Motor Trends Car Study

Jason Battles

October 15, 2016

Executive Summary

Motor Trend, a magazine about the automobile industry, requested this sample study to investigate the impact of several variables on on miles per gallan (MPG). This study answers a few questions.

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions.
3. Is there a more effective model for MPG prediction than transmission type?

This study reveals that vehicles with manual transmissions do ineed have a slightly better fuel efficienciency than those with automatic transmissions. We also determine a better model of MPG prediction than transmission type.

Environment Preparation

We configure the environment, load the data set, perform some necessary data transformations, and verify that the data is loaded correctly. This working data set has 32 observations of 11 variables.

```r
library(ggplot2)
library(corrplot)
data(mtcars)
mtcars.orig <- mtcars    # Maintain original copy for Correlation Matrix
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'))
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 "Automatic","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 Data Analysis

The focus of our study is `mpg` so let us first see how it is distributed. We find that the median and mean are close to each other, so we can assume a normal distribution of data.

```
summary(mtcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.42   19.20   20.09   22.80   33.90
```

In this analysis, we are interested in the effects of car transmission type (Automatic or Manual) on MPG so we start our exploration by creating a box plot *(Appendix 1 - Figure 1)*. This plot clearly depicts that manual transmissions tend to have higher MPG than automatic.

Independent Variable Selection

We now more closely explore the available variables to identify those which are most significant to any regression models that we may build. We conduct this exploration with a correlation matrix between all the variables in the data set. We learn that the variables `cyl`, `disp`, `hp`, `drat`, `wt`, `vs` and `am` *(transmission type)* have a strong correlation with dependent variable `mpg` *(sig.level = 0.0005)*. Variables `qsec`, `gear`, and `carb` are NOT significant and will be excluded from further analysis. **(Appendix 1 - Figure 2)**.

This data is further analyzed and discussed in regression analysis section by fitting a linear model.

Regression Analysis

In this section, we start building linear regression models to determine the optimal fit. We compare the models using `anova` tables.

Simple Linear Regression

We begin our regression analysis with our original assumption that `mpg` is mostly dependent on `am` (transmission type).

```
first.mdl <- lm(mpg ~ factor(am), data = mtcars)
summary(first.mdl)
```

```
##
## Call:
## lm(formula = mpg ~ factor(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 ***
## factor(am)Manual    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
```

We can see that the adjusted R squared value is only 0.338 which means that only **33.8%** of the regression variance can be explained by this univariate simple linear regression model. However, as outlined earlier by the Correlation Matrix, we must not forget there are several other predictor variables that we must take into account to see if any may lead to a better model for MPG.

Multivariate Regression Model

Although we identified the most highly correlated variables in the Correlation Matrix, we would like to sharpen our model further by conducting an **AIC** analysis using the `step` function.

Using this method, we build an initial model with all variables as possible predictors and then perform an iterative model selection process using the `step` method. This method runs `lm` multiple times to build multiple regression models and selects the best variables using both forward selection and backward elimination methods using the AIC algorithm. The full output is included in *Appendex 1 - Figure 3*.

```
init.mdl <- lm(mpg ~ ., data = mtcars)
best.mdl <- step(init.mdl, direction = "both")
```

The best model resulting in the lowest **AIC** value (AIC=61.65) is `lm(mpg ~ cyl + hp + wt + am, data = mtcars)`. This model depicts the variables `cyl`, `hp`, and `wt` as confounders and `am` as the independent variable. Details of the model are depicted below.

