The Impact of Vehicle Transmission on Miles per Gallon

Mark Dakkak

August 21, 2014

Executive Summary

This analysis is performed by Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of 32 cars, we explore the relationship between type of transmission and miles per gallon (MPG) (outcome). We consider various confounders and present a model that addresses the following two questions:

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

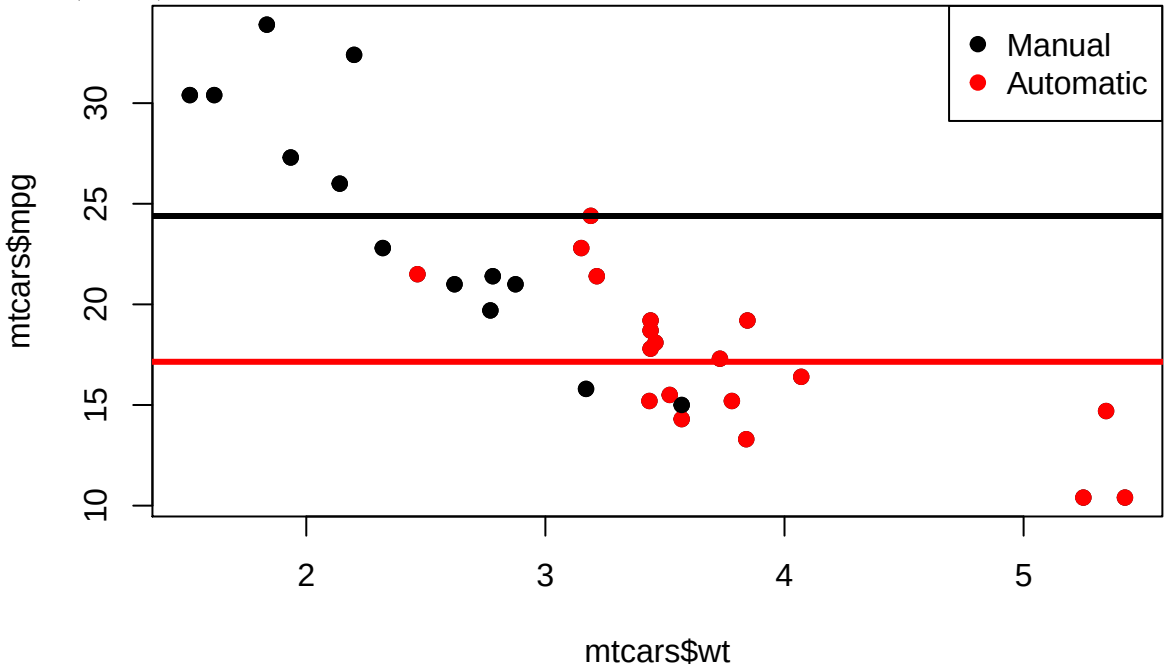
Data Processing

The data set mtcars is stored in the R package "datasets". The only changes we made to the variables are fixing variable types from numeric to factor.

```
library(datasets); data(mtcars)
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$am <- as.factor(mtcars$am)
```

Exploratory Data Analysis

To consider how vehicle transmission impacts miles per gallon, we investigated the relationship between MPG and weight. The graph below plots the 32 cars, with automatic cars plotted in red and manual cars plotted in black. The horizontal lines are at the mean MPG for automatic cars (red line) and the mean MPG for manual cars (black line). The distance between the two horizontal lines (7.2449) is equal to the average difference in MPG between manual cars and automatic cars.

