Motor Trend: Automatic vs Manual for Gas Mileage

Nikolai Alexander

Executive Summary

This study concludes that a manual transmission has 1.8 mpg more gas mileage that an automatic transmission with a 2.45 uncertaintly, keeping all other variables static. At first, we observed the correlation between transmission and gas mileage, ignoring all other variables; however, we only manages to fit 36% of the regression model. We then observed *mpg*'s correlation with number of cylinders, weight, displacement, and horsepower. Comparing these with the transmission gave us a 86.6% fit of the regression model, allowing us to come to our conclusion.

Analysis of Data

Gathering the Data

We are observing the *mtcars* data set, in compare gas mileage for **manual** transmission and **automatic** transmission for cars. So let's first look at the data set.

```
data("mtcars")
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
sapply(mtcars, class)
```

```
##       mpg       cyl      disp        hp      drat        wt      qsec 
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" 
##        vs        am      gear      carb 
## "numeric" "numeric" "numeric" "numeric"
```

As we can see, the class of all columns in the list are *numeric*. We want the categorical data (number of cylinders, transmission type, gear, etc) to be *factor*s to efficiently work with the data.

Manipulating the Data

```
for(i in c(2,8:11)){
  mtcars[,i] <- as.factor(mtcars[,i])
}

sapply(mtcars, class)
```

```
##       mpg       cyl      disp        hp      drat        wt      qsec 
## "numeric"  "factor" "numeric" "numeric" "numeric" "numeric" "numeric" 
##        vs        am      gear      carb 
##  "factor"  "factor"  "factor"  "factor"
```

We also would like *am* to be labled as **automatic** (0) or **manual** (1), for simplicity.

```
mtcars$am <- factor(mtcars$am,levels=c(0,1),labels=c("Automatic", "Manual"))

head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs        am gear
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0    Manual    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0    Manual    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1    Manual    4
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1 Automatic    3
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0 Automatic    3
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1 Automatic    3
##                   carb
## Mazda RX4            4
## Mazda RX4 Wag        4
## Datsun 710           1
## Hornet 4 Drive       1
## Hornet Sportabout    2
## Valiant              1
```

Hypothesis

Now that the data is in the form that we want, we can start analysing the correlation between transmission (*am*) and *mpg*. From observing a boxplot comparison of the transmission types (*Figure 1, Appendix A*), we hypothesize a very apparent increase in *mpg* when using an **manual** transmission. However, we must begin using Regression Analysis to prove this hypothesis futher.

Regression Models

T-Test

We can begin proving the strength of our hypothesis by using a t-test

We can separate the data by creating 2 subsets for **automatic** and **manual** transmissions

```
auto <- mtcars[mtcars$am == "Automatic",]
manu <- mtcars[mtcars$am == "Manual",]
```

Now onto the t-test…

```
ttest <- t.test(auto$mpg, manu$mpg)
ttest
```

```
##
##  Welch Two Sample t-test
##
## data:  auto$mpg and manu$mpg
## 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
```

We found the *p-value* to be **0.0013736**, which is well under the maximum accepted p-value of **0.05**. Therefore, our hypothesis is valid.

Simple Linear Regression

We can quantify our data using a simple linear model

```
slm <- summary(lm(mpg ~ am, data = mtcars))
slm
```

```
##
## 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 ***
## amManual       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
```

For this linear model, we can see that *Beta0*, or the estimate of mpg for an automatic transmission, is **17.1473684** and *Beta1*, or the difference in mpg between automatic and manual transmissions, is **7.2449393**. The *p-value* is **2.850207410^{-4}**, showing that this correlation is significant. However, R-squared is only **0.3597989**, meaning this linear model only fits 36% of the regression line. We can find a better fitted line by looking at the other variables

Multivariable Regression

Before adding variables to the regression model, we must determine what other data strongly correlates with mpg.

```
data(mtcars)    # resets all data to numerical to work with cor() function
cor(mtcars)[1,]
```

```
##        mpg        cyl       disp         hp       drat         wt 
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594 
##       qsec         vs         am       gear       carb 
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

As we can see, the variables that strongly correlate with mpg are number of cylinders (*cyl*), displacement (*disp*), gross horsepower (*hp*), and 1000 lbs of weight (*wt*). *Figure 2, Appendix A* gives a visualization of these correlations. Now that we found the other variables strongly affecting the *mpg*, we can create a new linear model

