Motor Trend Car Road Tests - Effects of automatic and manual transmission on MPG

Lokesh Aggarwal

November 11, 2016

Executive Summary

Motor Trend, an Automobile Trend Magazine is interested in exploring the relationship between a set of variables and Miles Per Gallon (MPG) outcome. In this project, we will analyze the mtcars dataset from the 1974 Motor Trend US magazine to answer the following questions:

-Is an Automatic or Manual transmission better for Miles Per Gallon (MPG)?

-How different is the MPG between automatic and manual transmissions?

Using Simple Linear Regression analysis, we determine that there is a signficant difference between the mean MPG for Automatic and Manual transmission cars. Manual transmissions achieve a higher value of MPG compared to Automatic transmission. This increase is approximately 1.8 MPG when switching from an Automatic transmission to a Manual one, with all else held constant.

Data Processing

We begin by loading in the mtcars dataset and transforming certain variables into factors.

```r
data(mtcars)
mtcars$cyl  <- factor(mtcars$cyl)
mtcars$vs   <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am   <- factor(mtcars$am,labels=c("Automatic","Manual"))
```

Exploratory Analysis

Now we explore various relationships between variables of interest. First, we plot the relationship between all the variables of the mtcars dataset. We learned from this plot that the variables cyl, disp, hp, drat, wt, vs and am have a strong correlation with mpg (Appendix - Figure 1).

In this analysis, we are interested in the effects of car transmission type on mpg (Appendix - Figure 2). So, we look at the distribution of mpg for each level of am (Automatic or Manual) by plotting box plot. This plot clearly depicts that manual transmissions tend to have higher MPG. This data is further analyzed and discussed in regression analysis section by fitting a linear model.

Regression Analysis

In this section, we build linear regression models using different variables in order to find the best fit and compare it with the base model using anova. After model selection, we also perform analysis of residuals.

Model building and selection

Our initial model includes all variables as predictors of mpg. Then we perform stepwise model selection in order to select significant predictors for the final, best model.The step function will perform this selection by calling lm repeatedly to build multiple regression models and select the best variables from them using both forward selection and backward elimination methods. This ensures that we have included useful variables while omitting ones that do not contribute significantly to predicting mpg.

```r
initialmodel <- lm(mpg ~ ., data = mtcars)
bestmodel <- step(initialmodel, direction = "both")
```

The best model obtained from the above computations shows that variables, cyl, wt and hp as Confounders and am as the independent variable. Details of the model are depicted below.

```r
summary(bestmodel)
```

```
##
## 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.

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.

```r
basemodel <- lm(mpg ~ am, data = mtcars)
anova(basemodel, bestmodel)
```

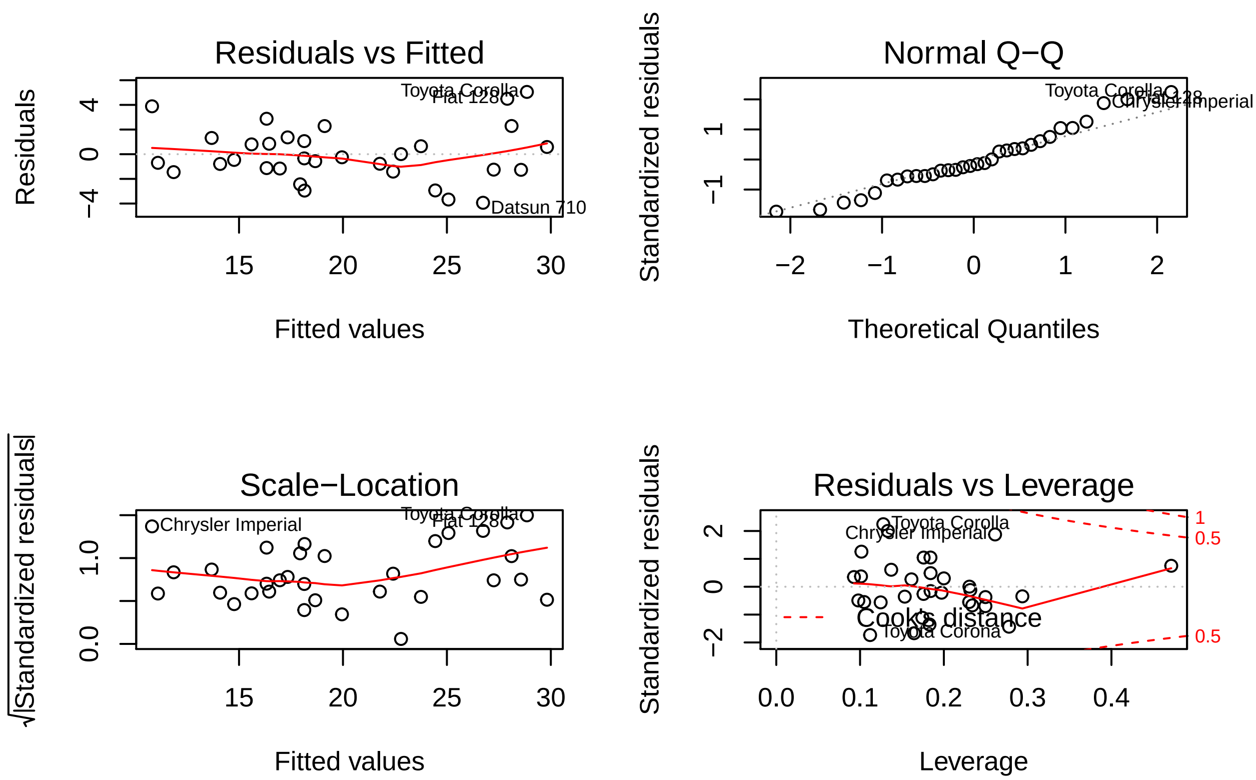
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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 and Diagnostics

In this section, we have the residual plots of our Regression Model along with computation of Regression diagnostics for our Linear 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(2, 2))
plot(bestmodel)
```



