Analysis on the mtcar dataset

Karla Nunez

Tuesday, July 14, 2015

Executive Summary

In this document we will describe the relationship between a set of variables and the fuel efficiency of 32 cars in miles/gallon. For this purpose we will use the data extracted from the 1974 Motor Trend US magazine. The 2 questions we will address are:
Is an automatic or manual transmission better for MPG?
Quantify the MPG difference between automatic and manual transmissions.

Brief analysis on the data

The dataset corresponds of 32 observations of 11 variables. Since the intention of this study is to understand relationships between variables that could have an impact on fuel efficiency we created a correlation matrix mapping each pairwise combination of variables `abs_cor_mtcars` (Appendix A). A reference dataframe `df_cor_mtcars`will be used throughout this document.

```
library(dplyr)
abs_cor_mtcars <- abs(cor(mtcars))
df_cor_mtcars <- as.data.frame(abs_cor_mtcars) %>%
  mutate(variable=rownames(abs_cor_mtcars))
```

Question 1: Is an automatic or manual transmission better for MPG?

```
mpg_auto <-mtcars[mtcars$am==0,c("mpg")]
mpg_manual <- mtcars[mtcars$am==1,c("mpg")]
```

An simple boxplot (Appendix B) can easily give us the answer we are looking for, in this case we can see that automatic transmission as a mean of 17.15MPG which is worse than the mean of 24.39MPG for manual transmission. However, we should create a hypothesis test to prove this.
H_0=There is no significant difference beween the cars with automatic and manual transmission. (Appendix C)

```
t_test_transmission <- t.test(mpg_auto, mpg_manual)
```

With a p-value of 0.0013736 we can reject the null hypothesis, therefore the conclussion is that there is a significant difference between the mileage on a car based on its transmission. given that the mean efficiency of manual transmission cars is better than that of automatic transmission cars we conclude that **manual transmission is better than automatic transmission for MPG when no other factors are considered**. If we want to take into account all of the factors that can contribute to this, we will have to look into a multi-variable regression model, explained in the next question.

Question 2: Quantify the MPG difference between automatic and manual transmissions.

Our first step is to understand the relationship between mpg and the transmissions. (Appendix D)

```
mpg_am_model <- lm(mpg~am, mtcars)
sum_mpg_am_model <- summary(mpg_am_model)
```

Given that R^2 has a value of 0.3597989 we can only explain 35.98% of the variance, therefore we need to create multi-variable linear regressions to get a more accurate estimate by adding all of them in selectively removing those with low statistical significance. This will be done using a stepwise search

```
mpg_all_model <- lm(mpg~., mtcars)
best_model <- step(mpg_all_model, trace=0)
sum_best <- summary(best_model)
sum_best
```

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

With the information above we can see that in this case R^2 has a value of 0.8496636 which explains 84.97 of the variance.