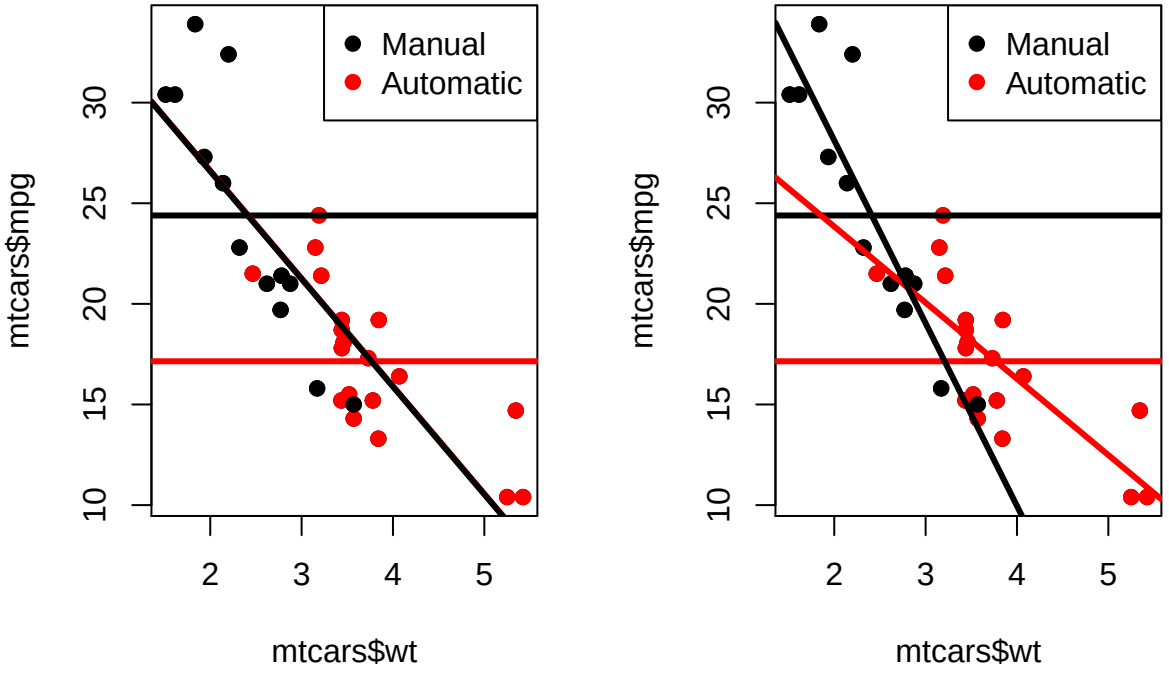


From the graph above, we can see that manual cars tend to be lighter and have higher MPGs, whereas automatic cars tend to be heavier and lower MPGs.

Model Selection

1. Impact on bivariate model
To start off, let's dig a little deeper into the relationship between weight and miles per gallon. Let's understand how vehicle transmission impacts this bivariate relationship. Below, we test whether there's a constant treatment effect (where the two regression lines would be parallel) or if the treatment effect depends on weight (where the two regression lines are not parallel).

```
lm_Parallel <- lm(mpg ~ wt + am, data = mtcars)
lm_Different <- lm(mpg ~ wt + am + wt * am, data = mtcars)
```



In the image above, the horizontal lines correspond with mean MPG for the types of vehicles. In the left image, the slope is assumed to be constant, meaning the relationship between weight and MPG is assumed to be the same between manual cars and automatic cars. We only see one line, because both lines are overlapping. Using automatic cars as the reference, the coefficient for manual cars is -0.0236 with a p-value of 0.9879. These values indicate that when the slope is assumed to be constant, transmission type does not significantly influence the relationship between weight and MPG.

However, once we relax this assumption, the right image illustrates that the relationship between MPG and weight differs dramatically between manual and automatic cars. The expected decrease in MPG for every additional 1000 pounds is -3.7859 for automatic cars and -5.2984 for manual cars. The p-value for the interaction term coefficient is 0.001, indicating that the two coefficients are significantly different. To confirm that the addition of the interaction term is necessary, we performed an likelihood ratio test using ANOVA. This gave a p value of 0.001017, indicating that including the interaction term has a significant impact on the model.

2. Impact in multivariate models
Now, let's step back to understand all the variables we should include in our model. The mtcars dataset includes 11 variables, including number of cylinders, rear axle ratio, weight, horsepower, engine displacement,