```
summary(best.mdl)
```

```
##
## 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
```

We observe that the R squared value of this updated model is 0.8401 which means that **84%** of the variability is now explained.

We now compare the base model (only transmission type as predictor) with the best model which we obtained earlier containing the additional confounder variables.

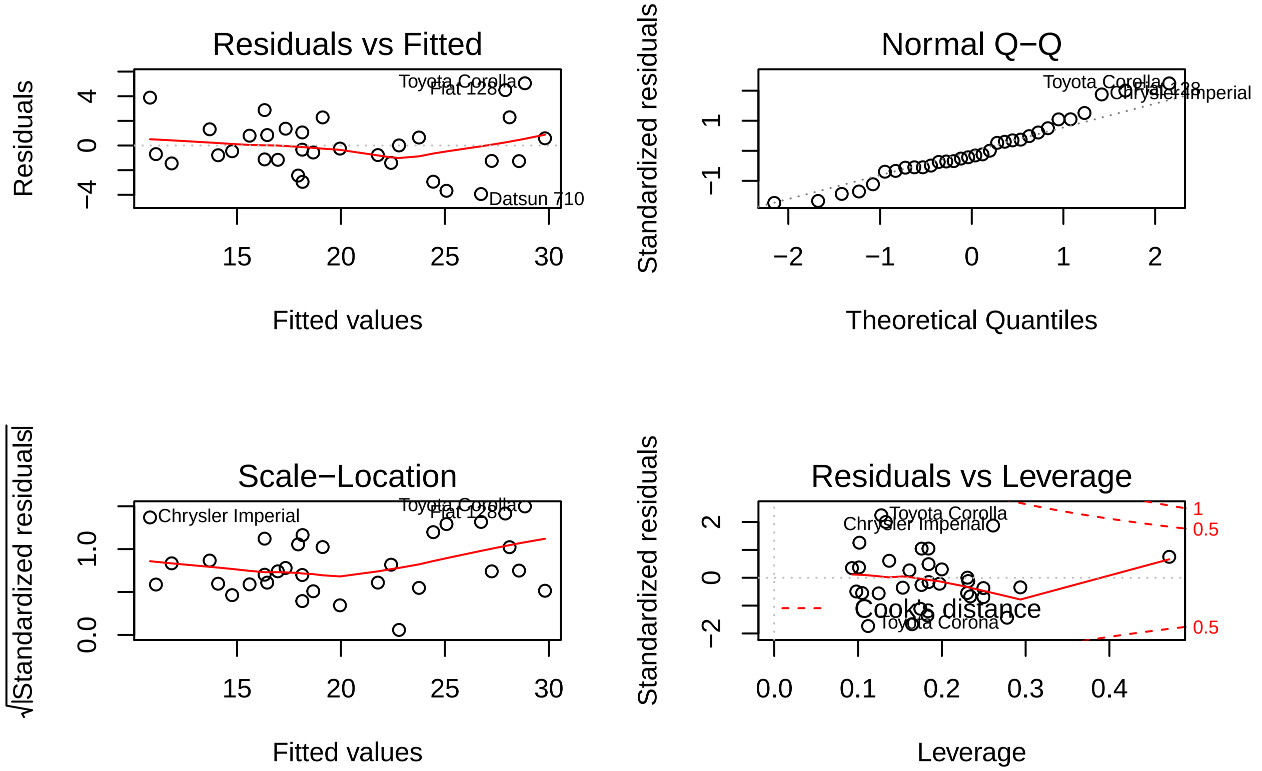
```
anova(first.mdl, best.mdl)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(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
```

The P-Value is highly significant so we may reject the null hypothesis that the confounder variables `cyl`, `hp`, and `wt` do not contribute to the accuracy of the model.

Model Residuals and Diagnostics

```
par(mfrow=c(2, 2))
plot(best.mdl)
```



* The points in the Residuals vs. Fitted plot are randomly scattered on the plot which 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.

Conclusions

1. Manual transmissions result in higher MPG than automatic transmissions.
2. Manual transmissions will achieve and additional 1.8mpg compared to automatic transmissions.
3. The optimal regression model includes `cyl`, `hp`, `wt`, `am` (transmission type)

   - Basline for the model is a 4 cylinders engine. 6 cylinders will decrease `mpg` by 3.03.

   - 8 cylinders will decrease `mpg` by an additional 2.16.
   - `mpg` will decrease by 2.49 for every 1000lb increase in `wt`
   - `mpg` will slightly decrease by 0.32 with every increase in 10hp

Appendix 1

Figure 1 - Box Plot (MPG vs. Transmission Type)