```
for(i in c(2,8:11)){    # Changes the class of all the categorical variables back to factors
  mtcars[,i] <- as.factor(mtcars[,i])
}
mtcars$am <- factor(mtcars$am,levels=c(0,1),labels=c("Automatic", "Manual"))

multivar <- lm(mpg ~ am + wt + cyl + disp + hp, data = mtcars)
multivar_lm <- summary(multivar)
multivar_lm
```

```
## 
## Call:
## lm(formula = mpg ~ am + wt + cyl + disp + hp, data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.9374 -1.3347 -0.3903  1.1910  5.0757 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 33.864276   2.695416  12.564 2.67e-12 ***
## amManual     1.806099   1.421079   1.271   0.2155    
## wt          -2.738695   1.175978  -2.329   0.0282 *  
## cyl6        -3.136067   1.469090  -2.135   0.0428 *  
## cyl8        -2.717781   2.898149  -0.938   0.3573    
## disp         0.004088   0.012767   0.320   0.7515    
## hp          -0.032480   0.013983  -2.323   0.0286 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344 
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

Now, including all other factors, we see that the difference in *mpg* between automatic and manual transmissions is **1.8060995**, much smaller than the **7.2449393** in the SLM. We see *Multiple R-squared* is **0.8664276**, meaning that this model fits 86.6% of the regression model. The residual values (*Figure 3 - Apendix A*) show us that the data follows the normal curve. This confirms a **1.8** mpg increase in gas milage, with a **2.4528254** uncertainty, when using a manual transmission vs an automatic transmission.

Appendix A

Figure 1: MPG by Transmission Type

```
boxplot(
  mpg ~ am, data = mtcars,
  xlab = "Transmission Type", ylab = "MPG"
)
```

