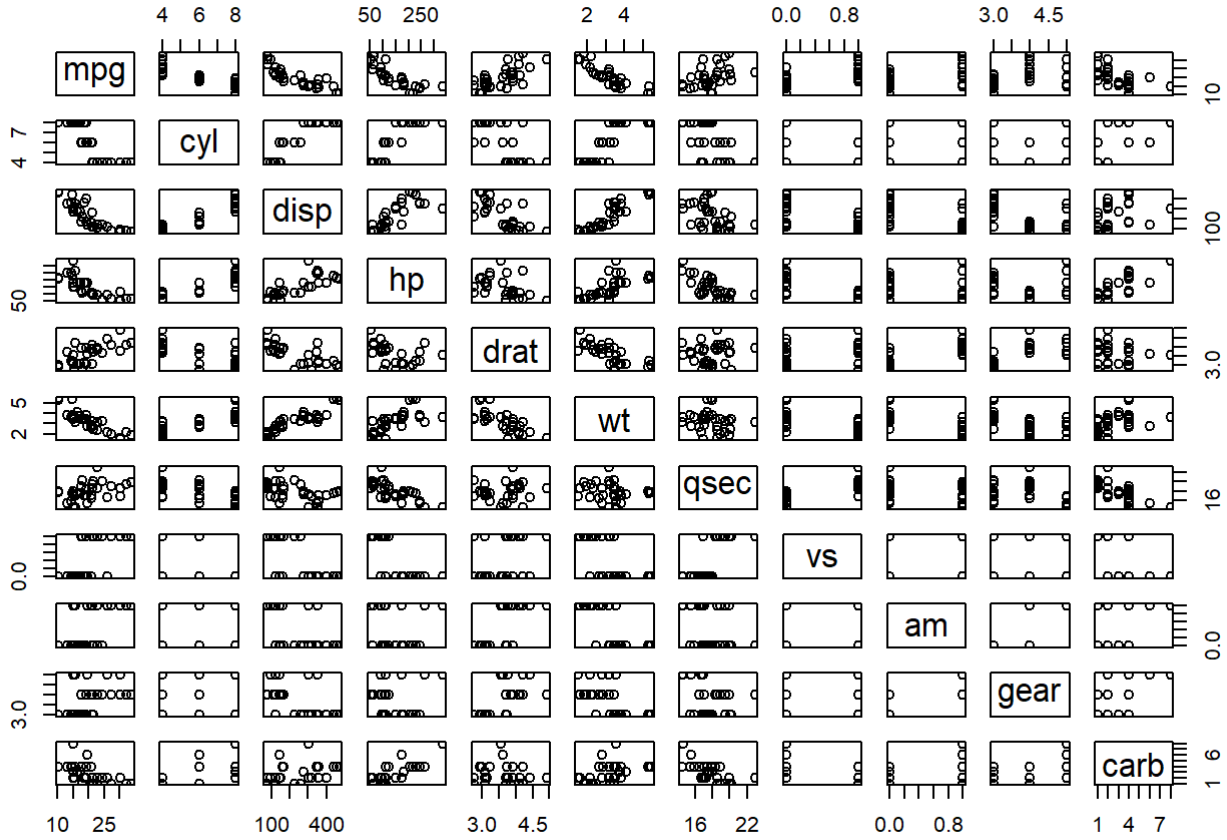
Based on our model, it is concluded that fuel efficiency of cars (mpg) is a function of the cars' quater mile time, weight and the type of transmission. On average, cars with manual transmission are better than cars with automatics transmission by 2.9358mpg.

Appendices

Appendix 1: Boxplot of fuel efficiency by transmission



Appendix 2: Pairwise analysis on all the dataset variables



Appendix 3: Summary of best model

```
##
## Call:
## lm(formula = mpg ~ qsec + wt + 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
## qsec          1.2259     0.2887   4.247 0.000216 ***
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## 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
```

Appendix 4: Analysis of Variance Table (ANOVA) on full model and best model

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

Appendix 5: Residual Dianostics of best model