Comparision between Automatic and Manual Transmission in MT-cars

Shreyas Shukla

June 23, 2016

EXECUTIVE SUMMARY

This report includes subheadings:Inference, Simple Regression, Multiple Regression and Model Residuals which draws the conclusion that cars manual transmission have on average significantly higher MPGs with the car with automatic transmission. It also shows that manual transmission vehicles have 2.94 mpg more than automatic transmission vehicles under our bestfit model.

Introduction

Looking at the sets of collection of the cars, we are interested in exploring the relationship between a set of variables and mpg (outcome). We are particularly interested in two questions: 1.Is an automatic or manual transmission better for MPG? 2.Quantify the MPG difference between automatic and manual transmissions

Data Loading and Cleaning

For the purpose of this analysis we use mtcars dataset which is a dataset that was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

```
data(mtcars)
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
mtcars$am <- factor(mtcars$am, labels= c("Auto","Manual") )
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
```

```
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "Auto","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

Exploratory Analysis

The dataset srtucture shows us that the data frame has 32 observations on 11 variables:

mpg - Miles/(US) gallon

cyl - Number of cylinders

disp - Displacement (cu.in.)

hp - Gross horsepower,

drat - Rear axle ratio

wt - Weight (lb/1000),

qsec - 1/4 mile time

vs V/S (0 = V engine, 1 = straight engine)

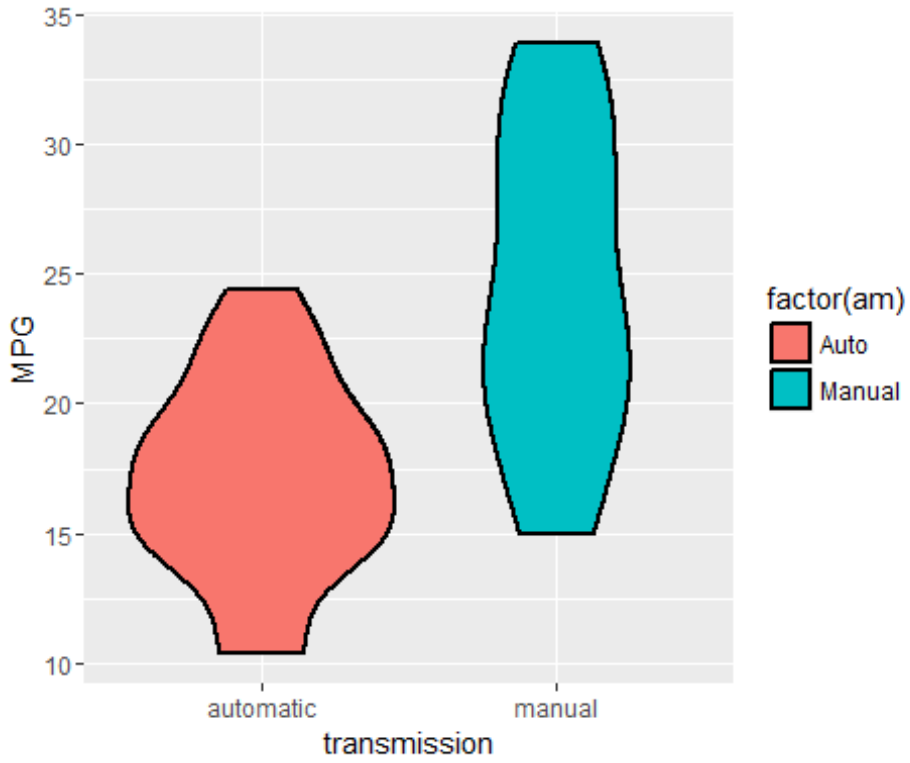
am - Transmission (0 = automatic, 1 = manual)

gear - Number of forward gears

carb - Number of carburetors.

It is also worthwhile check how MPG varies by automatic versus manual transmission. For that purpose we create a Violin plot of MPG by automatic and manual transmissions

```
library(stats)
library(ggplot2)
ggplot(mtcars, aes(y=mpg, x=factor(am, labels = c("automatic", "manual")),
fill=factor(am)))+geom_violin(colour="black", size=1)+xlab("transmission")
+ylab("MPG")
```



It appears that automatic cars have a lower miles per gallon, and hence a lower fuel efficiency, than manual cars. But it is possible that this apparent pattern happened by random chance- that is, that we just happened to pick a group of automatic cars with low efficiency and a group of manual cars with higher efficiency. So to check whether that's the case, we have to use a statistical test.

Inference

Let us first test the hypothesis that cars with an automatic transmission use more fuel than cars with a manual transmission

```
x <- t.test(mpg~am, data = mtcars)
x

##
##  Welch Two Sample t-test
##
## data:  mpg by am
## 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 in group Auto mean in group Manual
##             17.14737             24.39231
```

The confidence interval (95%) does not contain zero (-11.28,-3.21) and it can concluded that the average consumption, in miles per gallon, with automatic transmission is higher than the manual transmission. In this case, the mean analysis, it is possible to quantify the MPG difference between automatic and manual transmissions: 7.24 mpg greater, subtracting means. p-value that shows the probability that this apparent difference between the two groups could appear by chance is very low.

Simple Linear Regression model

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

```
##
## 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
```

Since the p-value = 0.000285 is less than 0.05 so we rejected null hypothesis. The adjusted R squared value is 0.3385 which means our model only explains 33.85% of the variance. We need to include other predictor variables.

Mulitiple Regression Model

Among available methods we decide to perform stepwise selection to help us select a subset of variables that best explain the MPG.

