Relationship Between MPG and Transmission

Jessica Bohning

February 24, 2017

Executive Summary

This anlaysis seeks to answer the questions: 1) Is an automatic or manual transmission better for MPG?; and 2) Quantify the MPG difference between automatic and manual transmissions. It will be shown that a manual transmission is better for MPG than an automatic transmission and that the automatic transmission gets at least 3.91 mpg less

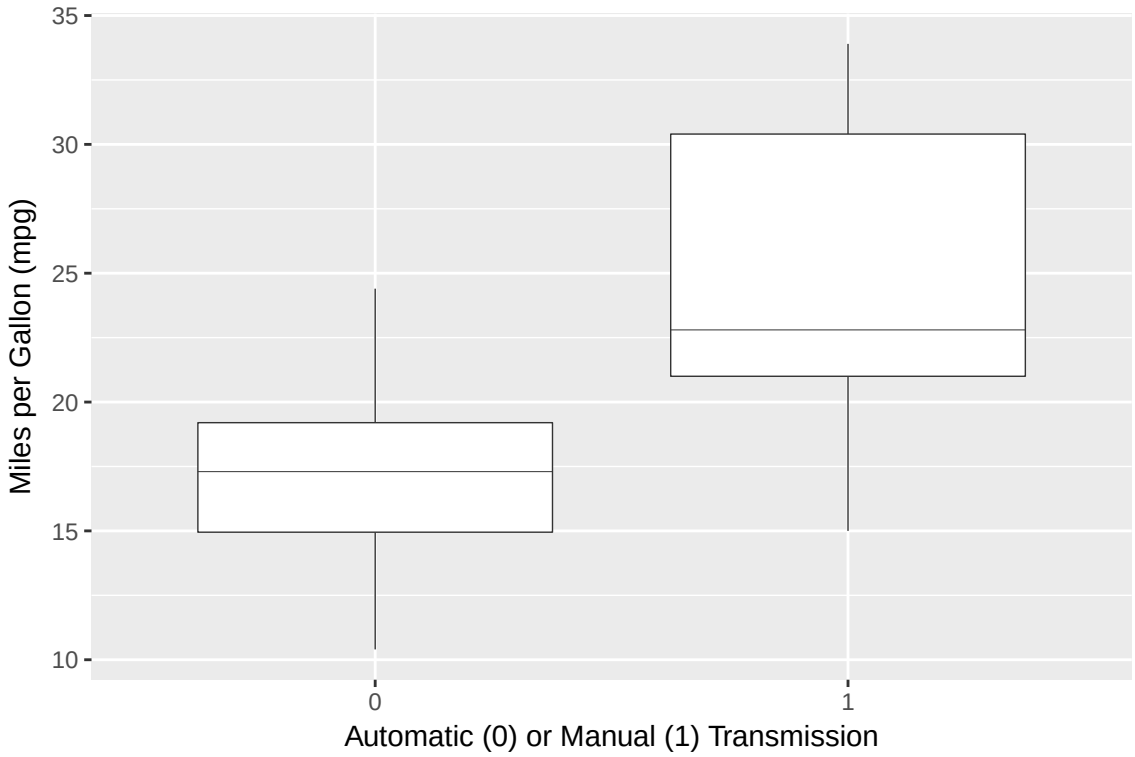
Exploratory Analysis

The following breaks what each variable referenced in this analysis stands for:

Variable	Definition
mpg	miles/(US) gallon
cyl	number of cylinders
disp	displacement (cu.in.)
hp	gross horsepower
drat	rear axle ratio
wt	weight (1000 lbs)
qsec	0.25 mile time
vs	V/S
am	Transmission (0 = automatic, 1 = maunal)
gear	number of forward gears
carb	number of carburetors

Before starting the data analysis, let's get a brief glimpse of the data using a boxplot.

```
library(UsingR)
library(ggplot2)
ggplot(mtcars,aes(as.character(mtcars$am),mtcars$mpg))+
        geom_boxplot(lwd=0.25,fatten=0.25,outlier.size =0.25)+
        labs(x="Automatic (0) or Manual (1) Transmission")+
        labs(y="Miles per Gallon (mpg)")
```