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.
- The Normal Q-Q plot consists of the points which mostly fall on the line indicating that the residuals are normally distributed.
- 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.

In the following section, we show computation of some regression diagnostics of our model to find out these leverage points. We compute top three points in each case of influence measures. The data points with the most leverage in the fit can be found by looking at the hatvalues() and those that influence the model coefficients the most are given by the dfbetas() function.

```
leverage <- hatvalues(bestmodel)
tail(sort(leverage),3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##           0.2777872           0.2936819           0.4713671
```

```
influential <- dfbetas(bestmodel)
tail(sort(influential[,6]),3)
```

```
## Chrysler Imperial          Fiat 128     Toyota Corona
##         0.3507458         0.4292043         0.7305402
```

Looking at the above results, we notice that our analysis was correct, these are the same cars as mentioned in the residual plots.

Statistical Inference

In this section, we perform a t-test on the two subsets of mpg data: manual and automatic transmission assuming that the transmission data has a normal distribution and tests the null hypothesis that they come from the same distribution. Based on the t-test results, we reject the null hypothesis that the mpg distributions for manual and automatic transmissions are the same.

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

```
##
##  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 Automatic    mean in group Manual
##                17.14737                24.39231
```

Conclusions

Based on the analysis done in this project, we can conclude that:

- Cars with Manual transmission get 1.8 more miles per gallon compared to cars with Automatic transmission. (1.8 adjusted for hp, cyl, and wt).
- mpg will decrease by 2.5 for every 1000 lb increase in wt.
- mpg decreases negligibly (only 0.32) with every increase of 10 in hp.
- If number of cylinders, cyl increases from 4 to 6 and 8, mpg will decrease by a factor of 3 and 2.2 respectively (adjusted by hp, wt, and am).