```
z <- step(lm(mpg~., data = mtcars),trace = 0, steps = 10000)
summary(z)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

The adjusted R-squared value of 0.84 which is the maximum obtained considering all combinations of variables. From these results we can conclude that more than 84% of the variability is explained by the above model. On average, manual transmission cars have 2.94 MPGs more than automatic transmission cars. However this effect was much higher than when we did not adjust for weight and qsec. Now we compare the base model with only am as the predictor variable and the best model which we obtained above containing confounder variables also.

```
anova(y,z)
```

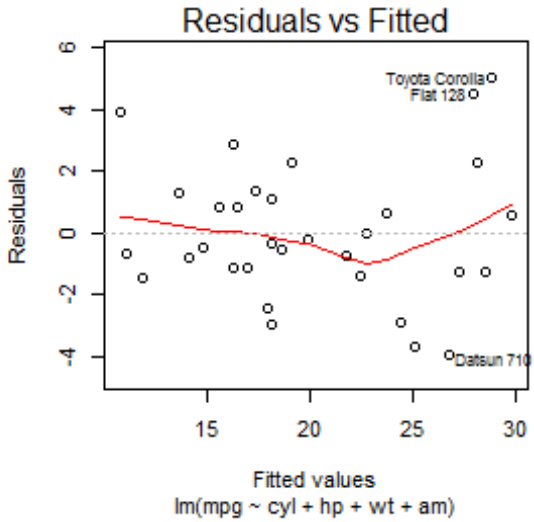
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

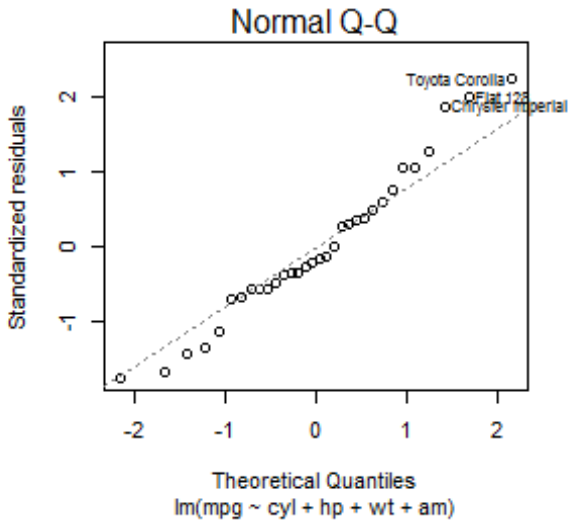
Looking at the above results, the p-value obtained is highly significant and we reject the null hypothesis that the confounder variables cyl, hp and wt don't contribute to the accuracy of the model.

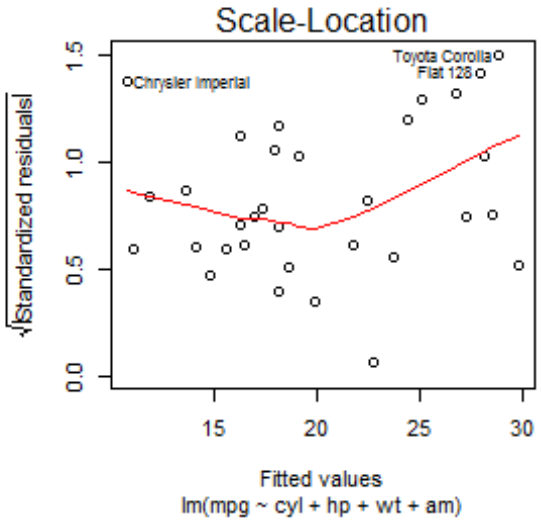
Model Residuals

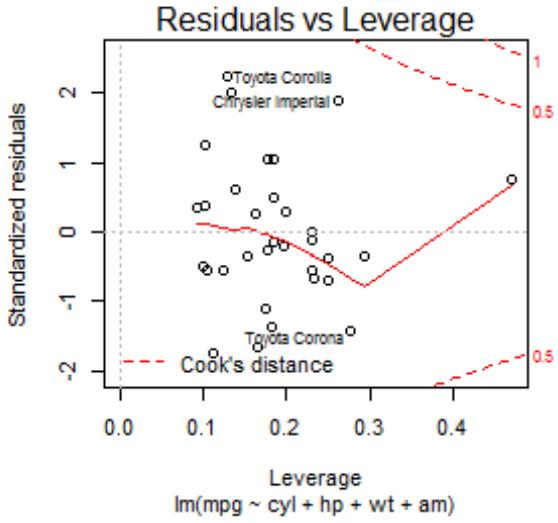
In this section, we have the residual plots of our regression model along with computation of regression diagnostics for our liner model. This excercise helped us in examining the residuals and finding leverage points to find any potential problems with the model.

```r
par(mfrow = c(4,1), fin = c(3,3))
i <- plot(z)
```









Following observations are made from the above plots.

- The points in the Residuals vs. Fitted plot are randomly scattered on the plot that verifies the independence condition.

- In, The normal Q-Q plot, we're trying to figure out the normality of the errors by plotting the theoretical quantiles of the standard normal distribution by the standardized residuals.
- The Scale-Location plot consists of points scattered in a constant band pattern, indicating constant variance.
- There are some distinct points of interest (outliers or leverage points) in the top right of the plots that may indicate values of increased leverage of outliers.