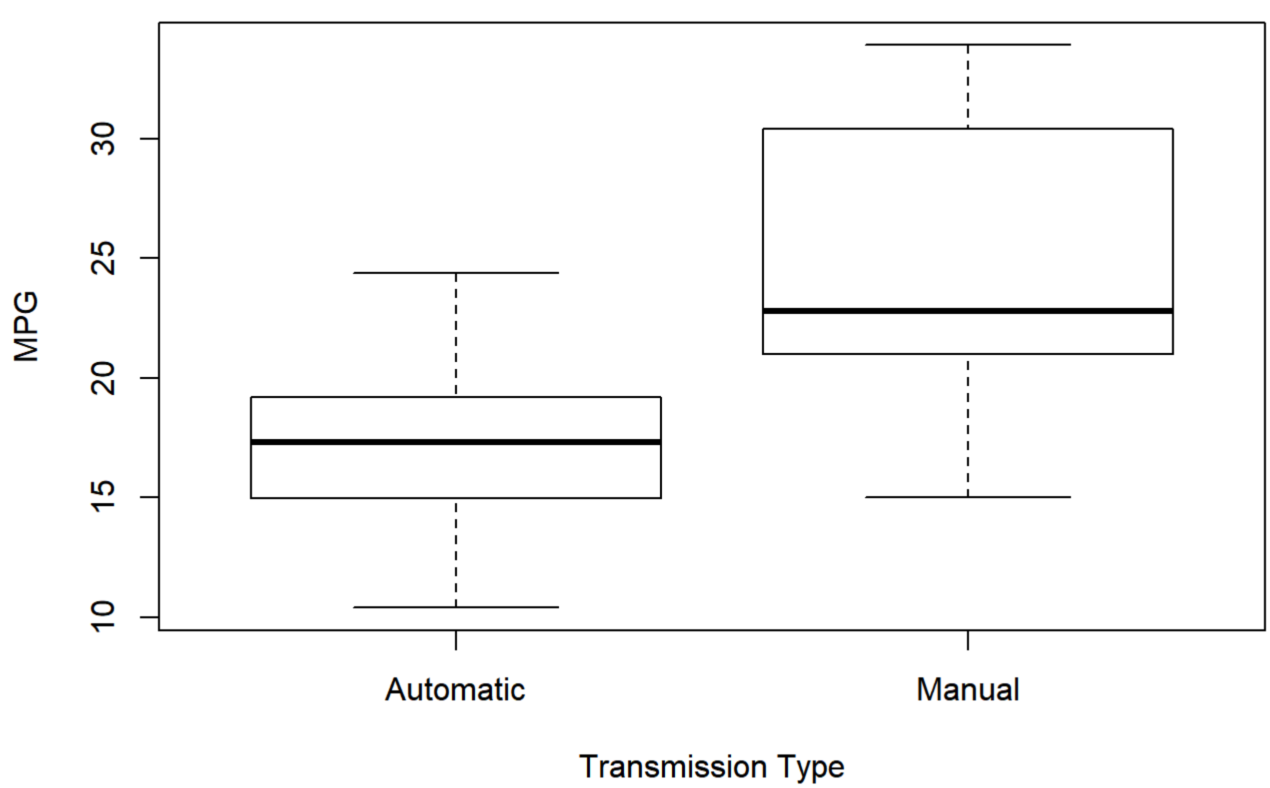


Figure 2: Correlation Between All Variables

```
pairs(mpg ~ ., data = mtcars)
```