Appendix

Figure 1 - Pairs plot for the “mtcars” dataset

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

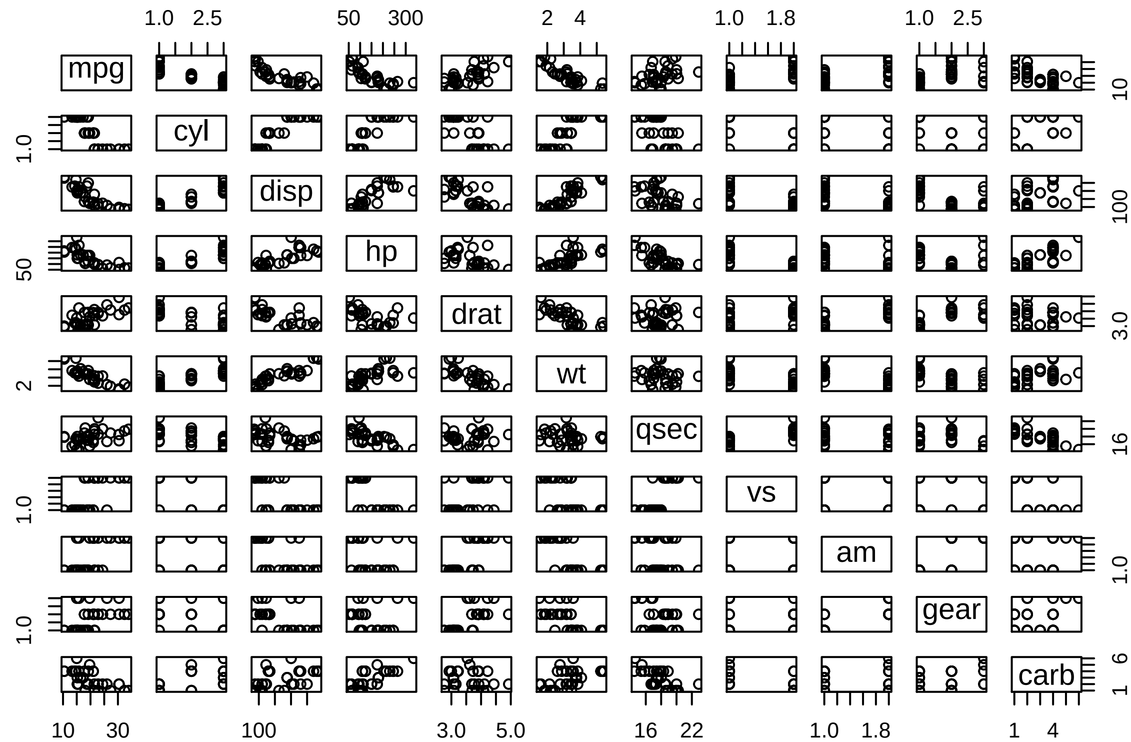
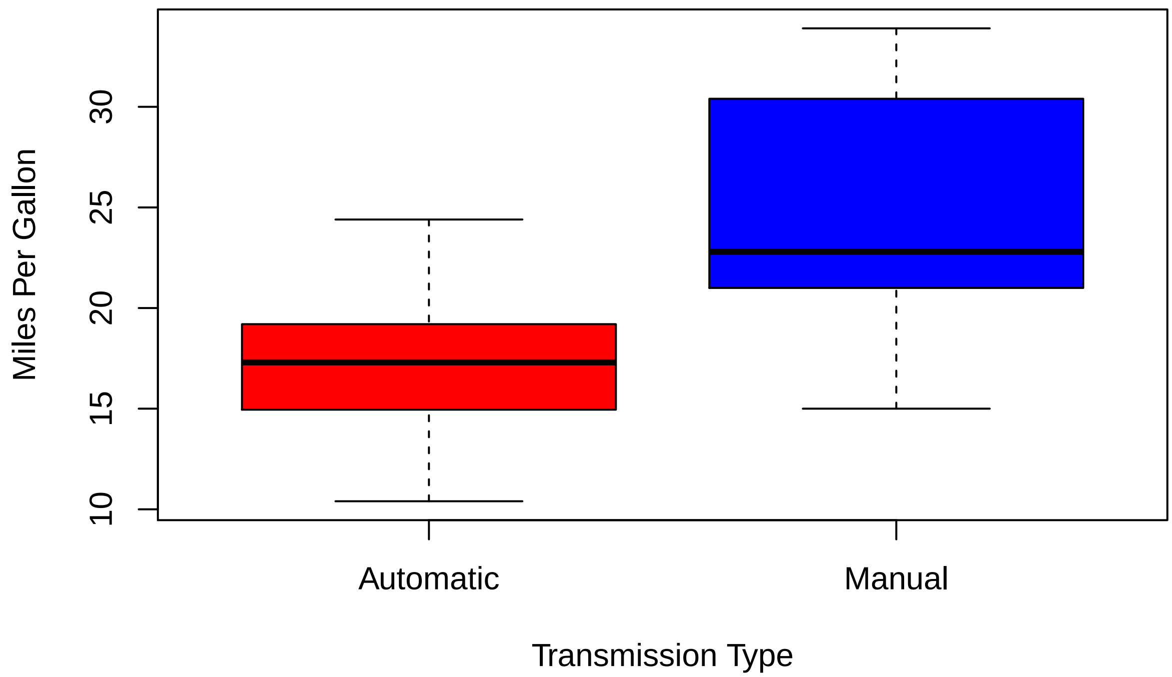


Figure 2 - Boxplot of miles per gallon by transmission type

```r
boxplot(mpg ~ am, data = mtcars, col = (c("red","blue")), ylab = "Miles Per Gallon", xlab = "Transmissi
```


Miles Per Gallon
10
15
20
25
30
Automatic
Manual
Transmission Type