quarter mile time, and transmission type. First, let's start with the relationship of interest, between transmission and miles per galon. From there, we will incrementally add variables to the model and then use the ANOVA function to test the importance of the added variables.

```
lm1 <- lm(mpg ~ am, data = mtcars)
lm2 <- lm(mpg ~ wt + am, data = mtcars)
lm3 <- lm(mpg ~ wt + am + wt*am, data = mtcars)
lm4 <- lm(mpg ~ (cyl-1) + wt + am + wt*am, data = mtcars)
lm5 <- lm(mpg ~ hp + (cyl-1) + wt + am + wt*am, data = mtcars)
lm6 <- lm(mpg ~ drat + hp + (cyl-1) + wt + am + wt*am, data = mtcars)
lm7 <- lm(mpg ~ disp + drat + hp + (cyl-1) + wt + am + wt*am, data = mtcars)

anova(lm1, lm2, lm3, lm4, lm5, lm6, lm7)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + am
## Model 3: mpg ~ wt + am + wt * am
## Model 4: mpg ~ (cyl - 1) + wt + am + wt * am
## Model 5: mpg ~ hp + (cyl - 1) + wt + am + wt * am
## Model 6: mpg ~ drat + hp + (cyl - 1) + wt + am + wt * am
## Model 7: mpg ~ disp + drat + hp + (cyl - 1) + wt + am + wt * am
##   Res.Df RSS Df Sum of Sq     F  Pr(>F)
## 1     30 721
## 2     29 278  1       443 78.11 7.4e-09 ***
## 3     28 188  1        90 15.94 0.00057 ***
## 4     26 138  2        50  4.41 0.02386 *
## 5     25 130  1         8  1.33 0.26113
## 6     24 130  1         0  0.01 0.90638
## 7     23 130  1         0  0.01 0.90849
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the incremental addition of variables significantly affects the model, up until the addition of horsepower in model 5. Because order is important when adding variables into the model, the sensitivity analysis below adds rear axle ratio and engine displacement instead of horsepower, but finds that none of these additional variables significantly improve the model.

```
lm5b <- lm(mpg ~ drat + (cyl-1) + wt + am + wt*am, data = mtcars)
lm5c <- lm(mpg ~ disp + (cyl-1) + wt + am + wt*am, data = mtcars)

anova(lm1, lm2, lm3, lm4, lm5b)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + am
## Model 3: mpg ~ wt + am + wt * am
## Model 4: mpg ~ (cyl - 1) + wt + am + wt * am
## Model 5: mpg ~ drat + (cyl - 1) + wt + am + wt * am
##   Res.Df RSS Df Sum of Sq     F  Pr(>F)
## 1     30 721
```

```
## 2     29 278  1       443 80.20 2.8e-09 ***
## 3     28 188  1        90 16.37 0.00044 ***
## 4     26 138  2        50  4.53 0.02093 *
## 5     25 138  1         0  0.01 0.94288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm1, lm2, lm3, lm4, lm5c)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + am
## Model 3: mpg ~ wt + am + wt * am
## Model 4: mpg ~ (cyl - 1) + wt + am + wt * am
## Model 5: mpg ~ disp + (cyl - 1) + wt + am + wt * am
##   Res.Df RSS Df Sum of Sq     F  Pr(>F)
## 1     30 721
## 2     29 278  1       443 80.58 2.7e-09 ***
## 3     28 188  1        90 16.44 0.00043 ***
## 4     26 138  2        50  4.55 0.02060 *
## 5     25 137  1         1  0.12 0.72746
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus, to understand the impact of transmission on miles per gallon, we will use model 4, which estimates mpg using the number of cylinders, weight, transmission, and an interaction term between weight and transmission. Here is a summary of that model:

