# Regression Model - Course Project

*Jordan Woloschuk*

*9/21/2019*

## Motor Trend: Analysis on Variables and the Impact on Miles per Gallon

## 1) Executive Summary

### 1.1) Overview

The purpose of this analysis is to examine a collection of cars and determine the relationship between vechicle variables (e.g. number of cyclinders) and fuel economy (i.e. miles per gallon).

This analysis will seek to answer the following questions:

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

### 1.2) Conclusion

The analysis conducted indicates that holding all other variables equal, a car with manual transmission will have a higher MPG and therefore better fuel economy.

The *final_model* determined that a car's MPG was dependent on the transmission type, weight of the car and the acceleration of the car (1/4 mile time). This model indicates that cars with manual transmission will have ~2.94 more miles per gallon comapred to a car with an automatic transmissions. However, since cars with automatic transmission tend to weigh more compared to cars with manual transmission this interation could be influencing the final result.

## 2) Data Processing

### 2.1) Loading Libraries

Load necessary libraries for data analysis and developing results.

```r
library(ggplot2)
library(dplyr)
library(ggfortify) # ggfortify for autoplot
```

### 2.2) Loading Data

Load necessary mtcars dataset.

```r
data(mtcars)
```

### 2.3) Modifying Data

The variables “cyl” (# cyclinders), “vs” (engine shape), “am” (transmission), “gear” (# of gears) will be converted to factor variables since they are not continous.

```
mtcars_data <- mtcars # creating a new mtcars dataframe to be modified
mtcars_data$cyl <- as.factor(mtcars_data$cyl)

mtcars_data$vs <- as.factor(mtcars_data$vs)
# Set levels for the engine shape (V-shaped or Straight)
levels(mtcars_data$vs) <- c("V-shaped", "Straight")

mtcars_data$am <- as.factor(mtcars_data$am)
# Convert 0 = Automatic and 1 = Manual
levels(mtcars_data$am) <- c("Automatic", "Manual")

mtcars_data$gear <- as.factor(mtcars_data$gear)
```

## 3) Exploratory Data Analyses

```
# Note following table and dim output hidden due to page limit. See Table 1.
head(mtcars_data,2) # Sample of the first 4 rows of data
dim(mtcars_data) # Dimensions of the mtcars_data
```

In order to understand the relationship between transmission type and fuel economy we will develop a box plot to show the impact automatic/manual has on MPG. Please refer to **Fig. 1** in the Appendix.

After reviewing **Fig. 1**, it can be seen that manual transmissions results in higher MPG. However, it is not clear if other variables are also influencing this outcome, and if manual transmission vechicles have higher MPG with all else equal. To do so a pair graph will be developed. Please refer to **Fig. 2**.

After reviewing **Fig. 2**, it can be seen that other vaiables have correlation; including “wt” (weight), “disp” (displacement), “cyl” (# of cyclinders), and “hp” (horsepower).

## 4) Modelling

### 4.1) Linear Regression Model

We will first develop a basic linear model to examine transmission type vs. MPG.

```
lm_linear <- lm(mpg ~ am, mtcars_data) # Basic linear model
summary(lm_linear)$coefficients[,4] # P-value coefficient
```

```
##  (Intercept)     amManual
## 1.133983e-15 2.850207e-04
```

```
summary(lm_linear)$r.squared
```

```
## [1] 0.3597989
```

Since the p-value is quite small and the R-quared value is equally small, it indicates that additional variables are needed to accurately model the relation between transmission type and MPG.

### 4.2) Multivariable Regression Model

We will now develop a multivaribale regression model with all possible variables to identify significant variables to use in the final model.

```
lm_all <- lm(mpg ~ ., mtcars_data) # Multivariable linear model
summary(lm_all) # Note summary table hidden due to page limit See Table 2.
```

A better match is obtained compared to the basic linear model (higher R-squared value). However, the coefficients are not significant.

