# Analysis of MPG per Transmission Type

*Tyler Peterson*

*11/18/2016*

## Executive Summary

Using the dataset mtcars from Motor Trend, the objective of this analysis is to examine the dataset and explore the relationship between a set of variables and miles per gallon (MPG) (outcome). In particular, we seek to answer the following two questions:

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

## Exploratory Data Analysis

In the EDA step, we attempt to explore the data and understand what the data points are expressing about the dataset. Please refer to the appendix for a more in depth understanding of how we explored the data and established normality.

We examine the expected values of the transmission types to see if they are significant.

```
# Determine the expected values (mean) of each transmission type
transMean <- with(mtcars, tapply(mpg, am, FUN = mean))
```

We calculate that the difference between mean automatic transmission mpg ratings and manual transmission mpg ratings is -7.2449. We need to do a t-test to determine if this difference is significant enough to confidently state that automatic transmissions overall get less gas mileage than manual transmissions. The null hypothesis of our t-test is that the mean values are the same, with the alternative hypothesis being that they are different.

```
autoSub <- subset(mtcars, am == "Automatic")
manualSub <- subset(mtcars, am == "Manual")
amTTest <- t.test(autoSub$mpg, manualSub$mpg)
```

From the t-test, the resulting p-value is 0.0013736 with a confidence interval of -11.2801944, -3.2096842. This confidence interval does not include 0, so we conclude from our initial assumptions that if the only thing we were measuring was mpg vs. transmission type, we can expect automatic transmissions to have a worse mpg rating than manual transmissions.

## Fitting the Model

As we look at the dataset, we see that there are many more independent variables than just the transmission type, and all of those can influence the mpg rating. Please see the appendix for a correlation matrix that indicates how much each variable in the dataset influences the mpg rating.

We now fit the model to validate and improve upon our assumptions.

```
fitNoIntercept <- lm(mpg ~ am - 1, data = mtcars)
fitWithIntercept <- lm(mpg ~ am, data = mtcars)
summary(fitNoIntercept)$coef
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## amAutomatic 17.14737   1.124603 15.24749 1.133983e-15
## amManual    24.39231   1.359578 17.94109 1.376283e-17
```

When we exclude the intercept from the model, we see that our conclusions about the expected value of automatic transmissions is validated - that the expected mpg rating of automatic transmissions is lower than the manual transmissions. However, including the intercept in the model and checking the R-squared value, we calculate that this model only accounts for 35.98% of the variation in the dataset. A multivariate approach should get us a better representation of the dataset.

```
fitAllVars <- lm(mpg ~ ., data = mtcars)
str(summary(fitAllVars))
```

```
## List of 11
##  $ call         : language lm(formula = mpg ~ ., data = mtcars)
##  $ terms        :Classes 'terms', 'formula'  language mpg ~ cyl + disp + hp + drat + wt + qsec + vs +
##   .. ..- attr(*, "variables")= language list(mpg, cyl, disp, hp, drat, wt, qsec, vs, am, gear, carb)
##   .. ..- attr(*, "factors")= int [1:11, 1:10] 0 1 0 0 0 0 0 0 0 0 ...
##   .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. ..$ : chr [1:11] "mpg" "cyl" "disp" "hp" ...
##   .. .. .. ..$ : chr [1:10] "cyl" "disp" "hp" "drat" ...
##   .. ..- attr(*, "term.labels")= chr [1:10] "cyl" "disp" "hp" "drat" ...
##   .. ..- attr(*, "order")= int [1:10] 1 1 1 1 1 1 1 1 1 1
##   .. ..- attr(*, "intercept")= int 1
##   .. ..- attr(*, "response")= int 1
##   .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. ..- attr(*, "predvars")= language list(mpg, cyl, disp, hp, drat, wt, qsec, vs, am, gear, carb)
##   .. ..- attr(*, "dataClasses")= Named chr [1:11] "numeric" "numeric" "numeric" "numeric" ...
##   .. .. ..- attr(*, "names")= chr [1:11] "mpg" "cyl" "disp" "hp" ...
##  $ residuals    : Named num [1:32] -1.6 -1.112 -3.451 0.163 1.007 ...
##   ..- attr(*, "names")= chr [1:32] "Mazda RX4" "Mazda RX4 Wag" "Datsun 710" "Hornet 4 Drive" ...
##  $ coefficients : num [1:11, 1:4] 12.3034 -0.1114 0.0133 -0.0215 0.7871 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:11] "(Intercept)" "cyl" "disp" "hp" ...
##   .. ..$ : chr [1:4] "Estimate" "Std. Error" "t value" "Pr(>|t|)"
##  $ aliased      : Named logi [1:11] FALSE FALSE FALSE FALSE FALSE FALSE ...
##   ..- attr(*, "names")= chr [1:11] "(Intercept)" "cyl" "disp" "hp" ...
##  $ sigma        : num 2.65
##  $ df           : int [1:3] 11 21 11
##  $ r.squared    : num 0.869
##  $ adj.r.squared: num 0.807
##  $ fstatistic   : Named num [1:3] 13.9 10 21
##   ..- attr(*, "names")= chr [1:3] "value" "numdf" "dendf"
##  $ cov.unscaled : num [1:11, 1:11] 49.883532 -1.874242 -0.000841 -0.003789 -1.842635 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:11] "(Intercept)" "cyl" "disp" "hp" ...
##   .. ..$ : chr [1:11] "(Intercept)" "cyl" "disp" "hp" ...
##  - attr(*, "class")= chr "summary.lm"
```