```
p <- ggplot(mtcars, aes(factor(am), mpg))
p + geom_boxplot() + geom_jitter() +
    labs(title="Box Plot of MPG vs. Transmission Type", x = "Transmission")
```

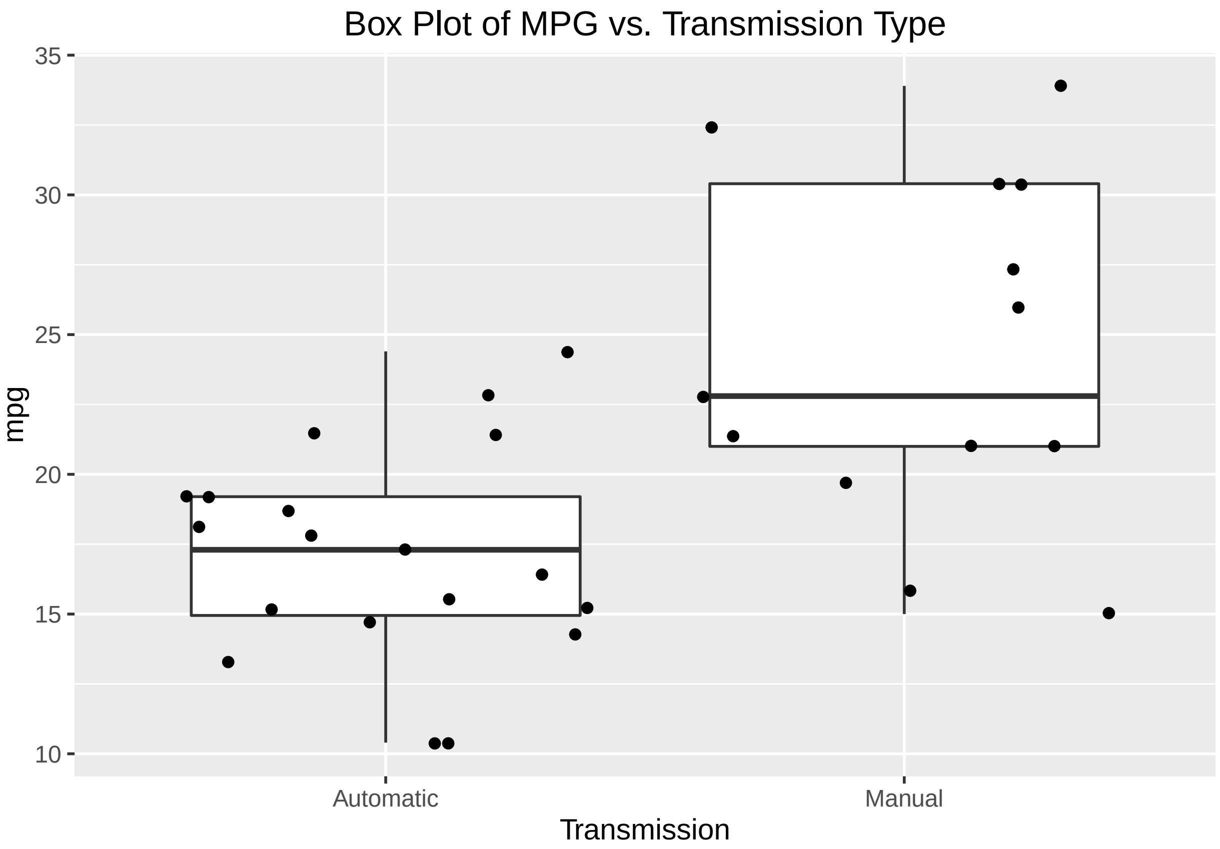


Figure 2 - Multivariate Correlation Matrix

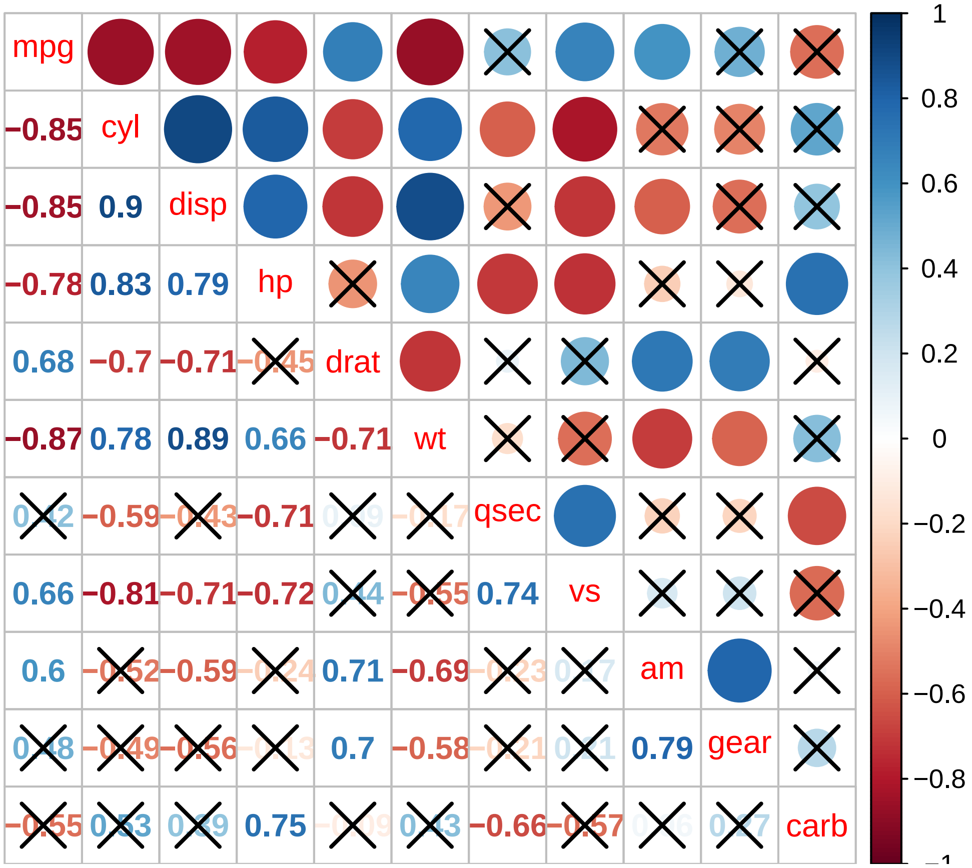


Figure 3 - Output from AIC Step Analysis

- This section removed due to page length restrictions*