This graph shows that the automatic tranmission (value of 0) gets lower gas mileage than a manual transmission (1). However, is the difference statistically significant? A t test should be able to determine if the means of the two sets of data is different. The null hypothesis is that the automatic (0) and manual (1) transmissions have the same mean MPG. The alternative hypothesis is that automatic transmissions have a smaller mean MPG than manual transmissions

Testing for Significance

To test for statistical significance, separate the data into automatic and manual (automatic has an am value of 0 and manual has an am value of 1)

```
automatic<-mtcars[mtcars$am==0,]
manual<-mtcars[mtcars$am==1,]
t.test(automatic$mpg,manual$mpg,alternative="less",paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  automatic$mpg and manual$mpg
## t = -3.7671, df = 18.332, p-value = 0.0006868
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -3.913256
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

The confidence interval does not include 0, and the p-value is less than 0.05. Therefore the null hypothesis is rejected and we can conclude that manual transmissions get better gas mileage. The 95% confidence interval shows that the automatic transmissions has a mean MPG that is lower than the manual by negative infinity

to -3.91 mpg. Obviously, in the context of the problem, the mean mpg can only be so large and cannot reach negative infinity. Using the max and min mpg of the data, the largest possible difference is 23.5 mpg

Next, the difference between automatic and manual transmissions will be quantified. First, I will look at a summary of all of a linear model using all of the variables and then I will graph all of the other variables vs. mpg to get an idea of the trends. The graph can be found in the Appendix as Figure 1

```
#Summary of linear model using all variables
summary(lm(mpg~.,data=mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

The summary table shows that "wt" is the most signicant variable on mpg (p-value of 0.063). From the graphs, cyl, disp and hp have an impact. There are multiple variabes that could be considered and so mutliple models will be looked at. I will add in the variables starting with transmission and then based on the p-value (from smallest to largest). Then an anova table will be created to determine which model might be best.

```
fit1<-lm(mpg~am,mtcars)
fit2<-lm(mpg~am+wt,mtcars)
fit3<-lm(mpg~am+wt+qsec,mtcars)
fit4<-lm(mpg~am+wt+qsec+hp,mtcars)
fit5<-lm(mpg~am+wt+qsec+hp+disp,mtcars)
fit6<-lm(mpg~am+wt+qsec+hp+disp+drat,mtcars)
fit7<-lm(mpg~am+wt+qsec+hp+disp+drat+gear,mtcars)
fit8<-lm(mpg~am+wt+qsec+hp+disp+drat+gear+carb,mtcars)
fit9<-lm(mpg~am+wt+qsec+hp+disp+drat+gear+carb+vs,mtcars)
fit10<-lm(mpg~am+wt+qsec+hp+disp+drat+gear+carb+vs+cyl,mtcars)
anova(fit1,fit2,fit3,fit4,fit5,fit6,fit7,fit8,fit9,fit10)
```

```
## Analysis of Variance Table
```

```
##
## Model  1: mpg ~ am
## Model  2: mpg ~ am + wt
## Model  3: mpg ~ am + wt + qsec
## Model  4: mpg ~ am + wt + qsec + hp
## Model  5: mpg ~ am + wt + qsec + hp + disp
## Model  6: mpg ~ am + wt + qsec + hp + disp + drat
## Model  7: mpg ~ am + wt + qsec + hp + disp + drat + gear
## Model  8: mpg ~ am + wt + qsec + hp + disp + drat + gear + carb
## Model  9: mpg ~ am + wt + qsec + hp + disp + drat + gear + carb + vs
## Model 10: mpg ~ am + wt + qsec + hp + disp + drat + gear + carb + vs +
##     cyl
##    Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1      30 720.90
## 2      29 278.32  1    442.58 63.0133 9.325e-08 ***
## 3      28 169.29  1    109.03 15.5240 0.0007497 ***
## 4      27 160.07  1      9.22  1.3127 0.2648040
## 5      26 153.44  1      6.63  0.9438 0.3423662
## 6      25 150.09  1      3.34  0.4762 0.4977087
## 7      24 148.53  1      1.56  0.2228 0.6417693
## 8      23 147.84  1      0.69  0.0976 0.7578163
## 9      22 147.57  1      0.27  0.0382 0.8468579
## 10     21 147.49  1      0.08  0.0114 0.9160874
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The table shows that the addition of each variable improves the model. Therefore, in selecting a linear model, you would pick fit10: the model with all of the variables. However, some of the later models only barely improve the model so if there is a limiting factor in how many variables could be picked, you'd at least want to used model 4, which gets the F value just above 1 and the Sum of Squares below 10.

Furthermore, Figure 2 in the Appendix shows that the residuals are randomly distributed and are approximately normally distributed. This means that Fit10 meets the conditions for linear modeling (independent data that is normally distributed). The following shows that fit10 has an R^2 value of 0.869 meaning that the model explains about 87% of the variance.