We will not display the results from running a linear regression using all variables, it is sufficient to say that from using all of the variables, the R-squared value grows to 0.869, but the p-values for the data are all fairly high and none of them represent a good t-score. We will need to find the most descriptive variables that provide the best representation of the data. To do this, we use the step method in r where we step through 100 different linear models to find the best variables to use.

```
bestFit <- step(lm(data = mtcars, mpg ~ .), direction = "both", trace = 0, steps = 100)
summary(bestFit)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## amManual     2.935837  1.4109045  2.080819 4.671551e-02
```

From using the step model, we see that using the wt, qsec, and am variables, we get an accurate linear model that represents roughly 85% of the data.

## Conclusion

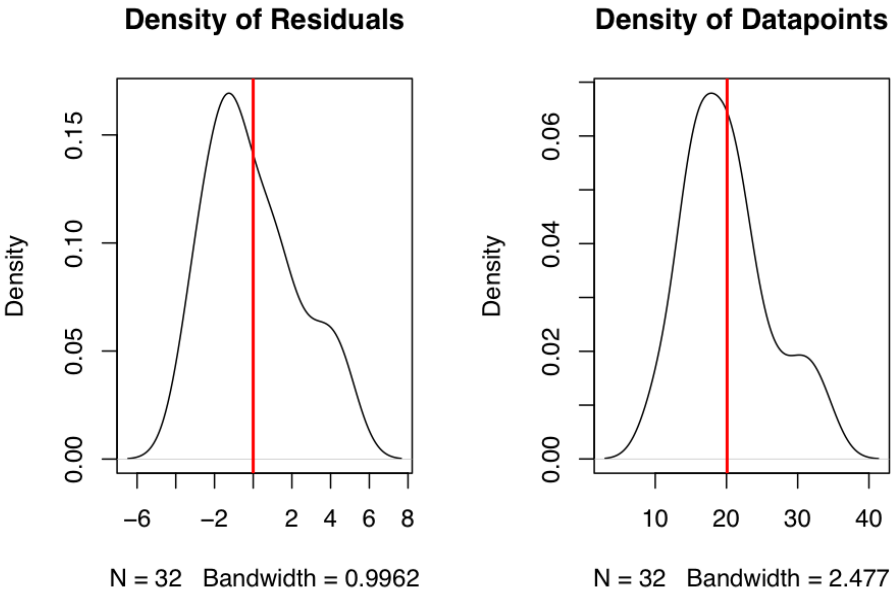
We compare our final multi-variate model to our initial single variate model using an ANOVA:

```
pander(anova(fitWithIntercept, bestFit))
```

Table 1: Analysis of Variance Table

| Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|
| 30 | 720.9 | NA | NA | NA | NA |
| 28 | 169.3 | 2 | 551.6 | 45.62 | 1.55e-09 |

From this test, we can see the p-value is sufficient to confidently reject the null hypothesis and conclude that the second model is superior. Furthermore, if we do a side-by-side density plot of the residuals from the fit model and the data points in the original dataset, we can see that they have nearly the same distribution.