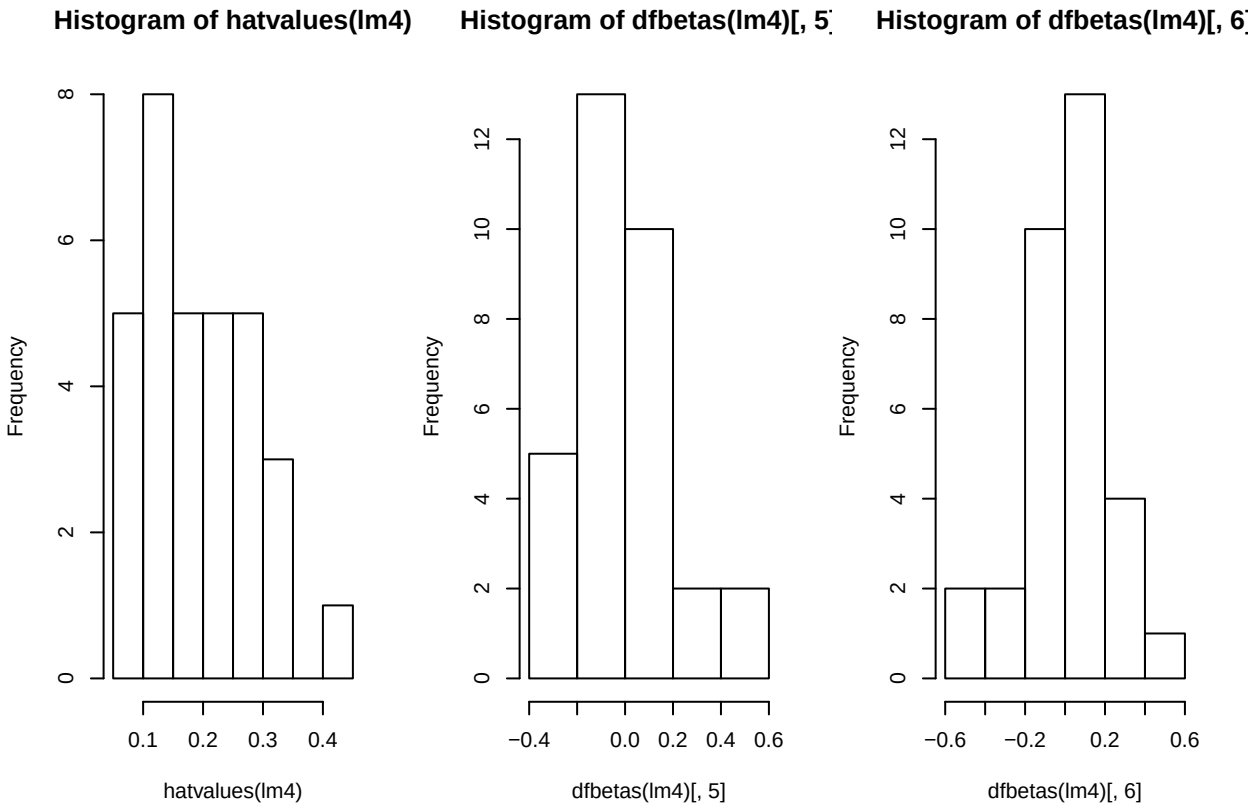
```
summary(lm4)$coefficients
```

```
##        Estimate Std. Error t value  Pr(>|t|)
## cyl4     29.775      2.840  10.483 7.871e-11
## cyl6     27.065      3.049   8.878 2.378e-09
## cyl8     24.999      3.485   7.174 1.283e-07
## wt       -2.399      0.844  -2.842 8.604e-03
## am1      11.569      4.088   2.830 8.854e-03
## wt:am1   -4.068      1.397  -2.911 7.296e-03
```

Because we included an interaction term in the model, the total impact of vehicle transmission on MPG is equal to the coefficient in front of the transmission variable (11.5688) plus the the coefficient in front of the interaction variable (-4.068) multiplied by the weight of the vehicle. Thus, **the impact of transmission on miles per galon depends on the weight of the car**. For a vehicle of average weight (3217.25 lb), manual transmission (compared to automatic transmission) would be expected to decrease MPG by -1.5189. For a vehicle of very high weight (5000 lb), manual transmission (compared to automatic transmission) would be expected to decrease MPG by -8.7711. Lastly, for a vehicle of very low weight (2000 lb), manual transmission (compared to automatic transmission) would be expected to increase MPG by 3.4328.

3. Diagnostics for multivariate model

Two useful diagnostic measures are hatvalues, which measure leverage, and dfbetas, which measure the change in coefficients when individual observations are deleted while fitting the model.



The histogram on the left shows the hatvalues for all 32 observations, and we can see that there is no observation with a very high hatvalue, indicating that there is likely no data entry errors. The histogram in the middle shows the dfbetas for the coefficient in front of the transmission variable term and the histogram on the right shows the dfbetas for the coefficient in front of the interaction term. It also appears that ignoring single observations does not significantly impact these coefficients.

Conclusion

Both sections in the model selection section above illustrate that the impact of vehicle transmission on MPG changes with the weight of the vehicle. Overall, for low weight vehicles, cars with manual transmission have improved MPG. However, for high weight vehicles, cars with automatic transmission have improved MPG. Therefore, there is no single value that can accurately quantify the difference in MPG between automatic and manual cars.