We will use the *Backward Selection Method* to determine the optimum model. This method will add/remove varaibles in order to find the optimum combination to include in the final model.

```
optimum_model <- step(lm_all, direction = 'both', trace = FALSE)
summary(optimum_model) # Note summary table hidden due to page limit. See Table 3.
```

This method identifies “wt” (weight), “qsec” (1/4 mile time), and “am” (transmission). Therefore, MPG is dependent on the transmission type, weight of the car and the acceleration of the car.

```
final_model <- lm(mpg ~ wt + qsec + am, mtcars_data)
summary(final_model) # Note summary table hidden due to page limit. See Table 4.
```

Please refer to **Fig. 3** and the following:

- The Residual vs Fitted chart appear random, with no pattern.
- The Noraml Q-Q chart shows the residuals arenormally distributed.
- The Scale-Location chart indicates the residuals are randomly distributed.
- The Residuals vs. Leverage indicates no outliers exist.

Please refer to section *1.2) Conclusion* for key takeaways and results.

## 5) Appendix

**Fig. 1**

```
Fig_1 <- ggplot(mtcars_data, aes(am, mpg))
Fig_1 + geom_boxplot(aes(fill = am)) + xlab("Transmission")
```

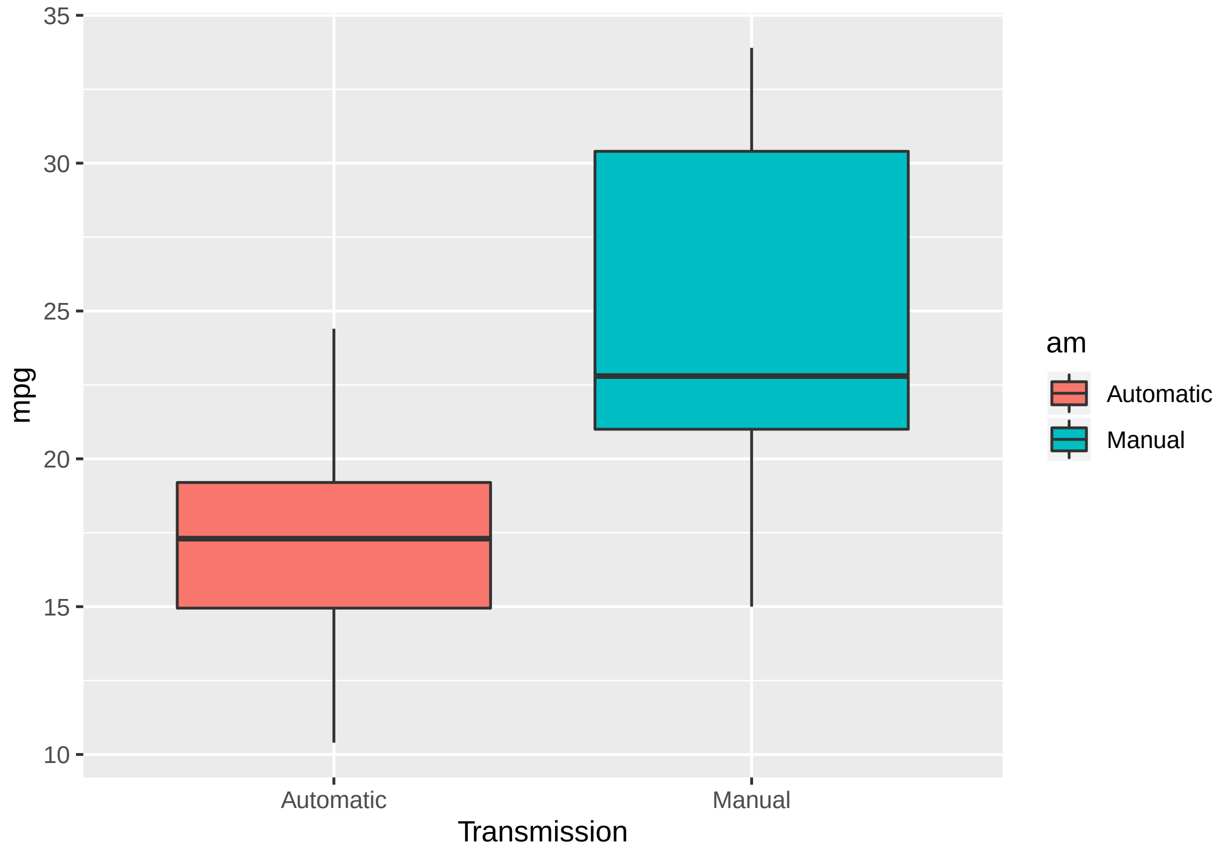

**Fig. 2**

```
pairs(mtcars_data, panel = panel.smooth, main="Pair Graph of Motor Trend Car Road Tests Data")
```

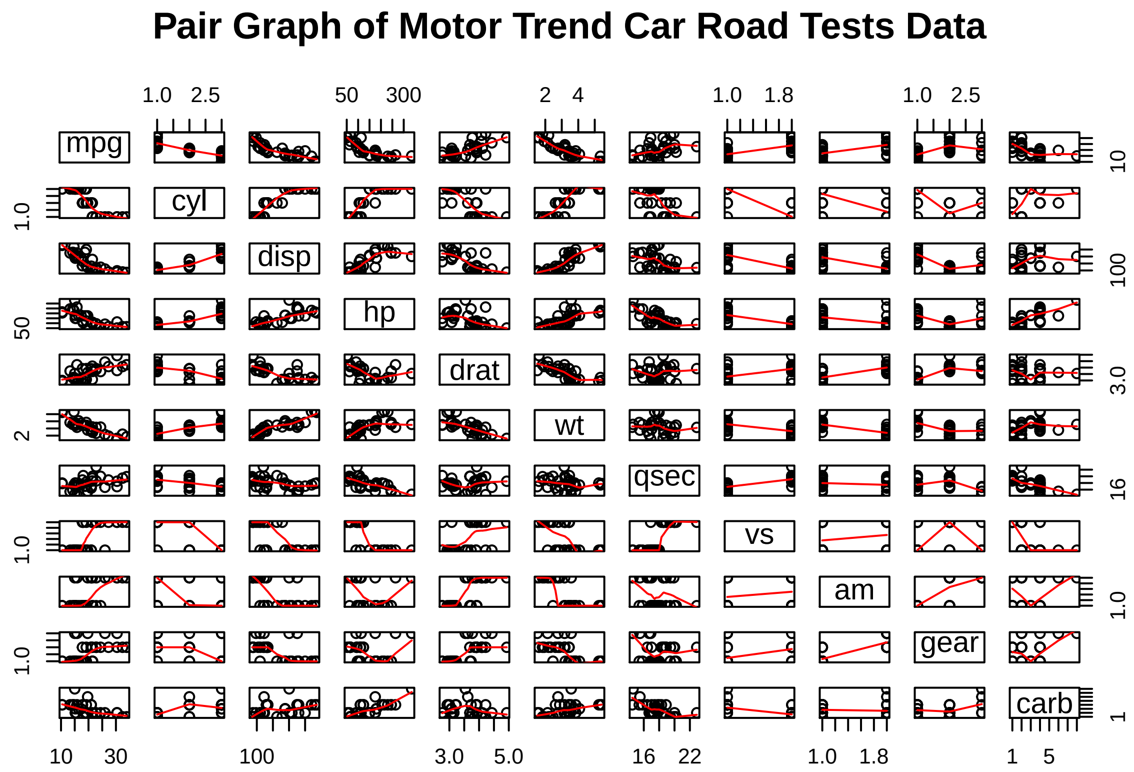

**Fig. 3**

```
autoplot(final_model)
```

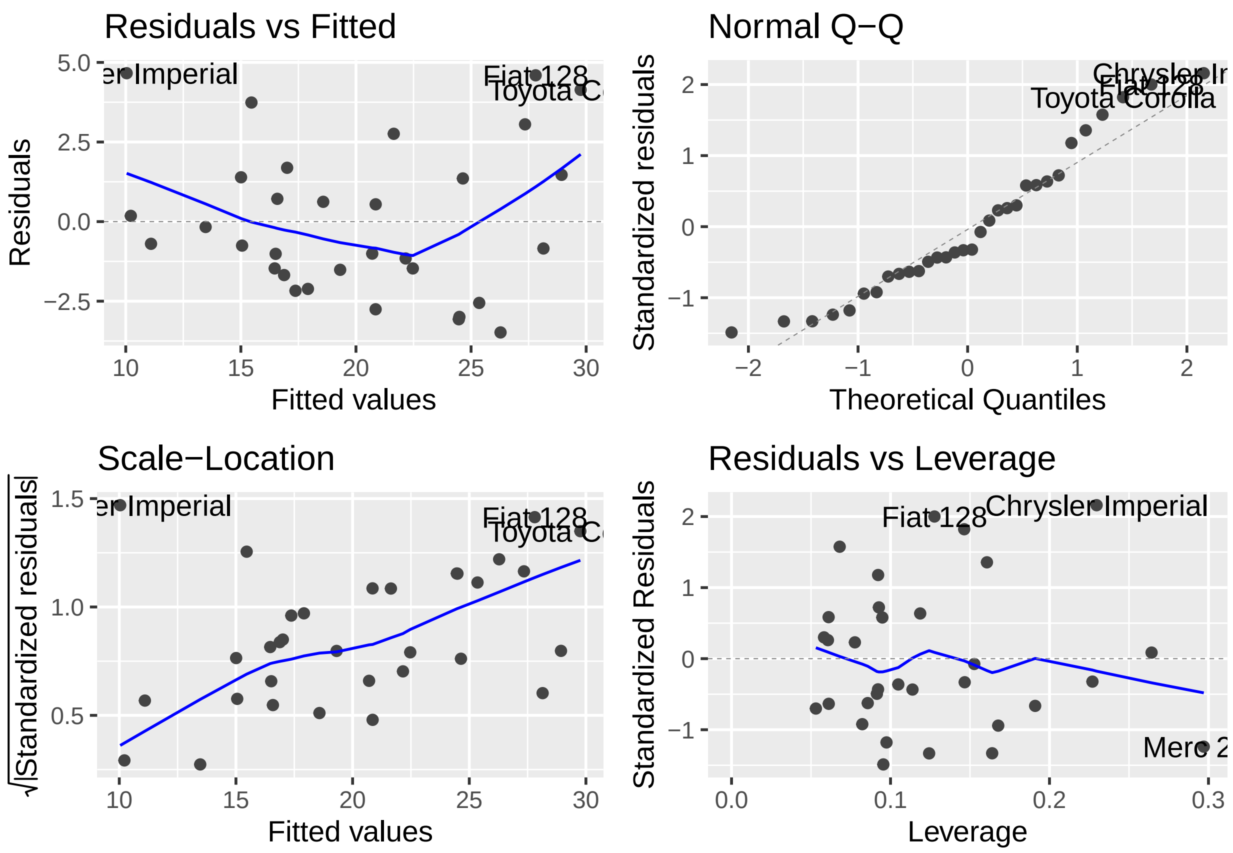

**Table 1**

```
head(mtcars_data,2) # Sample of the first 4 rows of data
```

```
##                mpg cyl disp  hp drat    wt  qsec       vs     am gear carb
## Mazda RX4       21   6  160 110  3.9 2.620 16.46 V-shaped Manual    4    4
## Mazda RX4 Wag   21   6  160 110  3.9 2.875 17.02 V-shaped Manual    4    4
```

```
dim(mtcars_data) # Dimensions of the mtcars_data
```

```
## [1] 32 11
```

**Table 2**

```
summary(lm_all)$coefficients
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 15.09261548 17.13627433  0.8807408 0.38946336
## cyl6        -1.19939698  2.38736481 -0.5023937 0.62116357
## cyl8         3.05491692  4.82986776  0.6325053 0.53459525
## disp         0.01256810  0.01774024  0.7084518 0.48726645
## hp          -0.05711722  0.03174603 -1.7991927 0.08789210
## drat         0.73576811  1.98461241  0.3707364 0.71493502
## wt          -3.54511861  1.90895437 -1.8570997 0.07886857
## qsec         0.76801287  0.75221895  1.0209964 0.32008122
## vsStraight   2.48849171  2.54014636  0.9796647 0.33956206
## amManual     3.34735713  2.28948094  1.4620594 0.16006890
## gear4       -0.99921782  2.94657533 -0.3391116 0.73824498
## gear5        1.06454635  3.02729599  0.3516492 0.72897110
## carb         0.78702815  1.03599487  0.7596834 0.45676696
```

**Table 3**

```
summary(optimum_model)$coefficients
```

```
##              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
```

**Table 4**

```
summary(final_model)$coefficients
```

```
##              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
```