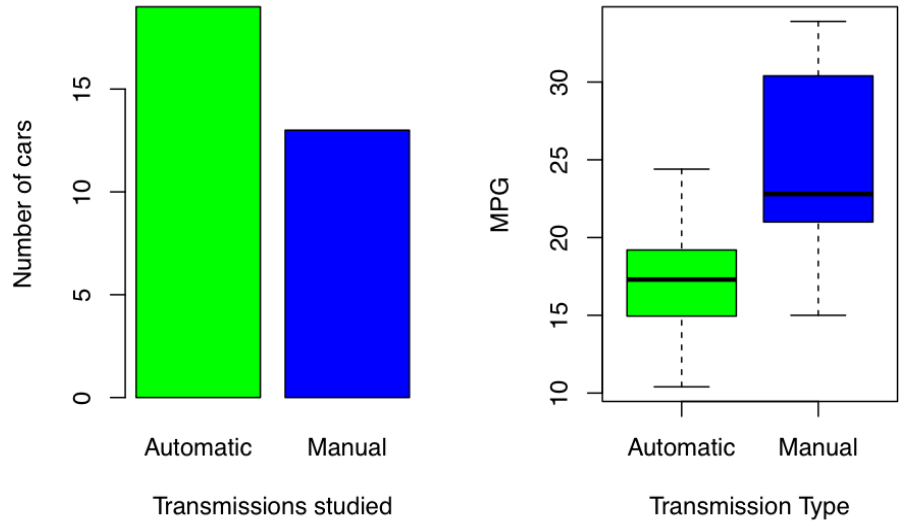

Therefore, we conclude that:

1. The multi-variate linear regression model closely models the actual data points in the dataset
2. There were other factors in the model that confounded the transmission selection and influenced it - namely the wt and qsec variables.
3. From the multi-variate linear regression we can say that on average, a manual transmission gets about 2.94 better mpg than automatic transmissions.

# Apendix

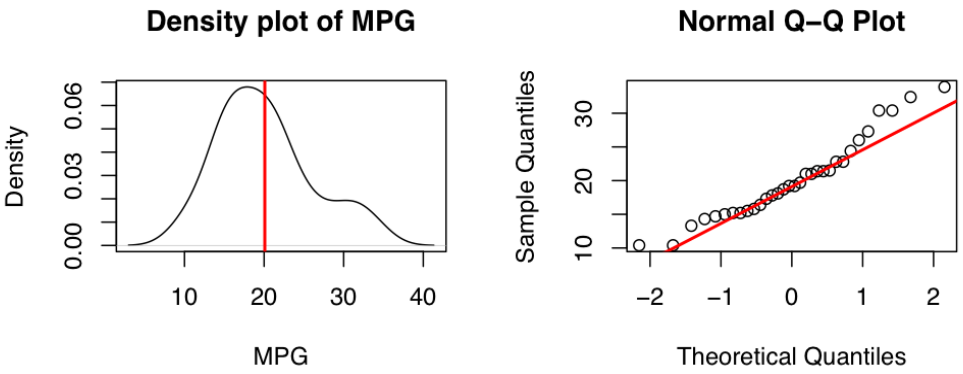
### Initial Assumpitions and Validation

We examine the number of cars in the study compared to their respective transmissions and also check their respective expected values.




From the histogram above, we can see that there are more automatic transmissions than manual transmissions in the study. This will likey skew any superficial suppositions we make about the data, thus requiring statistical testing to verify all assumptions. Furthermore, from the boxplot, it appears that we can assume that cars with an automatic transmission overall have a lower MPG rating. This confirms what we are attempting to determine.

### Datset test for normality.




The density plot seems to indicate that the data is not quite normally distributed. There appear to be a few outliers around the 30 to 40 mpg range that is skewing the data. Those data will need to be looked at closer

to see if we can rule them out or remove them from the analysis altogether. We run the Shapiro-Wilk test for normality with the null hypothesis that the data is normallly distributed

The Shapiro-Wilk test produces a p-value of 0.1228814 and we fail to reject the null hypothesis, but we created a QQ plot which gives us a good indication if the data follows a normal distribution.

We can see from the QQ plot that the mpg data does not appear to be normally distributed. Again, it appears that there are 3 or 4 cars in the 30+ mpg range that are skewing the data. Indeed we see that the mtcars dataset has the following cars with a mpg rating greater than 30

```
##                 mpg cyl disp  hp drat    wt  qsec vs     am gear carb
## Fiat 128       32.4   4 78.7  66 4.08 2.200 19.47  1 Manual    4    1
## Honda Civic    30.4   4 75.7  52 4.93 1.615 18.52  1 Manual    4    2
## Toyota Corolla 33.9   4 71.1  65 4.22 1.835 19.90  1 Manual    4    1
## Lotus Europa   30.4   4 95.1 113 3.77 1.513 16.90  1 Manual    5    2
```

We see that none of these have automatic transmissions. We therefore conclude that we cannot remove any data points from this dataset and accept that the linear model will likely not be able to account for this variance.

### Correlated Variables

Correlation between the different variables can be shown as follows:

```
data(mtcars)
mtcarsCor = t(as.data.frame(sort(cor(mtcars)[1,])[1:10]))
rownames(mtcarsCor) <- c("mpg")
corrplot(mtcarsCor, method = "number",
         title = "Correlation of MTCARS Variables to MPG",
         is.corr = FALSE, cl.pos = "n", mar = c(0,0,1,0))
```