In order to conclude we need to ensure that the variables we are choosing are not contributing to the model, for this we will consider H_0 = variables wt and qsec are not contributing to the model.

```
an_var <- anova(best_model, mpg_am_model)
an_var
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + qsec + am
## Model 2: mpg ~ am
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
```

```
## 1     28 169.29
## 2     30 720.90 -2   -551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

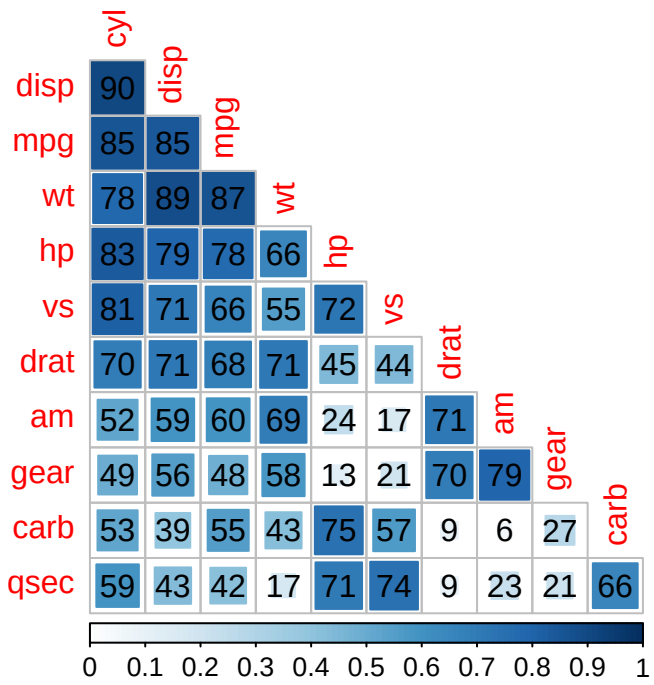
with a p-value of we can reject the null hypothesis, therefore concluding that the **variables wt, and qsec are contributing to the model**.

The remaining step consists of ensuring the residuals appear normally distribued and don't show signs of heteroskedasticity which can bee seen in appendix E. With these graphs we can see that the variables are independent and are normally distributed.

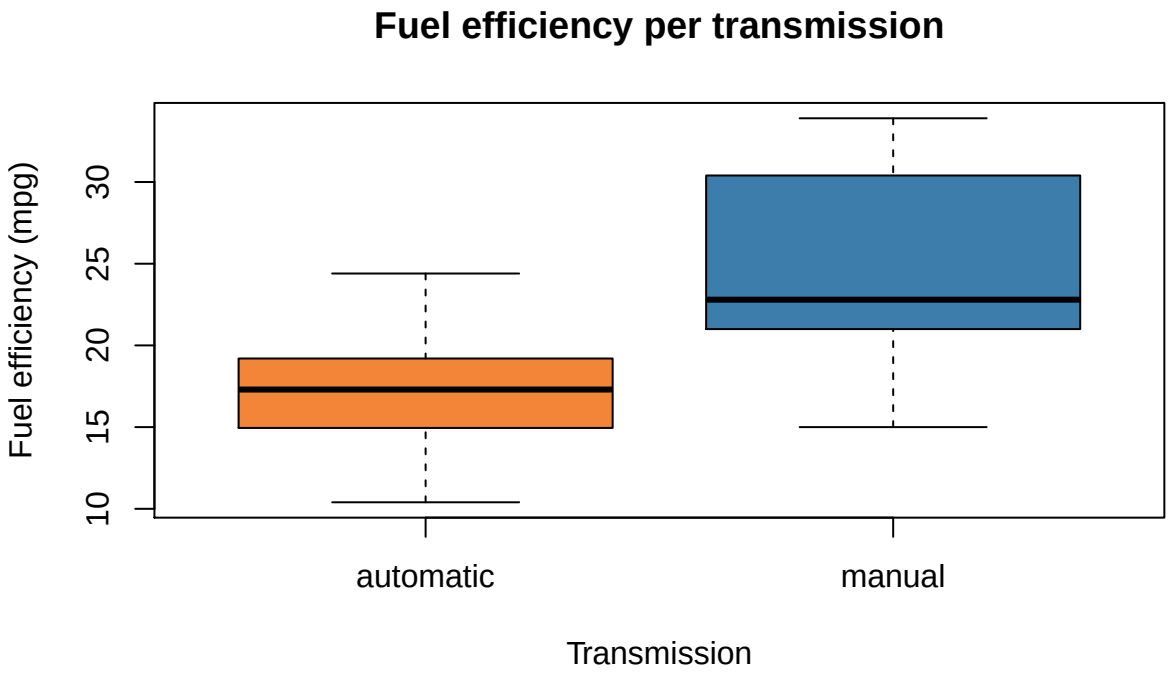
With all the information above we can conclude that **there is a difference in mileage between cars with manual transmission vs the ones with automatic transmission but there are other variables like the car's weight and qsec (1/4 mile time) contributing to these numbers. With our best model we can conclude that manual transmission cars can provide `2.9358` more MPG than their automatic counterparts.**

Appendix

Appendix A: Correlation matrix



Appendix B: Boxplot of fuel efficiency per transmission



Appendix C: t-test between automatic and manual transmission for MPG

```
t_test_transmission
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg_auto and mpg_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
```

Appendix D: linear regression of mpg on transmission

```
sum_mpg_am_model
```

```
##
## Call:
```

```
## lm(formula = mpg ~ am, data = mtcars)
##
## 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 ***
## am             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
```

Appendix E: Residual plots

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