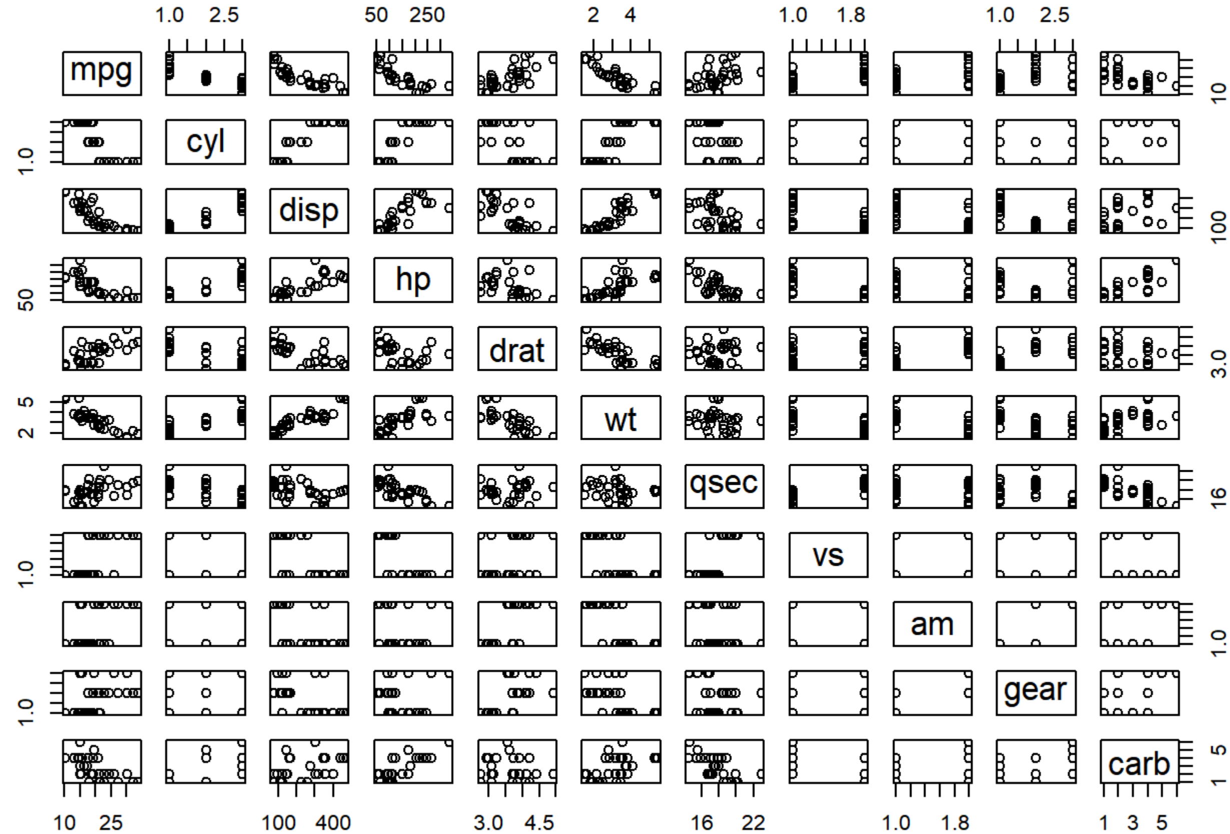
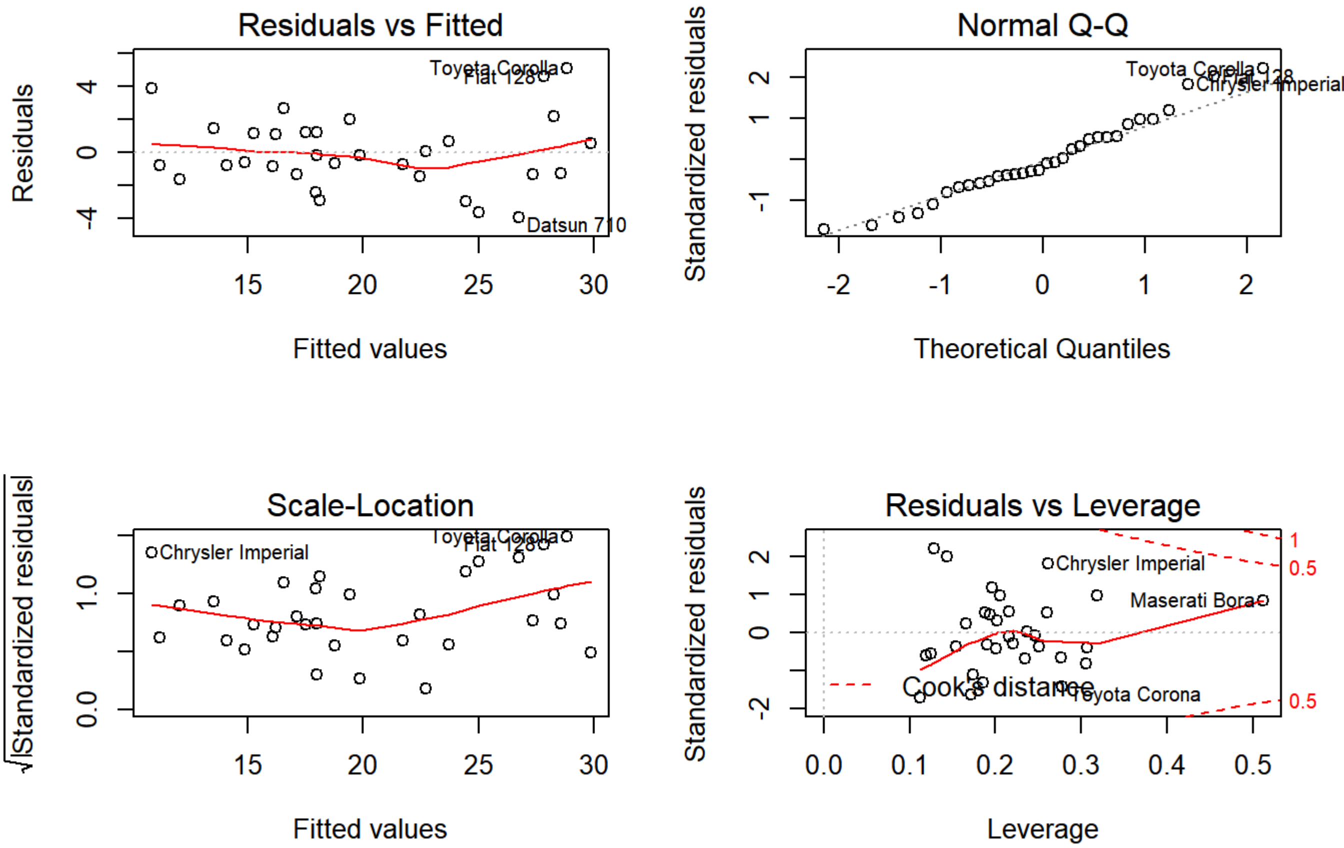


Figure 3: Residual Values

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


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Scale-Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Maserati Bora
Toyota Corona
Cook's distance
1
0.5
0.5