```
summary(fit10)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec + hp + disp + drat + gear +
##     carb + vs + cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## am           2.52023    2.05665   1.225   0.2340
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## hp          -0.02148    0.02177  -0.987   0.3350
## disp         0.01334    0.01786   0.747   0.4635
## drat         0.78711    1.63537   0.481   0.6353
```

```
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## vs           0.31776    2.10451   0.151   0.8814
## cyl         -0.11144    1.04502  -0.107   0.9161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Appendix

Figure 1

```
mtcars2<-mtcars[,-9]#We already know what the "am" data looks like
par(mfrow=c(3,3),mar=c(4,4,2,1))
for(i in 1:9){
        plot(x=mtcars2[,i+1],y=mtcars$mpg,main=names(mtcars2[i+1]))}
```

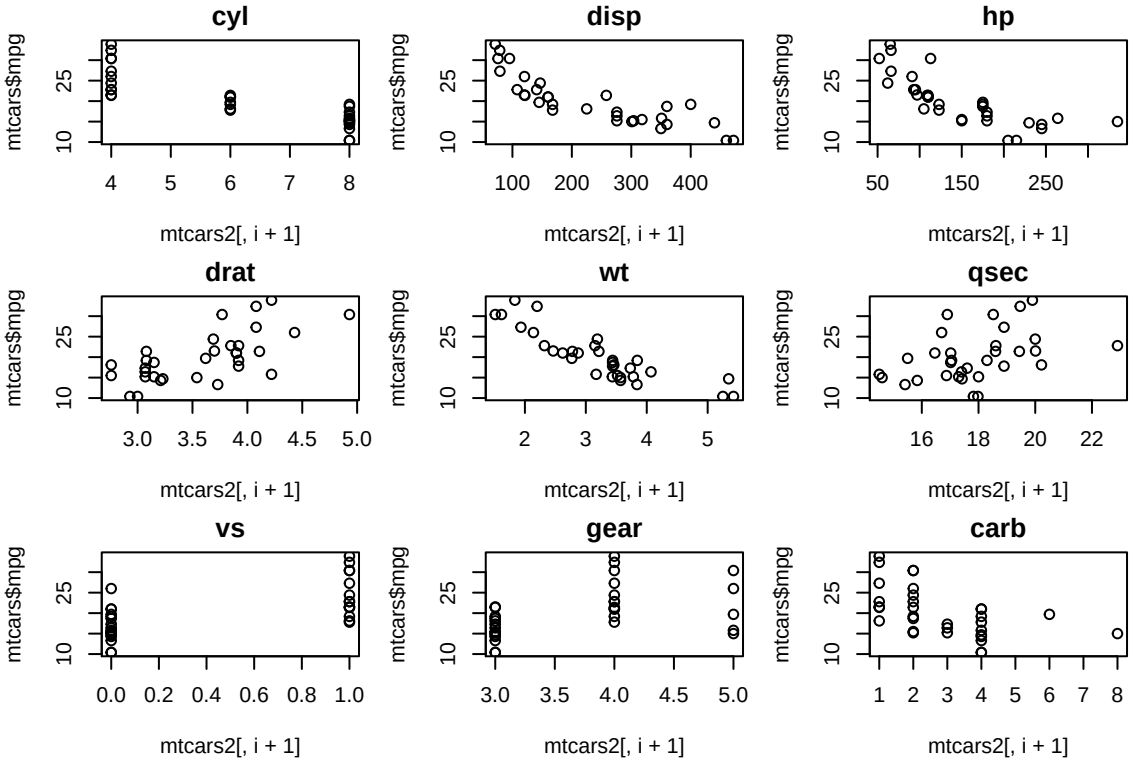
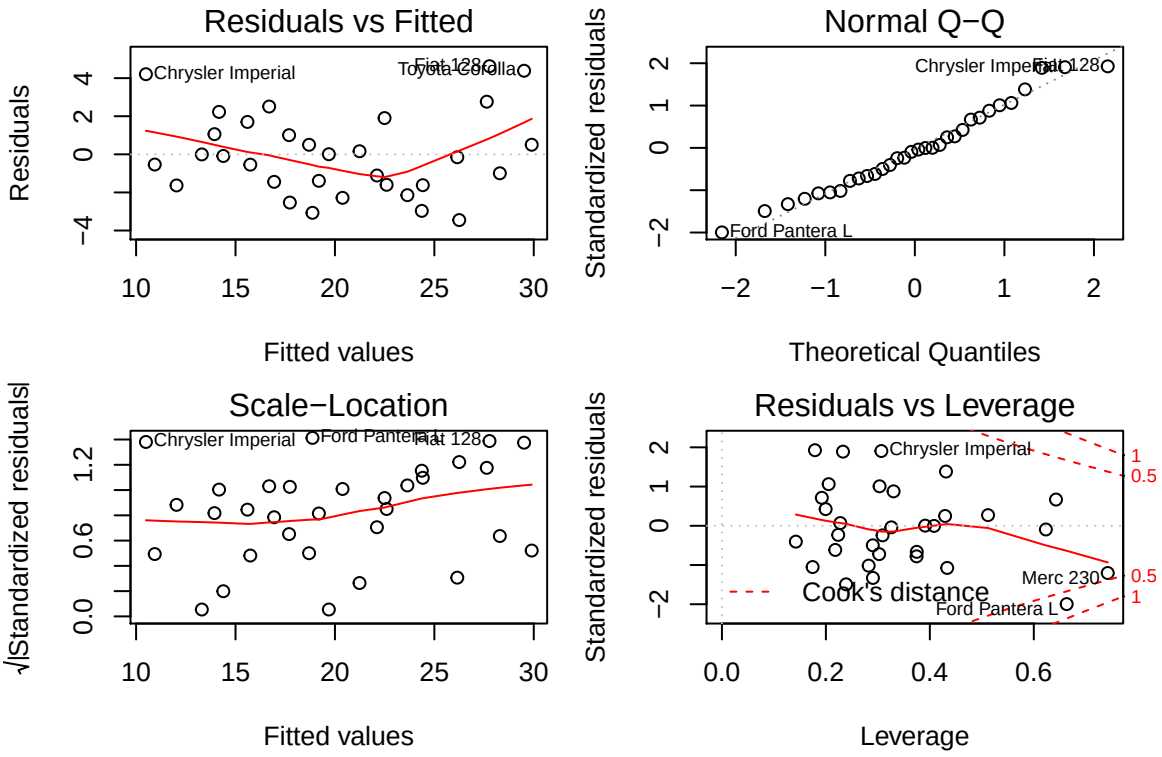


Figure 2

```
par(mfrow=c(2,2),mar=c(4,4,2,1));plot(fit10)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Fiat 128
Ford Pantera L
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Ford Pantera L
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
Chrysler Imperial
Merc 230
Ford Pantera L